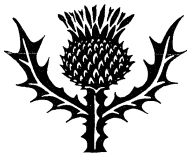
ENCYCLOPÆDIA
BRITANNICA



MACROPÆDIA

The Encyclopædia Britannica
is published with the editorial advice
of the faculties of the University of Chicago;
a committee of persons holding
academic appointments at the universities
of Oxford, Cambridge, London, and Edinburgh;
a committee at the University of Toronto;
and committees drawn from members of the faculties
of the University of Tokyo
and the Australian National University.

THE UNIVERSITY OF CHICAGO

"Let knowledge grow from more to more
and thus be human life enriched."

The New

Encyclopædia Britannica

in 30 Volumes

MACROPÆDIA
Volume 13

Knowledge in Depth

FOUNDED 1768
15TH EDITION

Encyclopædia Britannica, Inc.
William Benton, Publisher, 1943–1973
Helen Hemingway Benton, Publisher, 1973–1974
Chicago/London/Toronto/Geneva/Sydney/Tokyo/Manila/Seoul

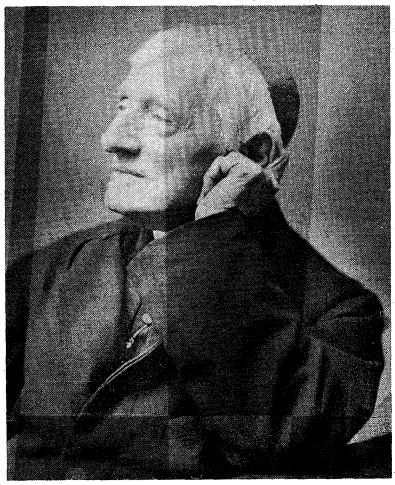
Printed in U.S.A.

Library of Congress Catalog Card Number: 77-94292
International Standard Book Number: 0-85229-339-9

Newman, John Henry

A leader of the Oxford Movement in the Church of England and then, after his conversion, a leader of the Roman Catholic Church, Cardinal Newman was one of the most remarkable churchmen and men of letters of the 19th century.

By courtesy of the Gernsheim Collection, the University of Texas at Austin



Newman, photograph by Herbert Barraud.

Life before conversion. Newman was born in London on Feb. 21, 1801. After pursuing his education in an evangelical home and at Trinity College, Oxford, he was made a fellow of Oriel College, Oxford, in 1822; vice principal of Alban Hall in 1825; and vicar of St. Mary's, Oxford, in 1828.

Under the influence of the clergyman John Keble and Richard Hurrell Froude, Newman became a convinced high churchman (one of those who emphasized the Anglican Church's continuation of the ancient Christian tradition, particularly as regards the episcopate, priesthood, and sacraments).

The Oxford Movement

When the Oxford Movement began he was its effective organizer and intellectual leader, supplying the most acute thought produced by it. A high church movement within the Church of England, the Oxford Movement was started at Oxford in 1833 with the object of stressing the Catholic elements in the English religious tradition, and of reforming the Church of England after the pattern of the Church of the first five centuries AD. Newman's editing of the *Tracts for the Times* and his contributing of 24 tracts among them were less significant for the influence of the movement than his books, especially the *Lectures on the Prophetical Office of the Church* (1837), the classic statement of the Tractarian doctrine of authority; the *University Sermons* (1843), similarly classical for the theory of religious belief; and above all his *Parochial and Plain Sermons* (1834–42), which in their published form took the principles of the movement, in their best expression, into the country at large. In 1838 and 1839 Newman was beginning to exercise far-reaching influence in the Church of England, because the stress upon the dogmatic authority of the church was felt to be a much-needed re-emphasis in a new liberal age, because he seemed so decisively to know what he stood for and where he was going, because in the quality of his personal devotion his followers found a man who practiced what he preached, and because he had been endowed with the gift of writing sensitive and sometimes magical prose.

Newman was contending that the Church of England represented true catholicity and that the test of this catholicity (as against Rome upon the one side and what he termed "the popular Protestants" upon the other) lay in the teaching of the ancient and undivided church of the Fathers. From 1834 onward this middle way was beginning to be attacked on the ground that it undervalued the Reformation; and when in 1838–39 Newman and Keble published Froude's *Remains*, in which the Reformation was violently denounced, moderate men began to suspect their leader. Their worst fears were confirmed in 1841 by Newman's *Tract 90*, which, in reconciling the Thirty-Nine Articles with the teaching of the ancient and undivided church, appeared to some to assert that the articles were not incompatible with the doctrines of the Council of Trent; and Newman's extreme disciple, W.G. Ward, claimed that this was indeed the consequence. Bishop Richard Bagot of Oxford requested that the tracts be suspended; and in the distress of the consequent denunciations Newman increasingly withdrew into isolation, his confidence in himself shattered and his belief in the catholicity of the English church weakening. He moved out of Oxford to his chapelry of Littlemore, where he gathered a few of his intimate disciples and established a quasi-monastery.

Newman's conversion

He resigned St. Mary's, Oxford, on Sept. 18, 1843, and preached his last Anglican sermon ("The Parting of Friends") in Littlemore Church a week later. He delayed long, because his intellectual integrity found an obstacle in the historical contrast between the early church and the modern Roman Catholic Church; but meditating upon the idea of development, a word then much discussed in connection with biological evolution, he applied the law of historical development to the Christian society and tried to show (to himself as much as to others) that the early and undivided church had developed rightly into the modern Roman Catholic Church and that the Protestant churches represented a break in this development, both in doctrine and in devotion. These meditations removed the obstacle and on Oct. 9, 1845, he was received at Littlemore into the Roman Catholic Church, publishing a few weeks later his *Essay on the Development of Christian Doctrine*.

Life after conversion. Newman went to Rome to be ordained to the priesthood and after some uncertainties founded the Oratory at Birmingham in 1848. He was suspect to the more rigorous among Roman Catholic clergy because of the quasi-liberal spirit that he seemed to have brought with him (his mode of expressing the idea of doctrinal development, his teaching on the nature of faith); and therefore, though in fact he was no liberal in any normal sense of the word, his early career as a Roman Catholic priest was marked by a series of frustrations, as he at least felt them to be. In 1852–53 he was convicted of libelling the apostate former Dominican priest Achilli. He was summoned to Ireland to be the first rector of the new Catholic university in Dublin, but the task was in the conditions impossible, and the only useful result was his lectures on the *Idea of a University* (1852). His role as editor of the Roman Catholic monthly, the *Rambler*, and in the endeavour of Lord Acton to encourage critical scholarship among Catholics, rendered him further suspect and caused a breach with H.E. Manning, once himself a Tractarian and soon to be the new arch-

bishop of Westminster. One of Newman's articles ("On Consulting the Faithful in Matters of Doctrine") was reported to Rome on suspicion of heresy. He attempted to found a Catholic hostel at Oxford but was thwarted by the opposition of Manning.

Apologia pro Vita Sua and *Grammar of Assent*

From the sense of frustration engendered by these experiences Newman was delivered in 1864 by an unwarranted attack from Charles Kingsley upon his moral teaching. Kingsley in effect challenged him to justify the honesty of his life as an Anglican. And though he treated Kingsley more severely than some thought justified, the resulting history of his religious opinions, *Apologia pro Vita Sua* (1864), was read and approved far beyond the limits of the Roman Catholic Church and by its fairness, candour, interest, and the beauty of some passages recaptured the almost national status that he had once held. Though the *Apologia* was not liked by Manning and those who thought as he did, because it seemed to show the quasi-liberal spirit that they feared, it assured Newman's stature in the Roman Catholic Church. In 1870 he expressed opposition to a definition of papal infallibility, though himself a believer in the doctrine. In the same year he published his most important book of theology since 1845, *An Essay in Aid of a Grammar of Assent* (commonly known as *The Grammar of Assent*), which contained a further consideration of the nature of faith and an attempt to show how faith can possess certainty when it rises out of evidence that can never be more than probable. In 1879 Pope Leo XIII made him cardinal-deacon of St. George in Velabro. He died at Birmingham on Aug. 11, 1890, and is buried (with his closest friend, Ambrose St. John) at Rednal, the rest house of the Oratory.

Mind and character

Newman's portraits show a face of sensitivity and aesthetic delicacy. He was a poet—most famous are his contributions in the *Lyra Apostolica* of his Anglican days, including the hymn "Lead, kindly light," written in 1833 when he was becalmed in the strait between Sardinia and Corsica, and *The Dream of Gerontius* (1865), based upon the requiem offices and including such well-known hymns as "Praise to the holiest in the height" and "Firmly I believe and truly"—and his thought as a philosopher or theologian was never far from the poetic apprehension. He was always conscious of the limitations of prose and aware of the necessity for parable and analogy, and logical theologians sometimes found him elusive or thought him muddled.

But his was a mind of penetration and power, trained upon Aristotle, David Hume, Bishop Joseph Butler, and Richard Whately, and his superficial contempt for logic and dialectic blinded some readers into the error of thinking his mind illogical. His intellectual defect was rather that of oversubtlety; he enjoyed the niceties of argumentation, was inclined to be captivated by the twists of his own ingenuity, and had a habit of using the *reductio ad absurdum* in dangerous places. Newman's mind at its best is probably to be found in parts of the *Parochial and Plain Sermons* or the *University Sermons*, at its worst in the *Essay on Ecclesiastical Miracles* of 1843.

His sensitive nature, though it made him lovable to his few intimates, made him prickly and resentful of public criticism, and his distresses under the suspicions of his opponents, whether Anglicans defending the Reformation or ultramontanes (exponents of centralized papal power) attacking his Roman theology, weakened his confidence and prevented him from becoming the leader that he was otherwise so well equipped to be. Nevertheless, as the effective creator of the Oxford Movement he helped to transform the Church of England; and as the upholder of a theory of doctrinal development he helped Catholic theology to become more reconciled to the findings of the new critical scholarship, while in England the *Apologia* was important in helping to break down the cruder prejudices of Englishmen against Catholic priests. In both the Catholic Church and the Church of England his influence has been momentous.

MAJOR WORKS

Theological Works:

Tracts for the Times, nos. 1–3, 6, 7, 8 (with R.H. Froude), 10, 11, 15 (with Sir W. Palmer), 19–21, 31, 33, 34, 38, 41, 45, 47, 71, 73, 74 (with B. Harrison), 75, 76, 79, 82, 83, 85, 88, 90, 1833–41 (all edited by Newman); *Parochial and Plain Sermons*, 1834–42; *Lectures on the Prophetical Office of the Church*, 1837; *University Sermons*, 1843; *An Essay on the Development of Christian Doctrine*, 1845; *On Consulting the Faithful in Matters of Doctrine*, 1859; *An Essay in Aid of a Grammar of Assent*, 1870.

Miscellaneous:

Remains of the Late R.H. Froude, 1838–39 (ed. with J. Keble); *The Idea of a University*, 1852; *Apologia pro Vita Sua*, 1864 (Newman's account of his conversion to Roman Catholicism); *Essays Critical and Historical*, 1872; *Meditations and Devotions of the Late Cardinal Newman*, 1893.

Novels:

Loss and Gain, 1848 (anon.; signed and with subtitle *The Story of a Convert*, 1853); *Callista: A Sketch of the Third Century*, 1856 (anon.). (VERSE): *Lead, Kindly Light*, 1834 (anon.; reprinted in Newman's *Verses on Various Occasions*, 1868; *The Dream of Gerontius*, 1865.

BIBLIOGRAPHY. The most important works by Newman have been reprinted, a few of them often; but as Newman was in the habit of making substantial alterations when re-editing, their text history needs care: see J. RICKABY, *Index to the Works of John Henry Cardinal Newman* (1914).

Many of his letters as an Anglican were edited by ANNE MOZLEY, *Letters and Correspondence of J.H. Newman During His Life in the English Church*, new ed., 2 vol. (1898); the *Correspondence of John Henry Newman with John Keble and Others 1839–1845* was edited at the Birmingham Oratory (1917). See also *Cardinal Newman and William Froude, F.R.S.: A Correspondence*, ed. by G.H. HARPER (1933). C.S. DESSAIN and V.F BLEHL (eds.), *Letters and Diaries of John Henry Newman*, vol. 11– (1961–), is definitive. The best edition of the *Apologia* is by M.J. SVAGLIC, *Apologia pro vita sua* (1967). The official biography is W.P. WARD, *The Life of John Henry Cardinal Newman, Based on His Private Journals and Correspondence*, 2 vol. (1912, reissued 1927); less critical, but well-based on the archives and very readable is M. TREVOR, *Newman*, 2 vol. (1962–63); a short modern life is C.S. DESSAIN, *John Henry Newman* (1966).

For Newman's thought and philosophical theology, see M. NEDONCELLE, *La philosophie religieuse de John Henry Newman* (1946); for theology in general, J.H. WALGRAVE, *Newman: le développement du dogme* (1957; Eng. trans., *Newman the Theologian*, 1960); for the idea of development and its background, O. CHADWICK, *From Bossuet to Newman* (1957); for Newman and education, A.D. CULLER, *The Imperial Intellect* (1955); and F. MCGRATH, *Newman's University: Idea and Reality* (1951). Useful articles appear in *Newman Studien* (1948–).

(W.O.C.)

New Mexico

A state of the American Southwest, New Mexico is an integral part of the "Old West" of cattle drives, cowboys, and clashes with Apache Indians. In the vast flatness of its Great Plains and the rough, weather-scored peaks of its mountain ranges, it still retains much of its frontier flavour. The severe tensions and increasingly frequent confrontations between its Spanish-American, Indian, and "Anglo," or English-speaking, population are a continuing reminder of the bitter antagonisms that characterized its long history and were little resolved when it became the 47th state in the Union in 1912.

The 121,666 square miles (315,115 square kilometres) of New Mexico make it the fifth largest of American states, with only 254 square miles (658 square kilometres) of water area. Rectangular in shape except for a small panhandle in the southwestern corner, New Mexico is bounded on the north by Colorado, on the east by Oklahoma and Texas, on the south by Texas and the Mexican state of Chihuahua, and on the west by Arizona, which, from 1850 to 1863, was part of the Territory of New Mexico. At its northwestern corner it joins Arizona, Utah, and Colorado in the only four-way meeting of states in the United States.

Despite the traditionally agrarian nature of the state, augmented by successful irrigation methods, by the early 1970s, New Mexico had become highly urbanized. Seventy percent of its population of 1,016,000 were living in urban areas—nearly one-third in Albuquerque and surrounding Bernalillo County. The capital, Santa Fe, is a much smaller city, but its founding in 1610 preceded that

of Albuquerque by 96 years, and it is the oldest continuously used seat of government in North America. It was also the southwestern terminus of the Santa Fe Trail, a wagon trail that was a major commercial and migration route from Missouri to the Southwest from 1821 to 1880, when the last section of the railroad was completed. (For information on related topics, see the articles UNITED STATES; UNITED STATES, HISTORY OF THE; NORTH AMERICA; ROCKY MOUNTAINS; and GREAT PLAINS.)

THE HISTORY OF NEW MEXICO

New Mexico was peopled by various Indian cultures, who farmed and hunted on the land for at least 10,000 years before white explorers appeared. The more peaceful agriculturists, including the later groups, whose pueblo ruins dot the state, held the longer sway and had well-developed irrigation systems by the time the more aggressive and nomadic Navajo and Apache arrived from the north, probably in the 15th century.

Spanish and Mexican rule. Reports of the fabled seven cities of gold brought the first white explorers into New Mexico in 1540, led by the Spanish adventurer Francisco Coronado. The journey fruitless, they returned to Mexico. After several decades of desultory exploration by soldiers and friars, Juan de Oñate was given contracts for colonization in 1595 and made the first permanent white settlements during the following years, founding Santa Fe in 1610.

Colonial period

For the next century, missionary work predominated, but the attempts to eradicate Indian religion and culture brought on an uprising and massacre in 1680 that cleared out the white man for some years. By 1700, Spanish arms had reasserted themselves, and for the next century there was considerable settlement. Albuquerque, founded in 1706, was the focal point in the south, and the older Santa Fe was that of the north. When New Mexico became a part of the Republic of Mexico, newly founded in 1821, it already had begun the trade with the United States over the Santa Fe Trail that would lead to still another allegiance 25 years later.

Territory and state. During the Mexican War, which began in 1846, New Mexico was taken by the Army of the West under Gen. Stephen Kearny. All residents were granted amnesty and citizenship in return for an oath of allegiance to the United States. The Territory of New Mexico was established by Congress in 1850. During the Civil War an invading Confederate force was driven out by the Colorado Volunteers.

The Navajo tribes were quelled and, in 1868, given a large reservation, but the Apaches, settled on two reservations in 1880, continued their struggles until 1886. The burgeoning cattle industry was the main development of these decades, and the territory often was bloodied by battles between cattlemen and sheepmen, large landowners and homesteaders. The legendary Billy the Kid and his lawman-nemesis, Pat Garrett, were products of this struggle. The Apache leaders Geronimo, Cochise, and Victorio, though mainly active in Arizona, made forays into southwestern New Mexico. The Atlantic and Pacific Railroad (later the Atchison, Topeka, and Santa Fe), which reached Albuquerque in 1880, brought new immigration, and farming grew rapidly with the development of new irrigation methods and resources.

Following admission as a state on January 6, 1912, New Mexico retained its agricultural bases until World War II, when atomic research opened a new era in the state's life. The most famous scientific installation is at Los Alamos, centre of the project that created the first atomic bomb in 1945.

The gradual development of mineral resources also helped the economy and income of the state.

THE NATURAL AND HUMAN LANDSCAPE

The natural environment. New Mexico has some of the flattest land in the world and also some of the most rugged mountains. Some portions are blessed with pine forests, rich meadows, and fish-laden mountain streams, while other areas are devoid of streams, and even cacti struggle to survive. The eastern third of the state is an extension of the Great Plains. The central third is the southern extension of the Rocky Mountains, with the ranges interspersed by valleys and running in a north–south direction. The western third is high-plateau country, but it also contains many short mountain ranges and plains.

Surface features

Mountains. Average elevation ranges from 5,000 to 8,000 feet (1,500 to 2,500 metres) above sea level in the northwest to less than 4,000 feet (1,200 metres) in the southeast, with 85 percent of the state in excess of 4,000 feet (1,200 metres). The highest mountain peaks, Wheeler Peak (13,161 feet [4,011 metres]) and Truchas Peak (13,102 feet [3,993 metres]), are in the Sangre de Cristo range in the north central part of the state. The numerous valleys between the ranges are indispensable to agriculture and grazing. Unique volcanic formations abound as reminders of past lava flows. The caverns near Carlsbad are among the most spectacular natural formations in the world.

Rivers. Five major river systems—the Rio Grande, Pecos, Canadian, San Juan, and Gila—drain the state. The Rio Grande, which has played an influential role in New Mexico's history, virtually bisects the state. Agriculture in its floodplain has been significant from prehistoric times, and, before American settlement, white men lived exclusively in its valleys and those of its tributaries. The Pecos, east of the Rio Grande and approximately parallel to it, was also a popular route for explorers. The Canadian River, rising in the Sangre de Cristo Range and flowing due east across the arid plains, was a useful avenue for explorers despite its deep canyons. The San Juan and Gila rivers lie west of the Continental Divide, in the northwest and southwest respectively. All but the Gila, which is not dammed in New Mexico, provide water for irrigation, recreation, and flood control.

Climate. Though New Mexico's mean annual temperature is 53° F (12° C), extremes range from 110° F (43° C) to −29° F (−2° C). Variations are caused more by altitude than latitude, temperatures rising or falling by 5° F with every 1,000 feet of elevation. Nighttime temperatures tend to fall sharply. The average annual rainfall is 15 inches (380 millimetres), though precipitation tends to increase with elevation. About 40 inches (1,000 millimetres) of rain falls in the higher mountains, whereas lower areas may get no more than eight to ten (200 to 250 millimetres). Generally, precipitation is greatest in the eastern third of the state, least in the western.

Vegetation. New Mexico has six vegetation zones, which are also determined mainly by altitude. The Lower Sonoran zone, in southern sections of the Rio Grande and Pecos valleys and in the southwestern corner, usually occurs at altitudes below 4,500 feet. It embraces nearly 20,000 square miles of the best grazing area and irrigated farmland. The Upper Sonoran, comprising about three-fourths of the state and including most of the plains, foothills, and valleys above 4,500 feet, is a zone of prairie grasses, low piñon pines, and juniper shrubs. At higher altitudes, better stands reflect the more abundant rainfall. The Transition zone, covering some 19,000 square miles, is identified chiefly with the ponderosa pine. The Canadian zone, embracing 4,000 square miles at elevations of 8,500 to 9,000 feet, contains blue spruce and Douglas fir. The Hudsonian and Arctic-Alpine zones, above 9,500 feet, are too small in area and too sparsely covered to be of great importance.

Zones of vegetation

Animal life. The diversity of natural vegetation and elevation affect the wildlife, and the inaccessibility of much of New Mexico has helped preserve its abundance. Mule deer, brown bear, bighorn sheep, mink, muskrat beaver, fox, mountain lion, and bobcat live in the mountain and forest areas above 7,000 feet, while at lower elevations antelope are still found, along with coyote and jackrabbit. Barbary sheep from North Africa were introduced into several mountain areas. Many species of trout are common in the mountain streams, and warm-water fish abound in lower streams. Approximately 300 species of birds can be found at almost any time, including various game birds. The rattlesnake and the black widow

spider are most feared by man among the many lower forms.

Patterns of human settlement. The first Spanish settlements were in the central Rio Grande Valley and its tributaries. The Spanish-speaking inhabitants are still concentrated in the north central portion of the state. The eastern third of the state is frequently referred to as "Little Texas" or the "East Side." It is an extension of the Great Plains of western Texas and was originally settled as a cattle frontier expanding westward from Texas after the Civil War. It has continued to attract Protestant Anglos from Texas as ranchers, farmers, or oil-field workers, and they often have been at odds with the Spanish-American Catholics of Albuquerque and Santa Fe. The southwestern corner of the state, settled by Anglo miners after the coming of the railroads, also has little in common with the central area. The northwest corner received Mormon settlers from Colorado, but the greatest growth of this area resulted from oil and natural-gas discoveries after World War II.

Early settlers remained along streams because of the scarcity of water elsewhere. In a typical community, adobe houses opened onto a plaza from which four streets ran outward, and the entire enclave was enclosed by a wall for defense. Nearby were small agricultural plots and orchards owned by individuals. Just beyond, in a great circle, was the ejido, land for communal grazing, recreation, or firewood. Despite fear of Indian attack, individual ranches away from settlements often were established. At the time of the American conquest, New Mexico was a self-sufficient agrarian community, with most people residing in small villages.

Spanish villages

After the Civil War, vast cattle ranches appeared on the East Side, their size limited only by the availability of water. The coming of the railroad in 1879 brought several waves of Anglo farmers, but frequent droughts ruined many who tried to till the soil as they had in their more humid homelands. Dry farming—tilling that uses drought-resistant crops or otherwise conserves soil moisture—saved many who remained, but today irrigated farming is the most important form of agriculture.

THE PEOPLE OF NEW MEXICO

Ethnic composition. The people of New Mexico are primarily Anglos, Spanish-Americans, or Indians, and there are few Negroes or Orientals. The original Spanish settlers intermarried with the Indians, and their descendants are designated as Spanish-Americans or Hispanos rather than Mexican-Americans, as elsewhere in the Southwest. The numerical superiority of Spanish-Americans lasted until the 1940s, but by 1970 persons with Spanish surnames made up only slightly more than 25 percent of the population. After World War II an influx of Anglos contributed to a widespread desertion of small agricultural villages by their traditional Spanish-speaking residents, who moved to urban centres or to California. Many such villages became ghost towns.

The Indian population has grown from 34,510 in 1940 to about 73,000 in 1970. A large Navajo reservation extends over the northwest corner into Arizona, and nearby Gallup is famous as "the Indian capital of America." There are also reservations for the Utes, Jicarilla, and Mescalero Apaches, and in 1968 more than 20,000 Pueblo Indians lived on 19 scattered land grants. The Indians preserve many of their ancient ways, tending flocks of sheep and producing numerous handicraft items. In recent years, however, dissatisfaction with their low income, inadequate housing, poor health standards, and lack of educational opportunity has led to a growing militancy and an increasing exodus from their reservations or pueblos to urban centres.

Indian population

Demography. Traditionally rural New Mexico has joined the national trend toward urbanization, and since 1920 its rate of population growth has exceeded that of the United States as a whole. A growth rate of nearly 40 percent during the 1950s was in large part the result of a proliferation of federal expenditures in the state for defense and research and of the discovery of oil and other minerals. The reduction of federal outlays and the leveling off of mining booms, however, caused the rate to plunge to only 7 percent in the 1960s, and in both decades half of the state's counties declined in population. Urbanization has involved a number of factors: the movement of Spanish-Americans away from their rural homes, the consolidation of farms and the increasing inclination of many farmers to abandon their isolation for the larger towns and commute to their fields and flocks.

THE STATE'S ECONOMY

New Mexico's economy is similar to that of the developing nations of the world in that it is largely at the mercy of forces over which it has little control. Relying heavily on the export of raw materials and on federal expenditures for programs of no certain permanence, it is subject to shifting demands from the outside. Over-all, government spending accounts for nearly one-third of the state's economy. New Mexico is a comparatively poor state, in 1970 ranking 44th nationally in per capita income. During the 1960s the income of the average New Mexican fell from about 85 percent to 79 percent of the national figure.

Agriculture. Though gross farm income continues to rise slowly, in 1970 it accounted for only 7 percent of the state's income. Under Spain and Mexico, Nueva Méjico was self-sufficient, growing beans, corn, cotton, and squash on the alluvial plain of the Rio Grande. Sheep prospered in the arid land and remained important until the 20th century. The Anglos brought cattle raising from Texas, and the sale of beef now accounts for more than one-half of marketing receipts from agricultural products. Cotton is the leading cash crop, with hay second. Wheat and sorghum grains are raised on the dry farms of the eastern part of the state, but the irregularity of rainfall makes this type of agriculture hazardous. Half of the total cropland of 1,117,000 acres (452,000 hectares) is under irrigation, and it is such lands that furnish the overwhelming share of the crop dollar.

Major crops

Mining. The mining industry, contributing 6 percent of the state's economy, brought many settlers and attracted outside capital in territorial days. Gold and silver mining began in the 19th century, reached its peak in 1915, and has declined since. Copper mining continues in importance, and coal mining, which went into decline with increased use of other fuels, expanded in the 1960s as the result of improved technology. New Mexico produces some 85 percent of the nation's potash and since discovery of uranium deposits in 1950, has led the nation in uranium production. Iron, lead, zinc, manganese, and molybdenum are also mined, but oil and natural gas account for some 65 percent of the state's mineral income. Production of the latter is mainly in the southeast corner and in the San Juan Basin in the northwest.

Manufacturing. Originally limited to the production of consumer goods, manufacturing has increased rapidly since World War II and accounts for 7 percent of the economy. Food processing, petroleum refining, smelting, construction materials, and railroad maintenance are leading manufacturing activities. Basic atomic research is carried on at the Los Alamos Scientific Laboratory, with testing taking place at Sandia Military Base, in Albuquerque, or at the White Sands Missile Range, near Alamogordo. An offshoot of this is private manufacturing of ordnance, electronics, precision instruments, and the like. Because of the limited development of industry, unionism has never been important, and it is confined largely to the mining, smelting, and petroleum industries.

Tourism and trade. In 1968 wholesale and retail trade contributed 21 percent and services 18 percent of the state's economy. The figure for services chiefly reflects the increasingly important development of recreation and tourism. The distinctive Indian and Hispanic cultures continue to draw countless visitors. State and national parks and monuments, historic sites, hunting and fishing, skiing, and Indian ceremonials are important attractions.

Transportation. Geographic isolation was a basic cause of New Mexico's slow economic development. In the Spanish and Mexican periods, it was often a six-month trip between Mexico City and Santa Fe. The Santa

Fe Trail route was much shorter, and American consumer goods helped prepare the way for conquest. This isolation was ended when the railroads reached Albuquerque and Santa Fe in 1880, and today an extensive rail network unites the state. New Mexico has highways linking major population centres, three of them part of the federal interstate system. Mountainous terrain has made road construction expensive, and, although secondary road building has lagged, it is adequate. Air transportation provides a vital link with the distant centres of the nation.

ADMINISTRATION AND SOCIAL CONDITIONS

Governmental structure. In most instances the state's constitution can be amended by a majority vote of the legislature and by a majority vote of the electorate. A public referendum on major issues is permitted, but public initiative on legislative matters is not. Nomination to office is by closed primary.

Executive. The governor has the usual powers of pardon, reprieve, and veto, but he has more authority than executives of most states. Since he appoints most of the state boards, departments, agencies, and commissions, he is the virtual master of patronage and the political organization. Like the lieutenant governor and other executive officials, he is elected for a two-year term and can serve two consecutive terms but is ineligible for all state elective offices for two years thereafter.

Legislature and judiciary. A legislature composed of 42 members of the Senate, elected to four-year terms, and 70 members of the House, elected to two-year terms, meets annually for a 60-day session. Heading the judiciary are five Supreme Court justices elected individually for eight years, with overlapping terms. Judges of the 11 judicial districts are elected for six years and serve ex officio as judges of juvenile courts.

Local government and politics. Each of New Mexico's 32 counties is administered by three commissioners elected for two years. Other county elective officers are assessor, clerk, sheriff, surveyor, treasurer, and probate judge. In the territorial era, citizens usually favoured Republicans, but since statehood Democrats have tended to dominate. The state voted for the winner in every presidential election from admission through 1972.

Education. A public school system was established in 1891, and, since statehood, educational improvement has been phenomenal. In 1968 the educational attainment of the Anglo adult population averaged 12.1 years—compared however, to only 8.1 for Spanish-Americans. Per-pupil expenditure in 1970 was more than $600 and average teaching salaries nearly $7,800. Most of the improvement in education has been in the urban centres, however, and many rural and small-town schools remain substandard. As Hispanos are dominant in these areas, this has meant a poorer education for their children. Legalized segregation for the Hispano minorities on the East side ended in the 1950s, but *de facto* segregation, primarily on the elementary level, remains.

Colleges and universities

Higher education is led by the University of New Mexico, in Albuquerque, established in 1888. Other state-supported institutions include New Mexico State University (1889), in University Park; Eastern New Mexico University (1934), in Portales; New Mexico Highlands University (1893), in Las Vegas; Western New Mexico University (1893), in Silver City; and New Mexico Institute of Mining and Technology (1889), in Socorro. Northern New Mexico State School at El Rito, originally established in 1909 to train Spanish-speaking teachers, is today a resident high school. In addition, the state universities have established branch campuses, while some cities have organized junior colleges. There are several private colleges as well.

Health and welfare. Other state institutions include a penitentiary, a hospital and training school, an industrial school for boys, a girl's welfare home, the Carrie Tingley Crippled Children's Hospital (1937), a school for the visually handicapped, a school for the deaf, and a miner's hospital. The state Department of Health, created in 1919, administers an extensive social-service program, often in collaboration with federal agencies. Medical services in rural areas are rarely adequate. State expenditures for health were $5.52 per capita in 1969.

Intergroup relations. The greatest single problem that New Mexico faces is the cultural clash between Spanish-Americans, once dominant in the state, and the now-dominant Anglo. Distrust and sometimes hostility, prejudice, and discrimination exist between the two groups. Efforts in 1967 of an Hispanic group, the Alianza de los Pueblos Libres (Alliance of Free City-States), to regain lost land was a result of the frustration growing out of their declining social, educational, and economic status, for which demands for land grants became a symbol. The Hispano has equality before the law, but in all practical matters he is a second-class citizen in his homeland. His income averages little more than half that of the Anglo, his education is much less, and he is under-represented in the professions. The percentage of Spanish-Americans living in substandard housing is not only above that of the Anglo but is also above that of the few blacks in the state. Desertion of the native villages for urban centres creates problems of social and economic adjustment that will take years to resolve.

Status of Spanish-Americans

CULTURAL LIFE AND INSTITUTIONS

The arts. Writers and architects have been influenced by New Mexico's Indian and Spanish heritages and the Anglo impact upon the two. The appearance of the cowboy and the miner and the conflicts of the frontier territory in the 19th century also have been dominant cultural themes. Painters have been concerned especially with the unique landscape, since no other area of the United States presents such a variety of scenery or so many modes of life side by side. Attracting artists in all fields from many parts of the nation and of the world, Taos was the first to have an important art community, but it is now rivalled by Santa Fe and Albuquerque. The state institutions of higher education, through their libraries and their departments of art, music, dance, and theatre, have played a key role in the dissemination of cultural knowledge. This has been accomplished directly and through the training of public-school teachers. The success of the Santa Fe Opera Association, organized in 1956, reflects the growth of musical appreciation. Performing in an outdoor theatre in the Sangre de Cristo Mountains near the city, the company has presented a repertory that has won it worldwide acclaim.

The historical atmosphere of New Mexico and its fusion of three cultures is represented by its unique architecture. Indian pueblo buildings were modified by the Spanish when they built Santa Fe and many of these original structures have been restored. The statehouse, most public buildings, and many private ones have recently been constructed in the modified Spanish mission style.

Folk culture. Local Indians produce pottery of high quality and authentic beauty. Each village has its own design to identify the work of its people. Navaho blankets are famous the world over. Today, many Indians make buttons, beads, pins, rings, necklaces, earrings, and belts, mainly for sale to the growing number of tourists. The United States Indian Arts and Crafts Board has attempted to preserve the authenticity of Indian jewelry by establishing standards in handworked silver. Individual pueblos preserve native dances by performing at numerous fiestas, the most important being the Inter-Tribal Ceremonial, which draws thousands of visitors to Gallup every summer.

Indian arts

Spanish folk art has been preserved largely by the Penitentes, a religious group within the Roman Catholic Church. In rural areas, medieval Spanish music and art have been preserved against modern influences.

Museums. The Museum of New Mexico, housed in the Palace of the Governors in Santa Fe, helps preserve archaeological sites, mementos, and folk arts of the past, as well as the state archives. The Museum of Navaho Ceremonial Art is in Santa Fe, and in Taos is the Kit Carson Home and Museum.

Communications. In 1971, New Mexico had 19 daily newspapers, 70 radio stations, and seven commercial and

educational television stations. The *New Mexico Historical Review*, a quarterly established in 1926, publishes scholarly articles on the state's history, while the University of New Mexico Press publishes books in many fields related to the Southwest.

Prospects. New Mexico's major problems lie, first, in its intergroup relations and, second, in its slight and uncertain economic base. Certain unique advantages are available, however, to meet these problems. The scientific talent at Los Alamos and elsewhere in the state could play a vital role in expanding peacetime uses of nuclear energy. The state's natural beauty and climate can be utilized better to expand the already important tourist industry. The success in luring movie producers to take advantage of the unequalled landscape and the ability of cities like Roswell to bring in new industries to offset the loss of federal installations suggest that existing economic problems can be solved. The social problem can be eased greatly by providing more economic and educational opportunity for the Hispano and Indian minorities, but this must be done in an atmosphere of trust and goodwill on all sides.

BIBLIOGRAPHY. VERNON BAILEY, *Life Zones and Crop Zones of New Mexico* (1913), the basic reference in this area, and *Mammals of New Mexico* (1931), the standard source; H.H. BANCROFT, *Arizona and New Mexico, 1530-1888* (1888, facsimile reprint, 1962), though dated in many areas, still a basic source, especially for its bibliography; W.A. BECK, *New Mexico: A History of Four Centuries* (1962), the only modern single-volume text available, and with YNEZ D. HAASE, *Historical Atlas of New Mexico* (1969), maps and text dealing with all phases of history and geography; T.C. DONNELLY, *The Government of New Mexico* (1947), a basic text long used in the teaching of the state's government; NANCIE L. GONZALEZ, *The Spanish Americans of New Mexico: A Heritage of Pride*, rev. and enl. ed. (1969), an up-to-date sociological study containing valuable statistics; B.L. GORDON *et al.*, *Regions of New Mexico* (1961), a brief geographical study; L. GREBLER, J.W. MOORE, and R.C. GUZMAN, *The Mexican-American People* (1970), a massive study suffering from social science jargon; J.E. HOLMES, *Politics in New Mexico* (1967), a complete analysis of the political environment since 1900, with good treatment of the Hispanic influence and regional voting patterns; R.W. LARSON, *New Mexico's Quest for Statehood, 1846-1912* (1968), the definitive study of this subject; WRITERS PROGRAM, *New Mexico: A Guide to the Colorful State*, rev. ed. by JOSEPH MILLER (1954), one of the superb American guide series; F.D. REEVE, *History of New Mexico*, 3 vol. (1961), the most complete work, with an excellent volume of biographies of the state's luminaries; A.M. SMITH, *New Mexico Indians* (1966), an analysis of the economic, educational, and social problems of the state's Indians, with a summary by tribes.

(W.A.B.)

New Orleans

Unquestionably one of the most distinctive cities of the New World, New Orleans was established at great cost in an environment of conflict. Its strategic position, commanding the mouth of the great Mississippi–Missouri river system, which drains the rich interior of North America, made it a pawn in the struggles of Europeans for the control of North America. As a result, its peoples evolved a unique culture and society, blending many heritages. Its citizens of African descent—who have represented from a fourth to a half of the population of the city since its inception—have provided a special contribution in making New Orleans the birthplace of Dixieland jazz. New Orleans is a city of paradox and contrast: while it shares the urban problems afflicting other North American cities, it has nevertheless preserved the exuberant and uninhibited spirit exemplified by its carnival season, culminating in the annual Mardi Gras, when more than 1,000,000 persons throng the streets. The city also has a solid economic base: it is the largest city in Louisiana, the second port of the United States in value of foreign commerce, a major tourist resort, and a medical, industrial, and educational centre.

Historical and cultural heritage

The city and Orleans Parish (county) are coextensive, covering an area of almost exactly 200 square miles (518 square kilometres). The boundaries are formed by the Mississippi River and Jefferson Parish to the west and Lake Pontchartrain to the north. The latter is connected by the Rigolets Channel to Lake Borgne on the east, and the southern boundary is made up of St. Bernard Parish and, again, the Mississippi River. The city is divided by the Mississippi, with the principal settlement on the east bank. The west bank, known as Algiers, has grown rapidly in recent years. It is connected to eastern New Orleans by the Greater New Orleans Bridge. The bridge, completed in 1958, has proved to be an increasingly painful bottleneck to the city's traffic.

The early city was located on the east bank along a sharp bend in the river, from which its popular name "Crescent City" is derived. The modern metropolis has spread far beyond this original location. Because its saucer-shaped terrain lies as low as five feet below sea level and experiences an average rainfall of 57.45 inches, a levee, or embankment, system and proper drainage have always been of prime importance to the city. New Orleans has a moderate climate, the average daily temperature from October through March being 60° F (16° C) and from April through September, 77° F (25° C). Freezing weather is rare, and the temperature goes above 95° F (35° C) only about six days a year. (For further information see LOUISIANA; UNITED STATES, HISTORY OF THE; and MISSISSIPPI RIVER).

THE GROWTH OF THE METROPOLIS

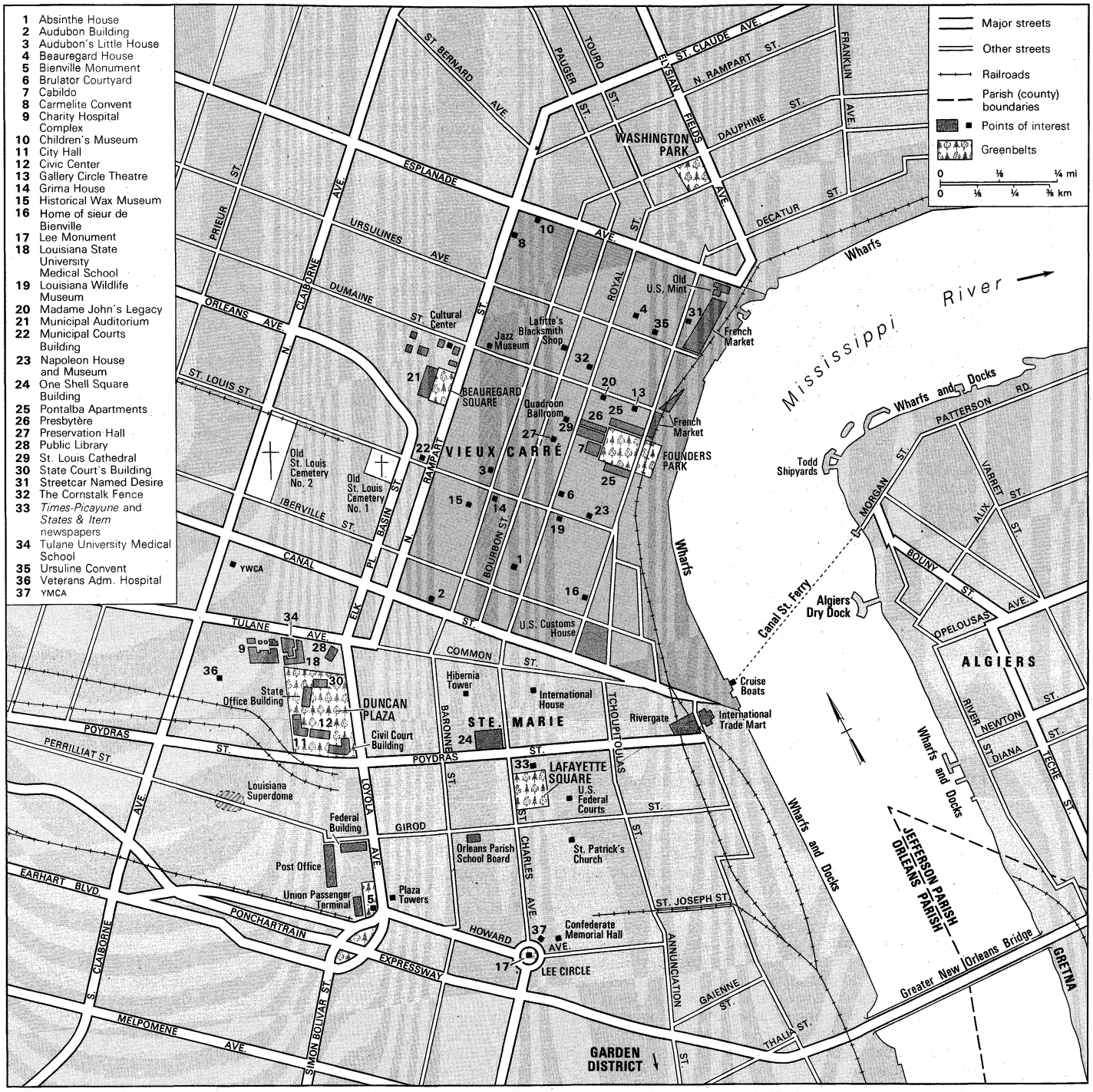
Foundation and early settlement. The decision to found New Orleans was made in Paris in 1717 by John Law's Company of the West, which had taken control of Louisiana that year. The colony's new proprietors envisioned New Orleans (named for the French regent, the Duc d'Orléans) as a "port of deposit," or transshipment centre, for future trade from upriver in the Mississippi Valley. Jean-Baptiste le Moyne, sieur de Bienville, the man who suggested the site, was entrusted with the actual foundation of the city. Clearing of underbrush for the new city probably began in March 1718. The engineers charged with this task met with problems arising from uncooperative convict labour, a shortage of supplies, two severe hurricanes (in 1721 and 1722), and the unpleasant physical conditions of swamps and mosquitoes as they set up the first crude dwellings covered with bark and reeds. An engineer, Adrien de Pauger, drafted the first plan for the town, encompassing the section known today as the Vieux Carré and consisting of 66 squares forming a parallelogram.

Hostility of the environment

The first residents were a colourful mixture of Canadian backwoodsmen, company craftsmen and troops, convicts, slaves, women of uncertain virtue, and indigents. In a census taken in November 1721, New Orleans had a population of 470 persons: 277 whites and 172 Negro and 21 Indian slaves. In 1722 New Orleans was designated the capital of Louisiana, and in 1731 the city returned to the control of the French crown. More respectable colonists began to arrive, but growth continued to be precarious. The main economic staples grown in the vicinity of New Orleans were tobacco and indigo for export and rice and vegetables for local consumption. Naval stores was also an export. French ships were, however, reluctant to call at New Orleans to pick up such cargo: its value did not match its bulk.

In 1762 France, ready to part with its unprofitable port, secretly agreed to cede Louisiana to Spain, and by the Treaty of Paris (1763) Spain received New Orleans and the Louisiana Territory west of the Mississippi. After a brief rebellion—which was sternly suppressed—the inhabitants of New Orleans enjoyed peace and a growing prosperity under Spanish law, while trade arose with the British colonies in spite of Spanish restrictions. At the same time English-speaking colonists were moving west to settle along the tributaries of the Mississippi. In the decade of the American Revolution, these "Kaintucks," as they were called, began floating their cargoes downriver to New Orleans: several times Spanish officials suspended the right of deposit of American goods at New Orleans in response to the boisterous conduct of American frontiersmen along the city's upper levee.

In 1800 Louisiana was secretly returned to Napoleon's



Central New Orleans.

France, and by 1803 the French emperor had negotiated for its sale to the United States. The ceremonies transferring Louisiana to France and later to the United States took place in New Orleans' Cabildo and main square, the Place d'Armes (now Jackson Square) in the winter of 1803.

The Louisiana Purchase

The early 19th century. New Orleans' population in 1803 was aproximately 8,000—4,000 whites and 2,700 enslaved and about 1,300 free "persons of colour." Its prosperity was reflected in its 1803 exports, which had a value approaching $2,000,000, and were bound mainly for American ports. In 1805, when it was incorporated as a municipality, New Orleans took on an identity separate from that of Louisiana's territorial government. As the city expanded out of its original limits, one of the first new tracts of land to be added was the Faubourg Ste. Marie, a suburb lying on the uptown side of the Vieux Carré and separated from it by a broad "commons" (today Canal Street, New Orleans' main street). It became the "American section" in the early 19th century and the hub of most business activities. Other *faubourgs* (lit., outskirts, or suburbs) were laid out above and below the two nuclear settlements and across the river and were finally absorbed into the city by the 1870s.

During the War of 1812, New Orleans was threatened by a British invasion force, which approached the city from the Gulf of Mexico. Gen. Andrew Jackson, with his army of frontiersmen and local volunteers, won a smashing victory on January 8, 1815, saving the city, though, unknown to him, the war already had been concluded.

The era of cotton

The next 40 years constituted the golden age of New Orleans as a great cotton port. The first steamboat to reach the city, in 1812, was appropriately called the "New Orleans." Mississippi River steamboats increased to 400

by 1840, and local commerce skyrocketed in value, reaching $54,000,000 by 1835. By 1840 the city was rated the fourth port in the world: after the 1840s canals and railroads diverted produce eastward to New York City.

German and Irish immigrants arrived in New Orleans in large numbers in the 1840s. By 1850 the city's total population had swelled to 116,375. New Orleans, however, had not learned to cope with the health hazards of its mushrooming growth; drinking water came from the river or cisterns; no sewerage system existed; drainage was deficient; and flooding was common after heavy rains. The results were sporadic outbreaks of cholera and yellow fever, the worst of which was the yellow fever epidemic of 1853, accounting for more than 8,000 deaths.

The Civil War and its aftermath. During the Civil War, the strategic location of the city was inadequately appreciated by the Confederate military. The Union fleet of Admiral David Farragut was able to capture New Orleans in April 1862. The city was placed under the military command of Gen. Benjamin Butler, and city officials were removed from office. Although Butler was replaced as commander by Nathaniel Banks by the end of 1862, his brief régime became infamous in local history for his roughshod handling of the population.

During the period of Reconstruction, 1865–77, racial tensions ran high. "Scalawags" (white Southerners who cooperated with Republican forces) and "carpetbaggers" (Northerners accused of exploiting the situation for personal gain) cooperated to gain political control of the city and state, with the support of black voters. By 1872 amnesty had been granted to the ex-Confederates, and the municipal government returned to traditional white control, although state government, and the city police force, remained under Radical Republican control until 1877. In the 1880s the debt of $24,000,000, incurred under carpetbag régimes, increased steadily with each subsequent administration. This municipal debt had to be paid before the city could undertake any new bond issues for sorely needed municipal improvements.

In the last 20 years of the 19th century, therefore, New Orleans made limited, though steady, progress. Between 1840 and 1900, it had dropped from third to 12th place in national rank, although its population had increased to 287,104.

Yellow fever was sharply curtailed after the Civil War through fumigation of ships at a quarantine station on the lower Mississippi and was finally eradicated by 1906. By the early 20th century, the river steamboats, unable to compete with railroads in bulk carried or in rates, disappeared; the Port of New Orleans, attracting less railroad freight than Eastern ports, reached a low ebb shortly before World War I. With the development of towboats and barges large enough to hold almost an entire trainload of cargo, however, cheap river transportation revived the river commerce, and the port again surged forward, becoming the second port in the nation after World War II. Substantial progress, at least in physical improvements, came to the city in the 1950s. During the administration of Mayor DeLesseps S. Morrison, a vast railroad consolidation program was achieved, and a $15,000,000 railroad terminal constructed. Streets were widened, railroad ground crossings were spanned with overpasses, and a $20,000,000 civic center built, which includes the 11-story City Hall.

20th-century growth

THE CONTEMPORARY CITY

The people. New Orleans' population in 1970 was almost 600,000, a 5.4 percent decline from 1960. Of this figure, 323,000 were white; 267,000, black; and 2,700 persons of other races. Whites accounted for scarcely more than half of the total, whereas in 1960 they had made up almost two-thirds. In contrast to the population decline in Orleans Parish, the adjacent parishes—which, together with Orleans, compose one of the national Standard Metropolitan Statistical Areas—showed big increases. The parishes of Jefferson, St. Bernard, and St. Tammany—the latter across Lake Pontchartrain from New Orleans—all showed an increase of around 60 percent. Since the black population in these three parishes is quite small, these figures indicate a white move to the suburbs, a trend common to major United States cities. The total population of the four parishes in the Greater New Orleans Standard Metropolitan Statistical Area was just over 1,000,000.

The shift in population to the suburbs was less motivated by racial tension (although this may play a part) than by desires for better and more modern living facilities. The fact that a large segment of the black population resides in declining neighbourhoods (some segregated, some integrated) has spurred both black and integrated political, social, and religious organizations to work either independently or with city and federal agencies in projects to improve the quality of life for lower-income citizens. The additional fact that New Orleans has an upper-class and a middle-class black population has been a significant factor in such projects.

The city's economy. *The port and its facilities.* New Orleans has always been primarily a commercial centre, with manufacturing playing a secondary role in its economic life. The busy harbour, besides adding to the city's cosmopolitan atmosphere, is the foundation of the metropolitan economy, influencing many aspects of urban life.

The era of the modern Port of New Orleans began in 1879 with the construction of jetties in South Pass, one of three passes that flow from the river into the Gulf. Sandbars had formed at intervals in these passes and had hindered ships entering the river since the city's founding. The jetties narrowed South Pass, forcing the river to cut a deeper channel to a depth of 30 feet. Later, a second channel, Southwest Pass, was deepened to 40 feet by installing jetties; it is now the main pass used by seagoing vessels entering and leaving the river. The distance from New Orleans to the Gulf is about 110 miles.

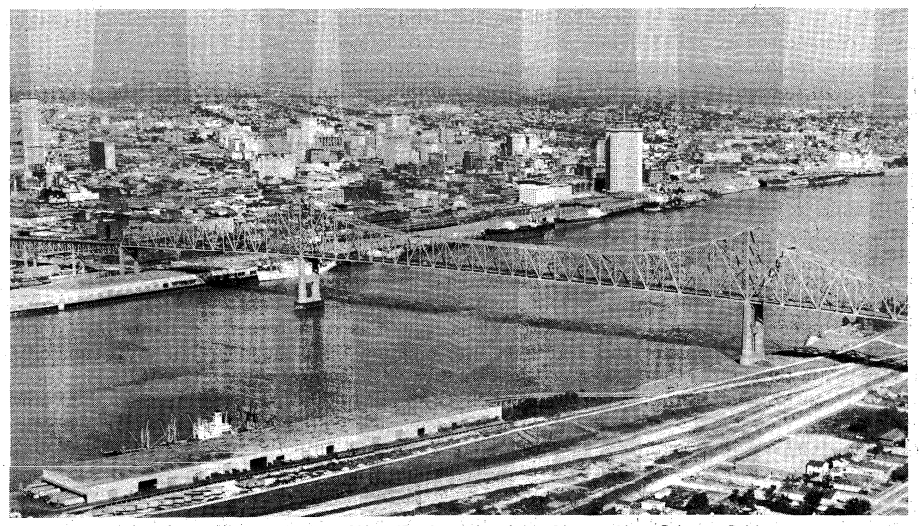
Another major step forward for the port was taken in 1896, when the state legislature removed wharf facilities from the control of private contractors and created the Board of Commissioners of the Port of New Orleans (the Dock Board), a body charged with administering the public wharves. In 1908 the Dock Board was authorized to issue negotiable bonds for the improvement of port facilities. The projects subsequently accomplished included the rebuilding and expansion of public wharves and the construction (in partnership with the Board of Levee Commissioners of the Orleans Levee District) of the five-and-a-half mile Industrial Canal, which links the river to Lake Pontchartrain, the Intracoastal Waterway, and the Mississippi River–Gulf Outlet. In 1963 the Mississippi River–Gulf Outlet, a ship channel shortening the passage to the Gulf by 40 miles, was opened to maritime traffic. Tonnage handled at this terminal grew at an average rate of 80 percent annually.

Legislative control of the port

The Dock Board has formulated a 30-year plan (1970-2000) called Centroport U.S.A. by which much of the port's activities will be switched from the Mississippi River to wharves and industrial complexes along the Gulf Outlet. The river frontage thus retired from maritime use will be diverted to such projects as high-rise apartments and public recreation areas.

New Orleans is a major grain port both in the United States and worldwide; other exports include raw and processed agricultural products, fabricated metals, chemicals, textiles, oils, petroleum and petroleum products, tobacco, and paper board. New Orleans' imports and exports have a high value, about twice the national average, which helps to account for its national ranking.

Manufacturing. The greater New Orleans' four-parish Standard Metropolitan Statistical Area is one of the fastest growing industrial areas in the nation. A concentration of petrochemical plants, representing a capital investment of about $1,000,000,000, has sprung up along the Mississippi River above New Orleans. The National Aeronautics and Space Administration established the Michoud Assembly Facility in New Orleans in 1961 to produce the giant Saturn rocket boosters used in flights to the moon. The major goods manufactured in the greater New Orleans area are food products, clothing and related items, stone, clay and glass articles, primary metal and fabricated metal items, and transportation equipment.



New Orleans from across the Mississippi River. To the right of the Greater New Orleans Bridge is the International Trade Mart.
Charles E. Rotkin—P.F.I.

In recent years, petrochemical industries along the Mississippi above New Orleans and offshore oil rigs in the Gulf have become serious polluters through oil-rig fires and oil slicks and discharges of mercury, arsenic, and lead, threatening the city's drinking water, ruining the taste of river fish, and endangering the ecology of the Gulf. Despite federal actions against the offending industries, much remains to be done.

Transportation. The transportation facilities of New Orleans include three airports. The New Orleans International Airport accounts for well over 300 commercial-flight arrivals and departures daily. The New Orleans Airport on the lakefront is devoted to private and corporate plane use, while the U.S. Naval Air Station serves air reserve units of the various armed services. The centre of railroad activity is the Union Passenger Terminal—eight railroads operate out of New Orleans with direct connections to cities in 28 states. Two passenger bus lines also use the Union Passenger Terminal, whereas more than 60 truck lines and barge lines operate to and from the city. Regular express sailings by more than 100 scheduled steamship lines also offer passenger- and cargo-carrying service. The major access bridges serving the Greater New Orleans area, in addition to the Greater New Orleans Bridge, are the Huey P. Long Bridge, which crosses the river above the city, and the Pontchartrain Causeway, a twin span that is the world's longest bridge, stretching over 23 miles.

Politics and government. *The political framework.* Both the political life and the municipal government of 20th-century New Orleans have been dominated by factions of the Democratic Party. The question of state interference in city affairs versus home rule is one of the major issues. In 1954, New Orleans finally received a strong home rule charter, which substituted a mayor-council form of government for the mayor-commission form that had existed since 1912.

Municipal management

In addition to the mayor and seven councilmen—five elected from districts and two at large—who serve four-year terms, the position of chief administrative officer to the mayor was created. The mayor is the top administrator over the 13 municipal departments and oversees the affairs of various commissions and boards. A chief administrative officer is charged with supervision of city departments, the preparation of the annual budgets, and the coordination of city relations with state and federal agencies. The seven councilmen are strictly a legislative body.

Political issues have changed. Gone is the antagonism between city and state governments that spanned the eras from governors Huey Long in the 1920s, through Earl Long in the 1940s. Political corruption is no longer an issue in city politics, and segregation also has deteriorated as a defensible position for whites to espouse. (The local White Citizens Council, which was strong in the 1950s, has all but disappeared.) On the other hand, blacks have become more politically articulate. Several black political organizations participate actively in local and state political campaigns.

Although most city and parish government has been consolidated in New Orleans, Orleans Parish officials continue to play an important role. These officials include the district attorney, the board of assessors, and the Orleans Parish School Board, who supervise public education under the state department of education.

Municipal services. In the 1960s, expansion of new residential areas in New Orleans, combined with the spiralling cost of services, caused the operating budget for municipal services almost to double. New Orleans government entered the early 1970s suffering from a grave lack of the funds necessary to carry out its work effectively and to provide an appropriate income for the employees of its various departments. One of the major problems is the low assessment of taxes on both residential and industrial property and the loss of taxpayers to the suburban parishes.

Drainage has always been the number-one problem among municipal services. Virtually surrounded by levees—25 feet high on the Mississippi River and 10 feet high on Lake Pontchartrain—the below-sea-level city has 14 major drainage pumping stations, with the average capacity of a single station being 2,500 cubic feet per second. The drainage machinery used at these stations is among the largest found in the world. Following the disaster of 1965, when Hurricane Betsy flooded the city's lower Ninth Ward, the Sewerage and Water Board operating the pumping stations drafted a plan to make these facilities hurricane safe. Further improvements in drainage canals and pumping equipment are planned in the older section of the city for completion by 1980.

Drainage and flood-control systems

Flood control along 129 miles of the river, Lake Pontchartrain, and secondary waterways in the city is under the direction of the Board of Levee Commissioners of the Orleans Levee District. In addition to its primary job of flood control, the board has, since the 1920s, reclaimed some 2,000 acres of Lake Pontchartrain bottom land and developed it into one of the most scenic lakefront areas in the United States. Approximately 60 percent of the land is dedicated to public facilities, which include an amusement park, sandy beaches, a marina, a cement seawall from which fishing and swimming can be enjoyed, picnic grounds, parkways with flower beds, fountains, and shelter houses. The remaining 40 percent of this reclaimed land has been turned into four residential subdivisions, which are among the finest in the city.

Police and crime

In the fight against the steadily rising crime rate in the city, the police department has, since 1961, introduced a guard-dog corps, reorganized its patrol system to increase

its effectiveness by 60 percent, created two new police districts, built new stations in older districts, more than doubled its automotive equipment, established a community relations division, and put into operation a communication van, which acts as a field command post in time of emergency.

In spite of the increase in the proportion of the black population and a high arrest rate for black youth, racial violence has been less troublesome in New Orleans than in many other cities. A racial confrontation with police occurred in 1970 when Black Panthers took over an unoccupied apartment in the densely populated black Desire housing project. Since 1970 the Urban Squad, about evenly divided between volunteer black and white policemen, has been markedly successful in securing the trust of Desire residents, and with the help also of a community centre and swimming pool, crime in the area has decreased significantly.

The social milieu. *Health.* New Orleans has become a medical and educational centre in the 20th century. Its 19-story Charity Hospital of Louisiana, with over 2,477 beds, is the teaching hospital for two adjacent institutions, the Tulane and the Louisiana State University medical schools, with which the nearby Veterans Administration Hospital is also affiliated. There are 21 hospitals within the metropolitan area containing a total bed capacity of more than 6,300. In addition to serving local residents of this area, specialists frequently treat patients from Latin America.

Education. New Orleans has nine institutions of higher learning: The Tulane University of Louisiana and its affiliate, The H. Sophie Newcomb Memorial College for Women; Loyola University; Louisiana State University in New Orleans; St. Mary's Dominican College; three primarily black schools, Dillard University, Xavier University of Louisiana, and Southern University in New Orleans; and a trade school, Delgado College. The city also has many private, parochial, and business schools. The public school system, which began in 1841 with 83 pupils and four teachers, had by the 1970s an enrollment in excess of 100,000 and some 4,500 teachers. In 1960 a public school crisis, attracting international attention, developed when an attempt was made at the token integration of two white schools. Despite this, within ten years the school system was substantially integrated at both student and teacher levels.

Racial relations. By 1970 the population of New Orleans was 45 percent black. Relations between the races, though certainly not ideal, proved less turbulent than in many other cities. Historically, racial segregation in New Orleans differed from that of other Southern cities; for example, free persons of colour always were an important minority in antebellum days. Mixed neighbourhoods, or scattered groups of blacks living near whites, have been the rule rather than the exception in housing. School integration has presented problems, particularly in the early 1960s, but these are gradually being solved. The city has an interracial Human Relations Committee, whose members work with the mayor. By the early 1970s two black judges were serving in local courts, one of whom was the first black man to be elected to the state legislature since Reconstruction. Other black Orleanians have served in administrative posts in city government.

Cultural life and institutions. The cultural life of New Orleans is a synthesis of contributions by both whites and blacks. The white American heritage is reflected in the public school system and the business and commercial life of the city, while the immigrant heritage—Irish societies, German Volkfests, Italian St. Joseph Day altars, and the like—added ethnic colour to urban conformity. The black heritage is particularly rich. In antebellum days, free persons of colour were musicians, poets, journalists, business entrepreneurs, and landlords. Both black freemen and slaves were famed for their craftsmanship in such trades as bricklaying, iron grillwork, and carpentry. The contribution of black musicians to the birth of jazz out of black blues and "field hollers" and white dance tunes and hymns is well known. New Orleans, therefore, is one American city in which the black as well as the white cultural contribution is abundantly clear and acknowledged.

Facilities for recreation and relaxation in New Orleans are justly famous. Often referred to as "the city that care forgot," New Orleans has always been a town for those seeking a good time. Its residents love music, dancing, and a "Continental Sunday" spent in amusements. The three factors that have contributed to its popularity with tourists are the Old World charm of the Spanish-French architecture in its Vieux Carré, the gay abandon of its carnival and Mardi Gras, and its reputation as the birthplace, between the 1880s and World War I, of Dixieland jazz.

The Vieux Carré: Creole architecture and ironwork

The Vieux Carré is a sightseer's delight. Its Creole architecture, creating the atmosphere of a foreign city, combines native architectural ingenuity with adaptations of French colonial traditions of eastern Canada and West Indian Spanish colonial styles. Typical are one-story cottages opening directly on the sidewalks, with high-pitched roofs and windows reaching to the ground. Another style is the L-shaped two-story dwelling with a side entrance to an inner patio. Also built to the sidewalk, it has a roof that extends out over balconies on both the street and patio sides. Iron grillwork, designs for which were created locally and executed to a high perfection by slave craftsmen, decorates these balconies and also supports the roof. Such houses are built side by side with no openings between them, but the patios offer space for trees, flowers, and fountains and ensure privacy for the occupants.

Central to the Vieux Carré is Jackson Square, facing which are the Cabildo and Presbytère (former governmental centres, but now state museums) and St. Louis Cathedral. All date from colonial times, but considerable stylistic changes have been made on these buildings since they were erected.

On either side of this square are the Pontalba Apartments, built between 1849 and 1851, while nearby is the historic French Market. Curio and antique collectors throng the many shops on Royal Street. Side streets are lined with art galleries, perfume shops, sidewalk cafés, and tearooms. Bourbon Street is famous for its nightclubs, where jazz and risqué floor shows are a specialty. Devotees of jazz may also visit Preservation Hall and Dixieland Hall, where revivals of traditional styles may be heard. The New Orleans Jazz Club has established a Jazz Museum, which is open to the public; and each spring the city puts on an International Jazz and Heritage Festival.

Every April the New Orleans Spring Fiesta Association sponsors walking tours of private homes and patios in the Vieux Carré and also of the spacious Garden District uptown, the elite 19th-century neighbourhood. Boats tour the extensive harbour facilities and the magnificent scenery of nearby waterways. The observation point atop the International Trade Mart (400 feet, or 122 metres) at the foot of Canal Street offers a panoramic view of the river and city. Adjacent to this commercial centre is The Rivergate, a mammoth exhibition hall. The world renowned Creole cuisine may be sampled in numerous restaurants, ranging from elegant dining rooms with French menus and waiters, to small cafés with checkered tablecloths, serving red beans and rice.

The sporting life and the lively arts

Sports share an honoured position with jazz and carnival activities in New Orleans. The city is the home of the New Orleans Saints, a member of the National Football League, and a dome stadium, to be known as the Superdome, is scheduled for completion in the 1970s. Located near the Civic Center, it will be one of the world's largest sports arenas. Crowds up to 82,000 may be seated in this structure. Racing has a 100-day season at the local Fair Grounds Race Track, while golf, a popular pastime, attracts top golfers every year to the $100,000 Greater New Orleans Open Golf Tournament held at Lakewood Country Club. Boating, fishing, and swimming are popular pastimes on the city's many waterways. The city's Southern Yacht Club, on Lake Pontchartrain, is the second oldest in the country. In addition to the lakefront, popular recreation areas include the city's two largest parks,

Audubon and City. The New Orleans Recreation Department operates more then 100 playgrounds and directs organized recreation activities for thousands of youngsters. At the end of each year the Mid-Winter Sports Carnival is held in the city, featuring amateur competition in all major sports. It concludes on New Year's Day with the Sugar Bowl football contest between outstanding college teams.

Since World War II, New Orleans has become an art centre, with many artists and art galleries offering original works to collectors. The New Orleans Museum of Art is a public museum displaying many art treasures. Live theatre is represented by several "little theatre" groups. Musical events include eight operas staged annually by the New Orleans Opera House Association, the 16 concerts a year given by the New Orleans Philharmonic-Symphony Society, a summer pops concert series, and concerts presented by the New Orleans Jazz Club and the New Orleans Recreation Department.

Mardi Gras

The New Orleans carnival season, one of the events lending character to the contemporary city, begins after Christmas with local carnival organizations holding balls almost every night until Mardi Gras, the "Fat Tuesday" before Ash Wednesday. The two weeks before Mardi Gras are filled with parades, both day and night, climaxing on Mardi Gras with the Rex parade. The first parading carnival group (called a "krewe") was the "Mystick Krewe of Comus," which appeared in 1857, though celebrations by masked students go back to 1827. The krewe of Rex came into existence in 1872. Krewes, parades and carnival crowds have been growing ever since.

In the latter half of the 19th century, there were about a half dozen leading newspapers including one in French. Today, through a process of gradual consolidation, New Orleans has two major daily newspapers: The *Times-Picayune* (morning) and the *New Orleans States-Item* (afternoon). Though both are owned by the same company, keeping competition minimal, they maintain high standards of journalism. Competitive journalism is kept alive among the city's five television stations, four of which are commercial and one educational. In addition, there are 15 local radio stations.

Problems and prospects. The major problem facing New Orleans is its lack of financial resources to maintain many vital city services, a situation directly linked to the faulty tax base heavily dependent upon traditionally low property assessments, and the move to neighbouring parishes of prospective taxpayers. In addition to the possibility of revisions in the total taxation structure, a further hope for the future is the development of tracts of untouched land within the city limits for new suburban-like neighbourhoods and the urban renewal of old neighbourhoods. The crime rate appears to be levelling off, but drug addiction, associated with such crimes as shoplifting, pickpocketing, and robbery, has increased and become a problem among high school students. The efforts of police, school officials, and drug clinics toward prevention and rehabilitation, however, are showing favourable results.

Racial tension is a problem that a city almost equally divided must live with for years to come. Aside from violence associated with black militants in 1970, the high crime rate among lower-income black youth that grows out of their slum environment, and some minor racial altercations in high schools, the two races coexist fairly amiably in New Orleans. With the continuing rise in black population and the broader opportunities opening up for blacks in education and professions, blacks are destined to play a much more prominent part in the future economic and political life of the city than ever before.

Tourism, one of New Orleans' major industries, faces two challenges for the future: it must regulate Mardi Gras more rigidly to ensure a safer and more pleasant holiday for tourists, and it must provide more entertainment for its youth. Already a mecca of sophisticated adult entertainment, the "Crescent City" hopes to add more child-oriented entertainment such as children's museums, amusement parks, and a modern zoo.

BIBLIOGRAPHY. Old but reliable general histories are WILL H. COLEMAN (comp.), *Historical Sketch Book and Guide to New Orleans* (1885); H. RIGHTOR (ed.), *Standard History of New Orleans, Louisiana* (1900); and G.E. KING, *New Orleans: The Place and the People* (1895, reprinted 1928). Popular general histories that include surveys of the modern city are O.W. EVANS, *New Orleans* (1959); and H. SINCLAIR, *The Port of New Orleans* (1942). A topical history that includes modern data is H. CARTER (ed.), *The Past As Prelude: New Orleans, 1718–1968* (1968). Detailed period studies are J.G. CLARK, *New Orleans, 1718–1812: An Economic History* (1970); R.C. REINDERS, *End of an Era: New Orleans, 1850–1860* (1964); G.M. CAPERS, *Occupied City: New Orleans Under the Federals, 1862–1865* (1965); and J.J. JACKSON, *New Orleans in the Gilded Age: 1880–1896* (1969). Port history and activity is discussed in R.W. BRADBURY, *The Water-Borne Commerce of New Orleans* (1937); A.I. WARRINGTON, *Economic Geography of New Orleans and the Middle South* (1952); *New Orleans Port Handbook and Manual: A Reference Book on a Major World Port* (1961); and the *Annual Report of the Board of Commissioners of the Port of New Orleans.* For politics and government, see L.V. HOWARD and R.S. FRIEDMAN, *Government in Metropolitan New Orleans* (1959); and M. INGER, *Politics and Reality in an American City: The New Orleans School Crisis of 1960* (1969). The best sources for detailed descriptions of historic places in the Vieux Carré are S.C. ARTHUR, *Old New Orleans* (1936) and *New Orleans City Guide* (1938). Detailed statistical information on city services is in the *Louisiana Almanac* (annual); the reports of the Mayor of New Orleans, of the Board of Levee Commissioners of the Orleans Levee District (1956), of the Sewerage and Water Board (1969), and in the *National Atlas of the United States of America* (1970). Population and manufacturing statistics are in the UNITED STATES DEPARTMENT OF COMMERCE, BUREAU OF THE CENSUS publications: *Louisiana General Population Characteristics* (1970); *Louisiana Final Population Counts* (1970), and *1967 Census of Manufactures, Louisiana* (1967).

(J.J.J.)

New South Wales

The first British colony in Australia, New South Wales is the oldest, richest, and most industrialized state of the Australian Commonwealth. Originally, the name New South Wales was applied to the entire east coast of Australia when the British explorer Capt. James Cook claimed the territory for the British crown in 1770. The separate colonies of Tasmania, South Australia, Victoria, and Queensland were proclaimed in the 19th century, and in 1911 and 1915 the Australian Capital Territory around Canberra and Jervis Bay was ceded to the commonwealth. New South Wales was thus gradually reduced to its present area of 309,433 square miles (801,428 square kilometres). It is bounded by the Pacific Ocean to the east, the states of Victoria to the south, South Australia to the west, and Queensland to the north. By no means the largest Australian state, it is the most populous, with over 4,500,000 inhabitants by the early 1970s. Its capital, Sydney, is the nation's largest city. Lord Howe Island, a dependency of New South Wales, is administered as part of the state.

New South Wales reflects the dynamism and the growth problems of a fast-developing country. It is Australia's focal point of commercial farming, industry, and cultural development. It is also plagued by an imbalance between its urban and rural populations and often chafes under the financial restrictions of the commonwealth government. (For a discussion of the capital city, see SYDNEY.)

THE LANDSCAPE

The Great Dividing Range

The natural environment. *Relief.* The dominant geographical feature is the Great Dividing Range which separates the narrow coastal strip from the great plains to the west. The coastal strip varies from 10 to 50 miles in width and is bounded along its entire length by a natural barrier of steep mountains. Beyond this barrier, however, the Great Dividing Range consists not of true mountains but of a series of high plateaus, or Tablelands, which slope gently to the west until they merge imperceptibly into the plains beyond. The average height of these Tablelands is about 2,500 feet. In some areas they rise to 5,000 feet, and they attain a height of 7,328 feet (2,234 metres) at Mt. Kosciusko—the highest mountain in Aus-

tralia—in the Australian Alps in the south. The gentle western side of the Tablelands is known as the Western Slopes. The plains cover nearly two-thirds of the state. Lying below 1,000 feet, they are interrupted only by the elevated country between Orange and Cobar in the east and by the Main Barrier and Grey ranges in the west.

Drainage. The Great Dividing Range is the state's main watershed. Numerous rivers flow eastward from the range to the Pacific Ocean. Though often beautiful, they are too short and rapid to be of much economic value. The major rivers that flow west—the Namoi, the Macquarie, the Lachlan, and the Murrumbidgee—cross some 500 miles of sunburned plains before joining the Murray and Darling rivers. These brown, muddy, inland rivers meander across the plains and lose a great deal of their water by evaporation. Over 1,600 miles of the Darling River, which rises in Queensland, flow to the southwest across the plains to join the Murray. Rising in Victoria, the Murray runs for over 1,200 miles along the southern border of New South Wales before crossing South Australia to reach the Pacific Ocean.

Soils. A great deal of New South Wales is naturally fertile, and the red and black soil plains are extremely rich. The coastal strip, however, consists mostly of poor and sandy soils. Agricultural potential is severely limited by inadequate and uncertain rainfall and intense evaporation.

Climate. About 19 percent of the state receives less than ten inches of rain a year, and approximately 23 percent receives only ten to 15 inches. The coastal districts have the most annual rainfall, varying from about 30 inches in the south to about 75 inches in the north. Precipitation diminishes progressively away from the coast. The average annual rainfall in the northwest is only eight inches, and some of the land beyond the Darling River merges into desert.

Rainfall

If the dry climate and brilliant sunshine present severe agricultural problems, they are yet attractive to the inhabitants of the coastal cities. Newcastle, Sydney, and Wollongong have a delightful climate. It is rarely too hot in summer or too cold in winter, and Sydney is without sunshine for an average of only 23 days a year. Inland it is both hotter in summer and colder in winter. Average temperatures range from 74° to 83° F (24° to 29° C) in the summer months and from 49° to 54° F (7° to 13° C) in winter. Temperatures of over 100° F (38° C) are frequent in summer, and frost at night is common in winter. Heavy snow falls on the southern mountains and, more rarely, on the northern and central Tablelands. On Mt. Kosciusko snowdrifts linger throughout the summer.

The seasons are fairly well defined; autumn begins in March, winter in June, spring in September, and summer in December.

Vegetation. The natural vegetation ranges from the dense semitropical forest of the northern coast to the sparse plant life of the western plains. Nearly one-tenth of the state is forested, and, except for the plains, a much larger area is covered with bush and scrub. The forests, concentrated mainly on the coast and Tablelands, give way to shrub eucalyptus on the Western Slopes and to salt bush and spinifex (a spiny grass) in the far west. The predominant tree is the eucalyptus, which is the state's chief source of hardwood. Smaller quantities of softwoods, such as the red cedar, the hoop pine, and the rosewood, are found on the northern coast, while the cypress pine grows on the Western Slopes. Native grasses are found everywhere but in the extreme west, and there are many wild flowers.

Animal life. The rich birdlife includes many species of parrot and cockatoo, the flightless emu, and the mound-building scrub turkey and mallee bird. Most of the species of marsupials, mammals that do not develop a placenta and that carry their young in an external pouch, are represented. These include the wombat, the koala, the common and ring-tailed opossum, the common and long-nosed bandicoot, and a variety of kangaroos and wallabies. Many of the smaller marsupials, diminished by agriculture and forestry, are retreating into uninhabited regions. The platypus and spiny anteater are common.

Marsupials

Several species of poisonous snakes, including black, brown, and tiger snakes, are found throughout the state. The best known fish is the Murray cod, which is found in the western rivers and is an excellent food source. Trout have been introduced into the streams of the Great Dividing Range.

Traditional regions. There are four regions of traditional activity. The coastal strip is mostly used for dairy farming, the Tablelands for sheep and cattle raising, and the Western Slopes for wheat cultivation. The plains are the site of the great sheep stations, for over 100 years the basis of the state's economy. They still provide almost 40 percent of Australia's wool.

There are also four distinct political regions. The northern Tableland is known as New England. Its population, almost entirely rural, tends to resent the government's preoccupation with Sydney. There has long been an unsuccessful movement to form a new state. Similar feelings can be found in the Riverina district, located between the Murray and Murrumbidgee rivers. There, too, people tend to favour separatism; in practical matters, they look toward Melbourne, the capital of Victoria, rather than to Sydney. The Monaro sheep country comprises the windy, upland district of the southern Tableland, including what is now the Australian Capital Territory. The sheep graziers of the Western plains, living a remote and lonely life, feel a common loyalty born of common interests.

The landscape under human settlement. *Rural settlement.* Climatic conditions have also dictated the size of landholdings. The smallest are in the coastal strip, where dairy and fruit farming are the chief occupations. Holdings are considerably larger—between 500 and 5,000 acres—on the Tablelands and Western Slopes, where mixed farming is the normal practice. Further west they grow larger still; in the dry lands of the Western Division, many sheep stations cover over 100,000 acres. The exception to this general rule is in irrigated areas along the Murrumbidgee and other inland rivers, where diversified cultivation makes small holdings possible.

Landholdings

The tenure of landholdings in New South Wales is mostly either freehold or leasehold from the crown. Tenancy, as understood in Europe, is uncommon, and, except in the Western Division, most land is occupied by the owners.

Because of the large size of the average field, or paddock, and the relatively uncultivated appearance of the land, a typical sheep station presents a characteristic appearance. The heart of each property is the homestead, with its cluster of low buildings and well-watered trees and gardens, surrounded by bare, brown paddocks. There are no villages, and the nearest country town may be 30 or 40 miles away. In spite of the severe problems faced by sheep graziers because of the unstable price of wool and the constant threat of drought, life on the larger properties is still considered a pleasant and privileged one.

Urban settlement. The largest industrial urban area is the coastal complex of Newcastle, Sydney, and Wollongong. There are several towns of 30,000 or fewer inhabitants in the interior, such as Albury and Wagga Wagga in the southeast and Orange and Tamworth in the east central region. These are essentially country towns serving the surrounding rural population. The only interior industrial town is Broken Hill, in the far west, which depends upon the rich mineral deposits in the Barrier Range.

THE PEOPLE

Population groups. *Ethnic origins.* The people of New South Wales are in no way different from those of Australia as a whole. Nearly 80 percent are of British origin, and over 20 percent are of Irish descent. The small remainder is composed mainly of continental Europeans. There are also some 24,000 native tribesmen, or Aborigines, and a few thousand Chinese.

Religious groups. Almost the entire population professing a religion is Christian. The largest groups are Anglicans (Church of England), 42 percent, and Roman Catholics, 31 percent. Other groups are Presbyterians, Methodists, Greek Orthodox, and Baptists. Lutherans and Congregationalists each comprise less than 1 percent of the population.

Demography. The birthrate and deathrate and other vital statistics do not vary substantially from those of the rest of Australia. Since 1960, however, New South Wales has had a slightly lower birthrate than the other states, presumably because of the higher proportion of people living in cities. The birthrate in 1965 had fallen to 18.6 per 1,000 but has since risen slightly. Since 1947, immigration has accounted for over one-third of the total population increase.

The most striking feature of the population is the disparity between those living in the rural areas and those in the cities. Three-fourths of the state's 4,500,000 people are crowded into 2 percent of its area in Newcastle, Sydney, and Wollongong. This urban population increased by over 400,000 from 1966 to 1971, while that of the rest of the state declined by more than 126,000. The state government has long been concerned by this trend, and various schemes have been proposed to correct it. If it continues, by the end of the 20th century 82 percent of the population will be crammed into the central-coast industrial area.

THE ECONOMY

Economically the most important state in Australia, New South Wales contains about one-half of the country's sheep, one-quarter of its cattle, and one-third of its pigs. It produces a large share of the nation's grain, dairy products, and wool and mines most of its black coal and silver, lead, and zinc. It is also the country's most industrialized area and produces over two-fifths of the nation's manufactured goods.

The extent and distribution of resources. *Biological resources.* In 1970 there were about 70,000,000 sheep, 150,000 horses, and 5,500,000 cattle grazing on the state's vast grasslands. There were also about 500,000 pigs. The state forests provide an important natural resource of hardwood timber, although the percentage of forest land —12 percent—is low by international standards. The Pacific Ocean provides valuable fish.

Mineral resources. The most important mineral resource is the vast black-coal deposits of the central coastal region. The main silver, lead, and zinc deposits are located at Broken Hill. Large copper deposits have been discovered at Cobar. Other mineral resources include tin from the Ardlethan and Tallebong fields and rutile (a red or black mineral cut into gems), found in the coastal sands. By the early 1970s, no iron ore, nickel, uranium, oil, or natural gas had been discovered.

Power resources. Coal is the most developed power resource, while the Snowy River offers important hydroelectric potential.

Sources of income. *Agriculture, forestry, and fishing.* Agriculture is spread throughout the state. About half of the acreage under crops is devoted to wheat, which is grown for both domestic consumption and export. Other grains grown include oats, maize (corn), and rice; fodder, potatoes, grapes, sugarcane, and fruit are also raised. Excellent wine is produced in the Hunter Valley and cotton is grown on the Namoi River. Sheep are raised mainly for their wool, which is also exported. Both slaughter and dairy cattle are important.

Wheat and wool

New South Wales is the most important timber-producing state, accounting for about one-third of Australia's production. Hardwoods and softwoods are exploited, and there is a regular pine reforestation program.

Tuna fishing is the most important marine industry, and mullet, shark, and Australian salmon are also caught in significant numbers. The state provides about one-third of the national fish catch, as well as all of the oysters consumed domestically.

Mining and quarrying. The most important coalfields are in Hunter Valley above Newcastle, around Wollongong, and at Lithgow. Silver, lead, and zinc are mined at Broken Hill. Copper is mined chiefly near Cobar, and tin is mined in the central and northern Tablelands.

Industry. About two-thirds of the manufacturing industries are located in Sydney. Other important industrial centres are Newcastle, Wollongong, Lithgow, and Broken Hill. The fastest growing industry is the manufacture of iron and steel and metal goods. Other important products are textiles, food, beverages, tobacco, chemicals, paints, paper, and printed material. Factories are mainly small, and only a few employ over 1,000 persons. A very few large concerns, such as the Broken Hill Proprietary Company Ltd. (iron and steel) and the Australian Consolidated Industries Ltd. (glass), employ over half the industrial workers.

Power. Electricity is generated and distributed by the Electricity Commission of New South Wales. Power is sold to local authorities, the state railways and tramways, and large industrial users. Almost all the state's electricity is generated by thermal-power stations. The Snowy Mountains hydroelectric scheme, begun in 1949 and to be completed in 30 years, has diverted the waters of the Snowy and other rivers westward into the Murrumbidgee. After its completion, about 20 percent of the state's electricity will be generated by waterpower.

Finance. Since 1942 the commonwealth has collected all income taxes and reimbursed the states on an agreed formula. As a result, New South Wales is almost entirely dependent upon commonwealth grants for its revenue. In 1971 the states were given the right to levy a payroll tax, but it was not clear that this "growth tax" would be sufficient for their mounting needs.

Management of the economy. Like the rest of Australia, the state has essentially a capitalist economy, with the great majority of industry owned by public companies. The state government, however, owns and manages the railways, some coal mines, and the production of electricity. There is a vigorous trade-union movement. The Chamber of Manufacturers and other employers' organizations represent the interests of private enterprise.

Transportation. New South Wales has excellent internal air services. The state railways provide an adequate link between Sydney and larger population centres. They have suffered, however, from competition with air and road transport. Sydney has Australia's only underground rail network, and electric rail services connect the city with many of its suburbs.

Metro services in Sydney

Roads. There are 131,330 miles of public roads, including 22,410 miles of state highways and main roads. Most of these are paved, but many, including the main highways to Brisbane and Melbourne, are too narrow for the traffic they now bear. Most country roads are surprisingly good, though not all are yet paved.

Inland waterways. There are no commercial waterways, although before the railways were built the Murray and Darling rivers and their tributaries were used extensively for carrying freight.

Ports. The four major ports are Sydney, Newcastle, Port Kembla, and Botany Bay. Together they handle approximately 40,000,000 tons of cargo each year. All ports are administered by the Maritime Services Board of New South Wales.

Railways. The railways cover over 6,000 miles and centre upon Sydney. They also provide adequate facilities to the other major urban centres including Broken Hill, 700 miles to the southwest. The lines run from north to south along the coast and roughly from east to west in the Tablelands and plains.

Air services. Air services, largely for passenger traffic, are provided by a number of lines. Air links are operated to all the major towns, and the airport at Sydney can accommodate jet aircraft.

ADMINISTRATION AND SOCIAL CONDITIONS

Government. *Constitutional framework.* The state government administers internal matters, while the national government (commonwealth) is responsible for defense, foreign policy, immigration, trade, customs and excise, post and telegraph services, and air and sea transport. Within those limitations, the state is sovereign and has powers to make laws for the welfare and good government of New South Wales. It has no armed forces other than the police.

The Parliament consists of two houses—the lower house, or Legislative Assembly, of 94 members who are directly elected by the people, and the 60-member upper

Houses of Parliament

house, or Legislative Council, one-fourth of whose members are elected by both houses every three years. The Cabinet is chosen from the party that commands a majority in the Legislative Assembly; it is headed by a premier. Parliament meets for three years but can be dissolved earlier.

The state also has its own governor, who is appointed by the Queen on the recommendation of the government. The titular head of the government, he is now always an Australian, and his duties are almost entirely formal.

Local government. New South Wales is divided into more than 200 local-government areas, which are controlled by councils. These councils are elected every three years by adult residents of their areas.

Elections. All elections are by universal adult suffrage; the voting age is 18 years. Voting for the Legislative Assembly is compulsory, but voting for local councils is not.

Political parties. Political parties are usually state branches of the federal political parties and tend to have the same policies and interests. The three principal parties are the Liberal Party, the Country Party, and the Australian Labor Party. There is also a branch of the Australian Democratic Labor Party, without a seat in either house, and a small Communist Party, which still retains some influence in a few trade unions. Elections are fought on state issues, but the fortunes of the parties are often affected by the fortunes of the same parties in the commonwealth.

Justice. State law and its administration are largely based upon the British system. Legal procedure includes trial by jury in criminal cases, the right of appeal, and an independent judiciary. The highest state court is the Supreme Court, against which appeals can be made to the High Court of Australia and, in certain cases, to the Privy Council in London. Minor offenses are dealt with by magistrates in the Courts of Petty Sessions, while more serious cases are brought before a judge and jury in the Courts of Quarter Sessions.

Administration. *Education.* Education is compulsory for all children between the ages of six and 15. Most children are educated in free, nondenominational primary and secondary schools. A minority go to private schools, most of them administered by the Roman Catholic Church. There are also a few denominational secondary schools for the children of the wealthy. There are five universities in the state; the cost of administration is shared with the commonwealth.

Health and welfare. The state government is responsible for the supervision of public health, hospitals, and medical services. There are almost 270 public and a large number of private hospitals. The national health-insurance scheme is financed and administered by the commonwealth, as are the social services such as family allowances, child endowment, unemployment benefits, and pensions.

Standard of living

Social conditions. New South Wales shares in the high standard of living and generous social services enjoyed by all Australians. The state has its own Industrial Commission to settle industrial disputes, though here, too, the Commonwealth Arbitration system has become predominant. The basic wage—adopted from time to time by the Commonwealth Conciliation and Arbitration Commission—is adopted for state awards and agreements. The 40-hour workweek is now standard.

CULTURAL LIFE AND INSTITUTIONS

The state cannot claim a unique culture that distinguishes it from the rest of Australia. It is perhaps the most representative and typical of all the states, and its greater population and wealth give it a leading position over all but Victoria.

The state of the arts. Sydney has the oldest and reputedly the best of the symphony orchestras. There are many small theatres, but no professional drama company of first-class quality was established by the early 1970s. Sydney's Opera House, which opened in 1973, is a major arts centre with a concert hall, a large theatre for opera and ballet, and a small theatre.

Many of Australia's leading poets and novelists were either born in the state or live there. Sydney is the centre of a school of contemporary painting that looks toward the United States and Europe for its inspiration. The Art Gallery of New South Wales has the best collection of Australian painting in Australia. In general, the cultural climate is marked more by a healthy and democratic vitality rather than by any special distinction. This is equally true of the popular arts, where there are large audiences for and keen interest in both Australian and foreign folk singers and jazz and rock groups.

The mass media. Sydney has three morning papers, including the *Australian*, which has a national circulation, and the *Sydney Morning Herald*, which is the oldest paper in Australia. There are also two evening papers and four Sunday papers. Many weeklies and magazines are also published there. Most country towns, even those of medium size, also have their own newspapers. In both television and radio broadcasting, the Australian Broadcasting Commission competes with a number of commercial stations.

Prospects for the future. Because of its wealth, geographical position, and mixed economy, New South Wales is well placed to share in the progress and prosperity of Australia. It is likely to remain, for some time to come, the most important state in the commonwealth. Whether it will also remain a distinct political entity is more doubtful and must depend on the future of federalism in the country. Though there is a certain loyalty to their state among New South Welshmen, this more often finds expression in sports than in politics. The Australian Labor Party openly and some Liberals secretly are centralists who would like to abolish the states or at least reduce them to administrative units. It can be argued that the state governments and parliaments are in any case likely to lose their reason for existence if they are not given back the power to raise their own revenue to meet their own expenditures.

BIBLIOGRAPHY. The best source for up-to-date facts and statistics on New South Wales is the *Official Year Book*, issued annually by the government of New South Wales in conjunction with the Commonwealth Bureau of Census and Statistics. Historical, as well as other, information on New South Wales may be found in books on Australia, such as C.M.H. CLARK, *A History of Australia*, 2 vol. (1962–68); G. GREENWOOD (ed.), *Australia: A Social and Political History* (1955 and 1966); and A.G.L. SHAW, *The Story of Australia* (1955).

(J.D.Pr.)

New Thought

New Thought, a mind-healing movement based on religious and metaphysical (concerning the nature of ultimate reality) presuppositions, originated in the United States in the 19th century. Involving both individuals and denominations, it maintains an influence beyond the particular organizations associated with the movement. The diversity of views and styles of life represented in various New Thought groups are difficult to describe because of their variety, and the same reason makes it virtually impossible to determine either membership or adherents. The influence of the various New Thought groups has been spread by its leaders through lectures, journals, and books not only in the United States but in the United Kingdom, in Europe, Asia, Africa, and Australia. Many adherents of New Thought consider themselves to be Christian, though generalizations about their relations to Christianity have been questioned.

History of the New Thought movement. *Origins.* The origins of New Thought may be traced to a dissatisfaction on the part of many persons with scientific empiricism and their reaction to the religious skepticism of the 17th and 18th centuries. The romanticism and idealism of the 19th century also influenced the New Thought movement, of which Phineas P. Quimby (1802–66) is usually cited as the earliest proponent. From Portland, Maine, Quimby practiced mesmerism (hypnotism) and developed his concepts of mental and spiritual healing and health based on the view that illness is a matter of the mind. Quimby's influence may be seen in the writings

Influence of Phineas P. Quimby

of Mary Baker Eddy and in the development of Christian Science (which she founded), although Mrs. Eddy retracted acknowledgment of dependence on her teacher. Quimby's influence has been readily acknowledged by Warren F. Evans (1817–89) and by Julius A. Dresser (1838–1893) and his son Horatio W. Dresser (1866–1954). Evans, a Methodist and then a Swedenborgian minister (leader of a theosophical movement based on the teachings of the 18th-century Swedish scientist and theologian Emanuel Swedenborg), published a number of works exploring and systematizing the ideas of Quimby. These included *Mental Cure* (1869), *Mental Medicine* (1872), and *Soul and Body* (1876). Julius Dresser was a popular lecturer who emphasized the theories of Quimby, and his son Horatio spread the elder Dresser's teachings and later edited *The Quimby Manuscripts* (1921).

Organization of New Thought groups. As the ideas of Quimby and his students began to spread, those interested began to gather in congregations. They organized New Thought congresses (the first in 1894), the National New Thought Alliance in 1908, and the International New Thought Alliance (INTA) in 1914. In the Alliance are (or have been) such groups as the Unity School of Christianity (founded by former Methodists Charles and Myrtle Fillmore), Divine Science (founded by Nona L. Brooks), the Church of Religious Science (founded by Ernest Holmes), and others, such as the Home of Truth, the Church of Truth, the Christ Truth League, the Society of the Healing Christ, and the Christian Assembly. The Alliance has depended for its continuity upon presidents, two of whom have had long tenures (James A. Edgerton, 1915–23, and again 1934–37; and Robert H. Bitzer, 1949–63), and upon the cooperation of the leaders of various schools of New Thought recognized by the Alliance. Its headquarters has been in Washington, D.C., in New York City, and in Hollywood, California, where it is now located. Smaller groups have often gathered around teachers noted for their powers of healing and have organized themselves in rather loose congregational patterns. There is no single organizational structure.

Member groups of the New Thought Alliance

New Thought teachings and practices. New Thought represents the influence of Platonism (based on the views of the 5th–4th century BC Greek philosopher Plato, involving the view that the realm of ideas is more real than that of matter), Swedenborgianism, (especially Swedenborg's view that the material realm is one of effects whose causes are spiritual and whose purpose is divine); Hegelianism (based on the views of the 18th–19th century German philosopher G.W.F. Hegel, especially those concerning the external world, mental phenomena, and the nervous organism as the meeting ground of the body and the mind); Orientalism (involving spiritually-oriented teachings of Eastern religions; *e.g.*, Hinduism); and, particularly, the Transcendentalism (a form of Idealism) of the 19th-century American philosopher and poet Ralph Waldo Emerson. Though a summary of New Thought beliefs is difficult to make, since it is to a large degree individualistic in outlook, it is possible to summarize some of the more prevalent views. As far as Christian Science is concerned, New Thought adherents do not accept Mrs. Eddy's teaching or any other formulation as the final revelation. Rather, truth is viewed as a matter of continuing revelation, and no one leader or institution can declare with finality what is its nature. Moreover, New Thought does not oppose medical science, as Mrs. Eddy did, and it is essentially positive and optimistic about life and its outcome.

Basic views. In 1916 the Alliance agreed upon a purpose that embraces some central ideas of most groups:

> To teach the Infinitude of the Supreme One; the Divinity of Man and his Infinite Possibilities through the creative power of constructive thinking and obedience to the voice of the indwelling Presence which is our source of Inspiration, Power, Health and Prosperity.

Views on the divinity of man and his potentialities

In 1917, at the St. Louis (Missouri) Congress, the Alliance adopted a "Declaration of Principles." It was modified in 1919 and was printed in *New Thought* until revised in the 1950s. This purpose and these Principles emphasized the immanence of God, the divine nature of man, the immediate availability of God's power to man, the spiritual character of the universe, and the fact that sin, human disorders, and human disease are basically matters of incorrect thinking. Moreover, according to New Thought, man can live in oneness with God in love, truth, peace, health, and plenty. Many New Thought groups emphasize Jesus as teacher and healer and proclaim his kingdom as being within a person. Reference to Jesus or the Christ is totally omitted in the Principles, however, as revised in 1954. New Thought leaders, unlike Quimby, it should be noted, have increasingly stressed material prosperity as one result of New Thought. New Thought implies a kind of monism (oneness of the world), but it also has strong Gnostic (an esoteric dualism in which matter is opposed to spirit) undertones; that is, though New Thought is open to all, spiritual healing and strength of mind and body are only available to those who have the insights and who have been initiated into the movement at some point. There are no established patterns of worship, although the services often involve explication of New Thought ideas, testimony to healing, and prayer for the sick.

Spread of New Thought ideas. New Thought conceptions have been spread by a wide variety of periodicals, the following of which are still being published: *New Thought* (INTA); *Unity*, *Weekly Unity*, and *Christian Business* (Unity School of Christianity); *Divine Science Monthly*, *Science of Mind*, *Religious Science*, and *Crusader*, (published by Brother Mandus, English leader of the World Healing Crusade). New Thought ideas have also been spread by a number of writers who have produced several religious books with significant popular appeal: Ralph W. Trine, *In Tune with the Infinite* (1897); Orison Swett Marden, *Pushing to the Front* (1894); Robert Collier, *The Secret of the Ages* (1926); and Emmet Fox, *Power Through Constructive Thinking* (1940), and *The Sermon on the Mount* (1934). More recently, clergymen outside the movement, such as Glenn Clark, in *How to Find Health Through Prayer* (1940), and Norman Vincent Peale, in *The Power of Positive Thinking* (1952), have contributed to the distribution of some New Thought ideas. Though it is very difficult to generalize about the social and ethical consequences of New Thought, there is evidence, as suggested by this popular literature, that New Thought has been individualistic in its ethical concerns. It has attempted to promote a positive personal attitude toward life and to encourage a success mentality in an industrialized and urbanized America. This has made it appealing to middle and upper class clientele and has reinforced the suspicion of its adherents concerning those who attempt to deal with the corporate problems of life by means of political power. The general Methodist orientation of persons such as Evans and the Fillmores suggests a source of the perfectionist aspects of New Thought views.

Promotion of positive personal attitudes

New Thought groups. *Unity*. Among the most prominent groups spreading New Thought motifs is Unity; *i.e.*, the Unity School of Christianity, which was organized by Myrtle (1845–1931) and Charles (1854–1948) Fillmore. Both were originally Methodists, and both suffered from physical and emotional troubles. The Fillmores came under the influence of Emma Curtis Hopkins, a former follower of Mrs. Eddy, at the Christian Science Theological Seminary in Chicago. They started publishing *Modern Thought* in 1889 and *Unity* in 1891 and expanded their influence greatly through a ministry they called Silent Unity, which involved an affirmative prayer and counselling service on request. At first this type of healing and spiritual ministry was carried on through correspondence and later by telegraph and telephone. From small beginnings in Kansas City, Missouri, Unity expanded, with Unity Centers developing elsewhere; the Fillmores also founded a Unity Village, several miles outside of Kansas City on about 1,400 acres of farmland, in order to coordinate the work of followers. Unity developed into a publishing organization of remarkable size, distributing the writings of the Fillmores (in books,

pamphlets, and magazines) on healing, prosperity, success, prayer, and numerous other subjects. Unity publications included *Unity*, *Weekly Unity*, *Daily Word*, and *Wee Wisdom* for children; the publications have been supplemented with an extensive use of radio programs. Unity stresses its agreement with Christianity, and the Fillmores attempted to harmonize its several distinctive teachings with orthodoxy. Unlike recent tendencies in INTA and within New Thought groups, they continued to stress the teaching of Jesus Christ. According to their teaching, they revived his healing methods, but, in contradistinction to orthodox Christian churches, they teach a belief in reincarnation and the regeneration of the body in a succession of rebirths. Salvation is finally attained by breaking out of the cycle of rebirths and by reaching a true spiritual nature and Christ-consciousness. Unity allows sexual intercourse for procreation only and urges abstention from the use of meat, tobacco, intoxicants, and drugs. In 1922 Unity left the INTA, although some individuals in local Unity Centers have remained in the Alliance. Since the 1920s, Unity has developed some denominational characteristics with an accredited ministry in association with the parent body in Kansas City. Since the death of Charles Fillmore in 1948, Unity has continued under the leadership of his son, Rick, and a grandson, Charles. Unity claims to have received 650,000 contacts for a prayer a year.

Teachings of Unity differing from orthodox Christian churches

Psychiana. Though Unity is considered as Christian by its adherents, other groups promoting New Thought motifs find themselves at odds with, and even hostile to, Christianity. One such group, Psychiana (a healing cult promoted by advertisement and direct mail), was founded by Frank B. Robinson (1886-1948) of Moscow, Idaho. Robinson, son of a Baptist minister, one time student at the Bible Training School in Toronto, Canada, and pharmacist, began to advertise "God-power" in 1929. Rejecting the orthodox Christian Trinitarian beliefs in God the Father, the Son, and the Holy Spirit—and also the established Christian organizations—he stressed that the power of God was freely available for individual use. He advertised, "I talked with God," and circulated a series of "Psychiana Lessons" (1932), which demonstrated how others might also bring from the realm of the spirit of God anything needed for complete physical and material happiness. In his book, *Your God-Power* (1943), Robinson systematized some of his thoughts and showed some influence from Theosophy (a religious movement containing Hindu, Buddhist, and Western religious ideas that originated in the 19th century under the Polish-American Madame Helena Blavatsky). He called himself the archbishop of Psychiana, and, though he did head a corporation that sent out an estimated 1,000,000 sets of lessons, he headed no permanent or widespread organization.

The I Am movement. The I Am movement, also associated with New Thought, was founded in the 1930s by Guy Ballard (1878–1939), a mining engineer, and his wife, Edna (d. 1971). The I Am movement was based upon the teachings expounded in Ballard's books, including *Unveiled Mysteries* (1934), *The Magic Presence* (1935), and *The I Am Discourses. The Voice of the I Am* was the monthly periodical of the movement. Drawing upon classical mythology, astronomy, astrology, the Bible, and theosophy, Ballard taught that there were numerous Ascended Masters, who themselves had broken the bonds of human limitations. Among these, at the beginning, was Jesus, and then, most important of all, Comte de St. Germain, an 18th-century European necromancer and occultist. Persons could turn to these various masters for vast cosmic power. Furthermore, through successive reincarnations every member of the movement could become an Ascended Master. By employing such affirmation every member could be assured of health, prosperity, and happiness. To these benefits was added freedom from the fear of death because of the presence of the Mighty I Am in every man. The Ballards were known for their superpatriotism, their condemnation of Liberal causes in the 1930s, and mail fraud, for which Edna served a jail sentence after her husband's death. The fact that the I Am Movement and Psychiana declined in influence and numbers shortly after their leaders died indicates the ephemeral character of some New Thought groups.

Conclusion. New Thought has spread widely throughout the United States and the world and has contributed to the increased interest in what has been called faith healing, or better, the healing ministry, within the larger denominations of Christians. The development of this interest within these bodies may curb some of the influence of New Thought, but New Thought seems to continue because of man's apparently keen desire to assert his own value and to overcome his psychic and physical limitations, frailties, and finitude. New Thought gives expression to the desire for health and wholeness and also to the desire to realize a harmony, according to its teachings, of man's little mind with the Great Mind. It allows its adherents and those influenced by its leadership and widespread literature a buoyant optimism about themselves and the cosmos.

Influence on the larger Christian denominations

BIBLIOGRAPHY. The most helpful bibliography of books, pamphlets, magazines, and New Thought archives, as well as the best volume for the history and thought of the movement is CHARLES S. BRADEN, *Spirits in Rebellion* (1963). Primary sources are: HORATIO W. DRESSER (ed.), *The Quimby Manuscripts* (1921), and *History of the New Thought Movement* (1919); WARREN FELT EVANS, *The Mental Cure: Illustrating the Influence of the Mind on the Body, Both in Health and Disease and the Psychological Method of Treatment* (1869); CHARLES FILLMORE, *Metaphysical Bible Dictionary* (1931); and RALPH WALDO TRINE, *In Tune with the Infinite* (1897). See also *New Thought* and *Unity*, two of the most important periodicals still being published.

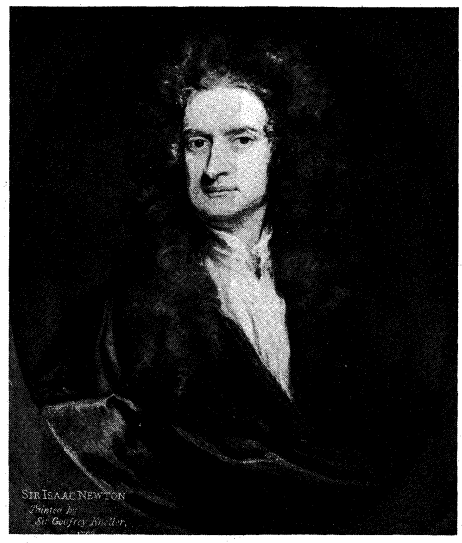
Secondary works that treat the subject sympathetically and place it in larger context are: DONALD MEYER, *The Positive Thinkers* (1965); ROBERT PEEL, *Christian Science: Its Encounter with American Culture* (1958); and LOUIS SCHNEIDER and SANFORD M. DORNBUSCH, *Popular Religion: Inspirational Books in America* (1958).

(J.H.Sm.)

Newton, Sir Isaac

Isaac Newton, English physicist and mathematician, was the culminating figure of the scientific revolution of the 17th century. In optics, his discovery of the composition of white light integrated the phenomena of colours into the science of light and laid the foundation for modern physical optics. In mechanics, his three laws of motion, the basic principles of modern physics, resulted in the formulation of the law of universal gravitation. In mathematics, he was the original discoverer of the infinitesimal calculus. Newton's *Philosophiae Naturalis Principia Mathematica* (*Mathematical Principles of Natural Philosophy*), 1687, was one of the most important single works in the history of modern science.

By courtesy of the National Portrait Gallery, London



Newton, oil painting by G. Kneller, 1702. In the National Portrait Gallery, London.

EARLY YEARS

Formative influences. Born on December 25, 1642 (January 4, 1643, new style), in the hamlet of Woolsthorpe, Lincolnshire, Newton was the only son of a local yeoman, also Isaac Newton, who had died three months before, and of Hannah Ayscough. That same year, at Arcetri near Florence, Galileo Galilei had died; Newton would eventually pick up his idea of a mathematical science of motion and bring his work to full fruition. A tiny and weak baby, Newton was not expected to survive his first day of life, much less 84 years. Deprived of a father before birth, he soon lost his mother as well, for within two years she married a second time; her husband, the well-to-do minister Barnabas Smith, left young Isaac with his grandmother and moved to a neighbouring village to raise a son and two daughters. For nine years, until the death of Barnabas Smith in 1653, Isaac was effectively separated from his mother, and his pronounced psychotic tendencies have been ascribed to this traumatic event. That he hated his stepfather we may be sure. When he examined the state of his soul in 1662 and compiled a catalog of sins in shorthand, he remembered "Threatning my father and mother Smith to burne them and the house over them." The acute sense of insecurity that rendered him obsessively anxious when his work was published and irrationally violent when he defended it accompanied Newton throughout his life and can plausibly be traced to his early years.

After his mother was widowed a second time, she determined that her first-born son should manage her now considerable property. It quickly became apparent, however, that this would be a disaster, both for the estate and for Newton. He could not bring himself to concentrate on rural affairs—set to watch the cattle, he would curl up under a tree with a book. Fortunately, the mistake was recognized, and Newton was sent back to the grammar school in Grantham, where he had already studied, to prepare for the university. As with many of the leading scientists of the age, he left behind in Grantham anecdotes about his mechanical ability and his skill in building models of machines, such as clocks and windmills. At the school he apparently gained a firm command of Latin but probably received no more than a smattering of arithmetic. By June 1661, he was ready to matriculate at Trinity College, Cambridge, somewhat older than the other undergraduates because of his interrupted education.

Influence of the scientific revolution. When Newton arrived in Cambridge in 1661, the movement now known as the scientific revolution was well advanced, and many of the works basic to modern science had appeared. Astronomers from Copernicus to Kepler had elaborated the heliocentric system of the universe. Galileo had proposed the foundations of a new mechanics built on the principle of inertia. Led by Descartes, philosophers had begun to formulate a new conception of nature as an intricate, impersonal, and inert machine. Yet as far as the universities of Europe, including Cambridge, were concerned, all this might well have never happened. They continued to be the strongholds of outmoded Aristotelianism, which rested on a geocentric view of the universe and dealt with nature in qualitative rather than quantitative terms.

Like thousands of other undergraduates, Newton began his higher education by immersing himself in Aristotle's work. Even though the new philosophy was not in the curriculum, it was in the air. Some time during his undergraduate career, Newton discovered the works of the French natural philosopher René Descartes and the other mechanical philosophers, who, in contrast to Aristotle, viewed physical reality as composed entirely of particles of matter in motion and who held that all the phenomena of nature result from their mechanical interaction. A new set of notes, which he entitled "Quaestiones Quaedam Philosophicae" ("Certain Philosophical Questions"), begun sometime in 1664, usurped the unused pages of a notebook intended for traditional scholastic exercises; under the title he entered the slogan "Amicus Plato amicus Aristoteles magis amica veritas" ("Plato is my friend, Aristotle is my friend, but my best friend is truth"). Newton's scientific career had begun.

Importance of the "Quaestiones"

The "Quaestiones" reveal that Newton had discovered the new conception of nature that provided the framework of the scientific revolution. He had thoroughly mastered the works of Descartes and had also discovered that the French philosopher Pierre Gassendi had revived atomism, an alternative mechanical system to explain nature. The "Quaestiones" also reveal that Newton already was inclined to find the latter a more attractive philosophy than Cartesian natural philosophy, which rejected the existence of ultimate indivisible particles. The works of the 17th-century chemist Robert Boyle provided the foundation for Newton's considerable work in chemistry. Significantly, he had read Henry More, the Cambridge Platonist, and was thereby introduced to another intellectual world, the magical Hermetic tradition, which sought to explain natural phenomena in terms of alchemical and magical concepts. The two traditions of natural philosophy, the mechanical and the Hermetic, antithetical though they appear, continued to influence his thought and in their tension supplied the fundamental theme of his scientific career.

Although he did not record it in the "Quaestiones," Newton had also begun his mathematical studies. He again started with Descartes, from whose *La Géometrie* he branched out into the other literature of modern analysis with its application of algebraic techniques to problems of geometry. He then reached back for the support of classical geometry. Within little more than a year, he had mastered the literature; and, pursuing his own line of analysis, he began to move into new territory. He discovered the binomial theorem, and he developed the calculus, a more powerful form of analysis that employs infinitesimal considerations in finding the slopes of curves and areas under curves.

Development of the calculus

By 1669 Newton was ready to write a tract summarizing his progress, *De Analysi per Aequationes Numeri Terminorum Infinitas* ("On Analysis by Infinite Series"), which circulated in manuscript through a limited circle and made his name known. During the next two years he revised it as *De methodis serierum et fluxionum* ("On the Methods of Series and Fluxions"). The word fluxions, Newton's private rubric, indicates that the calculus had been born. Despite the fact that only a handful of savants were even aware of Newton's existence, he had arrived at the point where he had become the leading mathematician in Europe.

Work during the plague years. When Newton received the bachelor's degree in April 1665, the most remarkable undergraduate career in the history of university education had passed unrecognized. On his own, without formal guidance, he had sought out the new philosophy and the new mathematics and made them his own, but he had confined the progress of his studies to his notebooks. Then, in 1665, the plague closed the university, and for most of the following two years he was forced to stay at his home, contemplating at leisure what he had learned. During the plague years Newton laid the foundations of the calculus and extended an earlier insight into an essay, "Of Colours," which contains most of the ideas elaborated in his *Opticks*. It was during this time that he examined the elements of circular motion and, applying his analysis to the Moon and the planets, derived the inverse square relation that the radially directed force acting on a planet decreases with the square of its distance from the Sun—which was later crucial to the law of universal gravitation. The world heard nothing of these discoveries.

CAREER

The optics. *Inaugural lectures at Trinity.* Newton was elected to a fellowship in Trinity College in 1667, after the university reopened. Two years later, Isaac Barrow, Lucasian professor of mathematics, who had transmitted Newton's *De Analysi* to John Collins in London, resigned the chair to devote himself to divinity and recommended Newton to succeed him. The professorship exempted Newton from the necessity of tutoring but im-

posed the duty of delivering an annual course of lectures. He chose the work he had done in optics as the initial topic; during the following three years (1670–72), his lectures developed the essay "Of Colours" into a form which was later revised to become Book One of his *Opticks*.

Beginning with Kepler's *Paralipomena* in 1604, the study of optics had been a central activity of the scientific revolution. Descartes's statement of the sine law of refraction, relating the angles of incidence and emergence at interfaces of the media through which light passes, had added a new mathematical regularity to the science of light, supporting the conviction that the universe is constructed according to mathematical regularities. Descartes had also made light central to the mechanical philosophy of nature; the reality of light, he argued, consists of motion transmitted through a material medium. Newton fully accepted the mechanical nature of light, although he chose the atomistic alternative and held that light consists of material corpuscles in motion. The corpuscular conception of light was always a speculative theory on the periphery of his optics, however. The core of Newton's contribution had to do with colours. An ancient theory extending back at least to Aristotle held that a certain class of colour phenomena, such as the rainbow, arises from the modification of light, which appears white in its pristine form. Descartes had generalized this theory for all colours and translated it into mechanical imagery. Through a series of experiments performed in 1665 and 1666, in which the spectrum of a narrow beam was projected onto the wall of a darkened chamber, Newton denied the concept of modification and replaced it with that of analysis. Basically, he denied that light is simple and homogeneous—stating instead that it is complex and heterogeneous and that the phenomena of colours arise from the analysis of the heterogeneous mixture into its simple components. The ultimate source of Newton's conviction that light is corpuscular was his recognition that individual rays of light have immutable properties; in his view, such properties imply immutable particles of matter. He held that individual rays (that is, particles of given size) excite sensations of individual colours when they strike the retina of the eye. He also concluded that rays refract at distinct angles—hence, the prismatic spectrum, a beam of heterogeneous rays, *i.e.*, alike incident on one face of a prism, separated or analyzed by the refraction into its component parts—and that phenomena such as the rainbow are produced by refractive analysis. Because he believed that chromatic aberration could never be eliminated from lenses, Newton turned to reflecting telescopes; he constructed the first ever built. The heterogeneity of light has been the foundation of physical optics since his time.

Analysis of colour

There is no evidence that the theory of colours, fully described by Newton in his inaugural lectures at Cambridge, made any impression, just as there is no evidence that aspects of his mathematics and the content of the *Principia*, also pronounced from the podium, made any impression. Rather, the theory of colours, like his later work, was transmitted to the world through the Royal Society of London, which had been organized in 1660. When Newton was appointed Lucasian professor, his name was probably unknown in the Royal Society; in 1671, however, they heard of his reflecting telescope and asked to see it. Pleased by their enthusiastic reception of the telescope and by his election to the society, Newton volunteered a paper on light and colours early in 1672. On the whole, the paper was also well received, although a few questions and some dissent were heard.

Controversy. Among the most important dissenters to Newton's paper was Robert Hooke, one of the leaders of the Royal Society who considered himself the master in optics and hence he wrote a condescending critique of the unknown parvenu. One can understand how the critique would have annoyed a normal man. The flaming rage it provoked, with the desire publicly to humiliate Hooke, however, bespoke the abnormal. Newton was unable rationally to confront criticism. Less than a year after submitting the paper, he was so unsettled by the give and take of honest discussion that he began to cut his ties, and he withdrew into virtual isolation.

In 1675, during a visit to London, Newton thought he heard Hooke accept his theory of colours. He was emboldened to bring forth a second paper, an examination of the colour phenomena in thin films, which was identical to most of Book Two as it later appeared in the *Opticks*. The purpose of the paper was to explain the colours of solid bodies by showing how light can be analyzed into its components by reflection as well as refraction. His explanation of the colours of bodies has not survived, but the paper was significant in demonstrating for the first time the existence of periodic optical phenomena. He discovered the concentric coloured rings in the thin film of air between a lens and a flat sheet of glass; the distance between these concentric rings (Newton's rings) depends on the increasing thickness of the film of air. In 1704 Newton combined a revision of his optical lectures with the paper of 1675 and a small amount of additional material in his *Opticks*.

Colours in thin films

A second piece which Newton had sent with the paper of 1675 provoked new controversy. Entitled "An Hypothesis Explaining the Properties of Light," it was in fact a general system of nature. Hooke apparently claimed that Newton had stolen its content from him, and Newton boiled over again. The issue was quickly controlled, however, by an exchange of formal, excessively polite letters that fail to conceal the complete lack of warmth between the men.

Newton was also engaged in another exchange on his theory of colours with a circle of English Jesuits in Liège, perhaps the most revealing exchange of all. Although their objections were shallow, their contention that his experiments were mistaken lashed him into a fury. The correspondence dragged on until 1678, when a final shriek of rage from Newton, apparently accompanied by a complete nervous breakdown, was followed by silence. The death of his mother the following year completed his isolation. For six years he withdrew from intellectual commerce except when others initiated a correspondence, which he always broke off as quickly as possible.

Influence of the Hermetic tradition. During his time of isolation, Newton was greatly influenced by the Hermetic tradition with which he had been familiar since his undergraduate days. Newton, always somewhat interested in alchemy, now immersed himself in it, copying by hand treatise after treatise and collating them to interpret their arcane imagery. Under the influence of the Hermetic tradition, his conception of nature underwent a decisive change. Until that time, Newton had been a mechanical philosopher in the standard 17th-century style, explaining natural phenomena by the motions of particles of matter. Thus, he held that the physical reality of light is a stream of tiny corpuscles diverted from its course by the presence of denser or rarer media. He felt that the apparent attraction of tiny bits of paper to a piece of glass that has been rubbed with cloth results from an ethereal effluvium that streams out of the glass and carries the bits of paper back with it. This mechanical philosophy denied the possibility of action at a distance; as with static electricity, it explained apparent attractions away by means of invisible ethereal mechanisms. Newton's "Hypothesis of Light" of 1675, with its universal ether, was a standard mechanical system of nature. Some phenomena, such as the capacity of chemicals to react only with certain others, puzzled him, however, and he spoke of a "secret principle" by which substances are "sociable" or "unsociable" with others. About 1679, Newton abandoned the ether and its invisible mechanisms and began to ascribe the puzzling phenomena—chemical affinities, the generation of heat in chemical reactions, surface tension in fluids, capillary action, the cohesion of bodies, and the like—to attractions and repulsions between particles of matter. More than 35 years later, in the second English edition of the *Opticks*, Newton accepted an ether again, although it was an ether that embodied the concept of action at a distance by positing a repulsion between its particles. The attractions and repulsions of Newton's speculations were direct transpositions of the occult sym-

pathies and antipathies of Hermetic philosophy—as mechanical philosophers never ceased to protest. Newton, however, regarded them as a modification of the mechanical philosophy that rendered it subject to exact mathematical treatment. As he conceived of them, attractions were quantitatively defined, and they offered a bridge to unite the two basic themes of 17th-century science—the mechanical tradition, which had dealt primarily with verbal mechanical imagery, and the Pythagorean tradition, which insisted on the mathematical nature of reality. Newton's reconciliation through the concept of force was his ultimate contribution to science.

The Principia. *Planetary motion.* Newton originally applied the idea of attractions and repulsions solely to the range of terrestrial phenomena mentioned in the preceding paragraph. But late in 1679, not long after he had embraced the concept, another application was suggested in a letter from Hooke, who was seeking to renew correspondence. Hooke mentioned his analysis of planetary motion—in effect, the continuous diversion of a rectilinear motion by a central attraction. Newton bluntly refused to correspond but, nevertheless, went on to mention an experiment to demonstrate the rotation of the Earth: let a body be dropped from a tower; because the tangential velocity at the top of the tower is greater than that at the foot, the body should fall slightly to the east. He sketched the path of fall as part of a spiral ending at the centre of the Earth. This was a mistake, as Hooke pointed out; according to Hooke's theory of planetary motion, the path should be elliptical, so that if the Earth were split and separated to allow the body to fall, it would rise again to its original location. Newton did not like being corrected, least of all by Hooke, but he had to accept the basic point; he corrected Hooke's figure, however, using the assumption that gravity is constant. Hooke then countered by replying that, although Newton's figure was correct for constant gravity, his own assumption was that gravity decreases as the square of the distance. Several years later, this letter became the basis for Hooke's charge of plagiarism. He was mistaken in the charge. His knowledge of the inverse square relation rested only on intuitive grounds; he did not derive it properly from the quantitative statement of centripetal force and Kepler's third law, which relates the periods of planets to the radii of their orbits. Moreover, unknown to him, Newton had so derived the relation more than ten years earlier. Nevertheless, Newton later confessed that the correspondence with Hooke led him to demonstrate that an elliptical orbit entails an inverse square attraction to one focus—one of the two crucial propositions on which the law of universal gravitation would ultimately rest. What is more, Hooke's definition of orbital motion—in which the constant action of an attracting body continuously pulls a planet away from its inertial path—suggested a cosmic application for Newton's concept of force and an explanation of planetary paths employing it. In 1679 and 1680, Newton dealt only with orbital dynamics; he had not yet arrived at the concept of universal gravitation.

Correspondence with Hooke

Universal gravitation. Nearly five years later, in August 1684, Newton was visited by the British astronomer Edmond Halley, who was also troubled by the problem of orbital dynamics. Upon learning that Newton had solved the problem, he extracted Newton's promise to send the demonstration. Three months later he received a short tract entitled *De Motu* ("On Motion"). Already Newton was at work improving and expanding it. In two and a half years, the tract *De Motu* grew into *Philosophiae Naturalis Principia Mathematica*, which is not only Newton's masterpiece but also the fundamental work for the whole of modern science.

Significantly, *De Motu* did not state the law of universal gravitation. For that matter, even though it was a treatise on planetary dynamics, it did not contain any of the three Newtonian laws of motion. Only when revising *De Motu* did Newton embrace the principle of inertia (the first law) and arrive at the second law of motion. The second law, the force law, proved to be a precise quantitative statement of the action of the forces between bodies that had become the central members of his system of nature. By quantifying the concept of force, the second law completed the exact quantitative mechanics that has been the paradigm of natural science ever since.

The quantitative mechanics of the *Principia* is not to be confused with the mechanical philosophy. The latter was a philosophy of nature that attempted to explain natural phenomena by means of imagined mechanisms among invisible particles of matter. The mechanics of the *Principia* was an exact quantitative description of the motions of visible bodies. It rested on Newton's three laws of motion: (1) that a body remains in its state of rest unless it is compelled to change that state by a force impressed on it; (2) that the change of motion (the change of velocity times the mass of the body) is proportional to the force impressed; (3) that to every action there is an equal and opposite reaction. The analysis of circular motion in terms of these laws yielded a formula of the quantitative measure, in terms of a body's velocity and mass, of the centripetal force necessary to divert a body from its rectilinear path into a given circle. When Newton substituted this formula into Kepler's third law, he found that the centripetal force holding the planets in their given orbits about the Sun must decrease with the square of the planets' distances from the Sun. Because the satellites of Jupiter also obey Kepler's third law, an inverse square centripetal force must also attract them to the centre of their orbits. Newton was able to show that a similar relation holds between the Earth and its Moon. The distance of the Moon is approximately 60 times the radius of the Earth. Newton compared the distance by which the Moon, in its orbit of known size, is diverted from a tangential path in one second with the distance that a body at the surface of the Earth falls from rest in one second. When the latter distance proved to be 3,600 (60 x 60) times as great as the former, he concluded that one and the same force, governed by a single quantitative law, is operative in all three cases, and from the correlation of the Moon's orbit with the measured acceleration of gravity on the surface of the Earth, he applied the ancient Latin word *gravitas* (literally, "heaviness" or "weight") to it. The law of universal gravitation, which he also confirmed from such further phenomena as the tides and the orbits of comets, states that every particle of matter in the universe attracts every other particle with a force that is proportional to the product of their masses and inversely proportional to the square of the distance between their centres.

When the Royal Society received the completed manuscript of Book I in 1686, Hooke raised the cry of plagiarism, a charge that cannot be sustained in any meaningful sense. On the other hand, Newton's response to it reveals much about him. Hooke would have been satisfied with a generous acknowledgment; it would have been a graceful gesture to a sick man already well into his decline, and it would have cost Newton nothing. Newton, instead, went through his manuscript and eliminated nearly every reference to Hooke. Such was his fury that he refused either to publish his *Opticks* or to accept the presidency of the Royal Society until Hooke was dead.

Charge of plagiarism

International prominence. The *Principia* immediately raised Newton to international prominence. In their continuing loyalty to the mechanical ideal, Continental scientists rejected the idea of action at a distance for a generation, but even in their rejection they could not withhold their admiration for the technical expertise revealed by the work. Young British scientists spontaneously recognized him as their model. Within a generation the limited number of salaried positions for scientists in England, such as the chairs at Oxford, Cambridge, and Gresham College, were monopolized by the young Newtonians of the next generation. Newton, whose only close contacts with women were his unfulfilled relationship with his mother, who had seemed to abandon him, and his later guardianship of a niece, found satisfaction in the role of patron to the circle of young scientists. His friendship with Fatio de Duillier, a Swiss-born mathematician resident in London who shared Newton's interests, was the most profound experience of his adult life.

Warden of the mint. Almost immediately following the *Principia*'s publication, Newton, a fervent if unorthodox Protestant, helped to lead the resistance of Cambridge to James II's attempt to Catholicize it. As a consequence, he was elected to represent the university in the convention that arranged the revolutionary settlement. In this capacity, he made the acquaintance of a broader group, including the philosopher John Locke. Newton tasted the excitement of London life in the aftermath of the *Principia*. The great bulk of his creative work had been completed. He was never again satisfied with the academic cloister, and his desire to change was whetted by Fatio's suggestion that he find a position in London. Seek a place he did, especially through the agency of his friend, the rising politician Charles Montague, later Lord Halifax. Finally, in 1696, he was appointed warden of the mint. Although he did not resign his Cambridge appointments until 1701, he moved to London and henceforth centred his life there.

In the meantime, Newton's relations with Fatio had undergone a crisis. Fatio was taken seriously ill; then family and financial problems threatened to call him home to Switzerland. Newton's distress knew no limits. In 1693 he suggested that Fatio move to Cambridge, where Newton would support him, but nothing came of the proposal. Through early 1693 the intensity of Newton's letters built almost palpably, and then, without surviving explanation, both the close relationship and the correspondence broke off. Four months later, without prior notice, Samuel Pepys and John Locke, both personal friends of Newton, received wild, accusatory letters. Pepys was informed that Newton would see him no more; Locke was charged with trying to entangle him with women. Both men were alarmed for Newton's sanity; and, in fact, Newton had suffered at least his second nervous breakdown. The crisis passed, and Newton recovered his stability. Only briefly did he ever return to sustained scientific work, however, and the move to London was the effective conclusion of his creative activity.

Nervous breakdown

As warden and then master of the mint, Newton drew a large income, as much as £2,000 per annum. Added to his personal estate, the income left him a rich man at his death. The position, regarded as a sinecure, was treated otherwise by Newton. During the great recoinage, there was need for him to be actively in command; even afterward, however, he chose to exercise himself in the office. Above all, he was interested in counterfeiting. He became the terror of London counterfeiters, sending a goodly number to the gallows and finding in them a socially acceptable target on which to vent the rage that continued to well up within him.

Interest in religion and theology. Newton found time now to explore other interests, such as religion and theology. In the early 1690s he had sent Locke a copy of a manuscript attempting to prove that Trinitarian passages in the Bible were latter-day corruptions of the original text. When Locke made moves to publish it, Newton withdrew in fear that his anti-Trinitarian views would become known. In his later years, he devoted much time to the interpretation of the prophecies of Daniel and St. John, and to a closely related study of ancient chronology. Both works were published after his death.

Leader of English science. In London, Newton assumed the role of patriarch of English science. In 1703 he was elected President of the Royal Society. Four years earlier, the French Académie des Sciences (Academy of Sciences) had named him one of eight foreign associates. In 1705 Queen Anne knighted him, the first occasion on which a scientist was so honoured. Newton ruled the Royal Society magisterially. John Flamsteed, the Astronomer Royal, had occasion to feel that he ruled it tyrannically. In his years at the Royal Observatory at Greenwich, Flamsteed, who was a difficult man in his own right, had collected an unrivalled body of data. Newton had received needed information from him for the *Principia*, and in the 1690s, as he worked on the lunar theory, he again required Flamsteed's data. Annoyed when he could not get all the information he wanted as quickly as he wanted it, Newton assumed a domineering and condescending attitude toward Flamsteed. As president of the Royal Society, he used his influence with the government to be named as chairman of a body of "visitors" responsible for the Royal Observatory; then he tried to force the immediate publication of Flamsteed's catalog of stars. The disgraceful episode continued for nearly ten years. Newton would brook no objections. He broke agreements that he had made with Flamsteed. Flamsteed's observations, the fruit of a lifetime of work, were, in effect, seized despite his protests and prepared for the press by his mortal enemy, Edmond Halley. Flamsteed finally won his point and by court order had the printed catalog returned to him before it was generally distributed. He burned the printed sheets, and his assistants brought out an authorized version after his death. In this respect, and at considerable cost to himself, Flamsteed was one of the few men to best Newton. Newton sought his revenge by systematically eliminating references to Flamsteed's help in later editions of the *Principia*.

Difficulties with Flamsteed

In Gottfried Wilhelm Leibniz, the German philosopher and mathematician, Newton met a contestant more of his own calibre. It is now well established that Newton developed the calculus before Leibniz seriously pursued mathematics. It is almost universally agreed that Leibniz later arrived at the calculus independently. There has never been any question that Newton did not publish his method of fluxions; thus, it was Leibniz's paper in 1684 that first made the calculus a matter of public knowledge. In the *Principia* Newton hinted at his method, but he did not really publish it until he appended two papers to the *Opticks* in 1704. By then the priority controversy was already smouldering. If, indeed, it mattered, it would be impossible finally to assess responsibility for the ensuing fracas. What began as mild innuendoes rapidly escalated into blunt charges of plagiarism on both sides. Egged on by followers anxious to win a reputation under his auspices, Newton allowed himself to be drawn into the centre of the fray; and, once his temper was aroused by accusations of dishonesty, his anger was beyond constraint. Leibniz's conduct of the controversy was not pleasant, and yet it paled beside that of Newton. Although he never appeared in public, Newton wrote most of the pieces that appeared in his defense, publishing them under the names of his young men, who never demurred. As president of the Royal Society, he appointed an "impartial" committee to investigate the issue, secretly wrote the report officially published by the society, and reviewed it anonymously in the *Philosophical Transactions*. Even Leibniz's death could not allay Newton's wrath, and he continued to pursue the enemy beyond the grave. The battle with Leibniz, the irrepressible need to efface the charge of dishonesty, dominated the final 25 years of Newton's life. It obtruded itself continually upon his consciousness. Almost any paper on any subject from those years is apt to be interrupted by a furious paragraph against the German philosopher, as he honed the instruments of his fury ever more keenly. In the end, only Newton's death ended his wrath.

Priority dispute with Leibniz

Final years. During his final years Newton brought out further editions of his central works. After the first edition of the *Opticks* in 1704, which merely published work done 30 years before, he published a Latin edition in 1706 and a second English edition in 1717–18. In both, the central text was scarcely touched, but he did expand the "Queries" at the end into the final statement of his speculations on the nature of the universe. The second edition of the *Principia*, edited by Roger Cotes in 1713, introduced extensive alterations. A third edition, edited by Henry Pemberton in 1726, added little more. Until nearly the end, Newton presided at the Royal Society (frequently dozing through the meetings) and supervised the mint. During his last years, his niece, Catherine Barton Conduitt, and her husband lived with him. He died on March 20 (March 31, N.S.), 1727, in London.

MAJOR WORKS

Philosophiae Naturalis Principia Mathematica (1687; *Mathematical Principles of Natural Philosophy*, 1729); *Opticks* (1704); *Arithmetica Universalis* (1707; *Universal Arithmetick*, 1720); *The Chronology of Ancient Kingdoms Amended*

(1728); *Observations Upon the Prophecies of Daniel and the Apocalypse of St. John* (1733).

BIBLIOGRAPHY. The standard biography of Newton is DAVID BREWSTER, *Memoirs of the Life, Writings, and Discoveries of Sir Isaac Newton*, 2nd ed., 2 vol. (1855). A more modern work by LOUIS T. MORE, *Isaac Newton: A Biography* (1934), does not succeed in replacing Brewster. FRANK MANUEL, *A Portrait of Isaac Newton* (1968), offers a fascinating Freudian analysis of him. The best general treatment of the major problems in Newtonian science is found in I. BERNARD COHEN, *Franklin and Newton* (1956). Two recent works have explored the development of Newton's mechanics: JOHN W. HERIVEL, *The Background to Newton's Principia* (1966); and RICHARD S. WESTFALL, *Force in Newton's Physics* (1971). I. BERNARD COHEN, *Introduction to Newton's Principia* (1971), a history of the development and modification of Newton's major work, is the first volume of Cohen's edition of the *Principia* with variant readings. A.I. SABRA, *Theories of Light from Descartes to Newton* (1967), is the leading authority on Newton's work in optics. ALEXANDRE KOYRE, *Newtonian Studies* (1965), contains a collection of essays by one of the master historians of science. In addition to Cohen's edition of the *Principia*, there are collections of Newtonian materials, all with valuable introductory essays: *The Mathematical Papers of Isaac Newton*, ed. by D.T. WHITESIDE, 5 vol. (1967–72, in progress); *Unpublished Scientific Papers of Isaac Newton*, ed. by A.R. and M.B. HALL (1962); *Isaac Newton's Papers & Letters on Natural Philosophy and Related Documents*, ed. by I. BERNARD COHEN (1958); and *The Correspondence of Isaac Newton*, ed. by H.W. TURNBULL and J.F. SCOTT, 4 vol. (1959–).

(R.S.W.)

New York (State)

One of the 13 original states of the United States, New York was until the 1960s the nation's first state in nearly all population, cultural, and economic indexes. Its displacement by California around mid-decade was caused by the enormous growth rate that had persisted on the West Coast for several decades, rather than by any decline by New York. Although the state's percentage growth between the 1960 and 1970 censuses was well below the national average, it gained more people—to more than 18,000,000 inhabitants—than any of the states except California, Florida, and Texas, and its gross economic product exceeded that of all but a handful of nations in the world.

An overview of the state

This great population and economic base exists in a region of substantial contrast—from the Atlantic Ocean shores of Long Island and the skyscrapers of Manhattan through the rivers, mountains, and lakes of Upstate New York to the plains of the Great Lakes. By means of this pathway from ocean to inland seas, the canals, railroads, and highways of New York made it a principal gateway to the west for the Middle Atlantic and New England states and a continuing point of union for much of the nation. Along the way grew the cities—from New York City through Albany (the state capital), Utica, and Syracuse to Buffalo and Rochester on the lakes—that, with their suburbs, today hold more than four-fifths of all New Yorkers.

New York's central role in the development of the American nation was a slow-maturing phenomenon, and both the New England and Southern colonies had more to do with the movement toward revolution and with stabilizing the new nation in its early decades. Once under way, however, New York's growth attained a breakneck pace. Today, the state—and New York City, in particular—remains the focus of much of the nation's economy and finance, as well as of many formative impulses in U.S. arts and culture. The overwhelming presence of New York City actually has tended to divide the state socially and politically, causing long-standing problems for both city and state, but the influence and image of the state is a major element of national political life.

The 49,576 square miles (128,401 square kilometres) of New York are bounded, from west to north, by Lake Erie, the Canadian province of Ontario, Lake Ontario, and Quebec Province; on the east by the New England states of Vermont, Massachusetts, and Connecticut; on the southeast by the Atlantic Ocean and New Jersey; and on the south by Pennsylvania. (For information on related topics, see the articles UNITED STATES OF AMERICA; UNITED STATES, HISTORY OF; AMERICAN REVOLUTION; NEW YORK (CITY); NORTH AMERICA; APPALACHIAN MOUNTAINS; GREAT LAKES; and NIAGARA RIVER AND FALLS.)

THE HISTORY OF NEW YORK

Two major groupings of Indian tribes were living in the region when the white man arrived: the Mohegan (Mohican) and Munsee tribes of the Algonkian family near the coast and, inland, the five tribes of the Iroquois—Mohawk, Oneida, Onondaga, Cayuga, and Seneca—which formed the League of Five Nations about 1570. This confederacy, with advanced social and governmental institutions, reached the height of its power around 1700. When these tribes later aligned themselves with the British against the French and the Algonkians, they probably provided the balance of power needed for the British to emerge victors in the nearly 150 years of struggle between the two European powers in North America.

Settlement and colonial period. New York was settled originally as a colony of the Netherlands following Henry Hudson's exploration in 1609 of the river later named for him. In 1624 the Dutch established Fort Orange at present-day Albany as the first permanent European settlement in New York. One year later, a similar colony, New Amsterdam, was established at the foot of Manhattan Island. To legalize the settlement, Peter Minuit, Dutch governor, paid the Indians 60 Dutch guilders—about $24—in the form of merchandise. Although the Dutch established several settlements along the Hudson, their interest was more in trade than in permanent agricultural development. Thus, while these trading posts prospered and aided the general expansion of the empire of the Netherlands, no deep roots of permanent colonization were planted. The likely explanation for this lies in the general economic prosperity and social stability of the homeland. The Dutch citizens had no strong economic motivations to move overseas, nor were there sufficient religious quarrels at this time to promote such movement. When an English fleet sailed into New York harbour in 1664, Peter Stuyvesant, the governor, surrendered without a fight. Although controversy ensued for several years, the colony was clearly in English hands by 1669. It was renamed New York after the Duke of York.

Slowness of growth in colonial period

Despite this change in ownership and sovereignty, however, the colony did not develop rapidly. Like the Dutch, the English crown granted large tracts of land to private individuals. This system of landownership was not attractive to settlers such as the farmer-colonists who had settled New England, and agricultural development, particularly in the Hudson Valley, remained slight.

Further, the European war between France and England had its counterpart in North America. The French, established along the St. Lawrence and in Quebec, made many forays into northern and central New York. The relatively strong Five Nations federation of the Iroquois aligned itself with the English in New York and New England because of earlier French aid to the rival Algonkians. This warfare discouraged settlement beyond Albany. The military situation was brought to a conclusion by the Treaty of Paris in 1763, which confirmed English dominance of the New York region. A gradual but steady movement of settlers from New England began New York's population explosion. The New Englanders moved across the borders of Connecticut and Massachusetts, some remaining on the east bank of the Hudson, others passing through Albany to the interior.

In 1698 the colony's population was about 18,000, two-thirds of it in and around New York City. By the eve of the American Revolution, however, the numbers had grown to 163,000, and the concentration was nearly exactly reversed; but New York still ranked only seventh among the colonies. The cosmopolitan and heterogeneous character of its population was already well established. Dutch culture remained strong in New York City and Albany, while most of the settlements in the interior possessed—and possess to this day—a flavour and dialect of the New England Yankee, with the addition of several German communities. This emerging pattern of cultural

heterogeneity was to have a considerable impact on the politics of the state, as were the waves of immigration from Europe that followed the war and continued well into the 20th century.

Revolution, statehood, and growth. New York contains many of the battlegrounds of the American Revolution, suffering invasions from both north and east. The war in New York took on many of the characteristics of a civil war, since the state probably had a higher proportion of persons loyal to the crown than did any other.

After the war, a part of the state's political leadership aligned itself with like-minded leaders from other colonies to urge establishment of a strong central government for the new nation rather than the loose confederation then existing. New York delegates participated vigorously in the Constitutional Convention, one of the leaders of which was New Yorker Alexander Hamilton. Despite the role of Hamilton and other New York delegates in drafting the new document, the politics of ratification within the state legislature were intense and bitter; and New York was the 11th state to endorse the Constitution.

Improvements in transportation, and the movement westward

Both the American Revolution and the War of 1812 interrupted New York's expansion westward, but, thereafter, the movement began in earnest. Turnpikes spread westward from Albany and other points along the Hudson River, and settlers spread across the state. The opening of the Erie Canal in 1825 confirmed New York's position as the gateway to the west from the Atlantic Coast. Railroads followed in quick order and tended to follow the pattern of trade established by the turnpikes and the canal.

By 1800 New York state had become the second largest state in the Union, trailing only Virginia, and 10 years later it had surpassed all other states. Its leadership was not only in population, size, and growth but also in manufacturing, trade, and transportation—and in the increasing heterogeneity of its population.

Growth and change were reflected as well in the political and governmental history of the state. The original state constitution restricted the suffrage to property holders and established a governing system dominated by large property holders and leading commercial interests. The change in population composition, as well as shifting political attitudes in the nation, soon caused New York to move in a more democratic direction. In the 1830s, a vigorous campaign was launched against the system of landownership in the Hudson Valley, with renters eventually being given the opportunity to own the land they tilled. The constitutional convention of 1846 confirmed these democratic moves by expanding the suffrage and restricting the power of both legislature and governor.

Emergence of political divisions. Over the next century and more, New York grew in virtually every measurable dimension, but its political development centred on the increasing chasm of interest and affection between New York City and Upstate. The issue of "home rule," the demands of the city for total powers of self-government against claims that the city was the creature of the state, remained central to the conflict that continued into the 1970s.

As early as the 1780s, an organization eventually to be known as Tammany Hall was formed in the city to combat attempts by propertied Revolutionary leaders to limit the franchise in the new state. Largely middle class in membership, it did not extend its democratic principles to the lower classes or immigrants. By mid-19th century, however, through workers' and equal-rights parties, Irish politicians dominated the organization and the office of mayor, culminating in the control of the Democratic machine after 1868 by William "Boss" Tweed.

Well into the 20th century, the name Tammany was an international byword for municipal corruption at the highest levels. City–state antagonism was fuelled by Democratic domination in the city and Republican domination of areas Upstate and, in most years, of the statehouse and legislature as well. Investigations of Tammany Hall and city politics in general were highlighted by those of the Seabury Commission (1931–32), which brought about the resignation of Mayor James Walker and led to the reform administration of Fiorello H. La Guardia (1933–45) and the efforts of subsequent mayors to tread the line between Tammany's power in municipal elections and an image of political incorruptibility.

Much of Tammany Hall's power was based on its social services to the waves of immigrants that inundated New York City until changes in immigration laws slowed the tide in the 1920s. When state and federal government began to take over such services as social security, workmen's compensation, and unemployment, welfare, and health benefits, notably during the Depression of the 1930s, Tammany's hold began slowly to erode. By the 1970s the state and city remained in difficult confrontation no matter what the party of mayor and governor, though ironically home rule and increased state financial assistance had become joint demands. It seemed likely that the focus would remain on the need for the state to provide a growing proportion of services.

THE NATURAL AND HUMAN LANDSCAPE

Although to many non-New Yorkers—and actually to many residents of the New York City metropolitan area as well—New York state and New York City are synonymous, the state has a great range of geographical and climatic conditions. During at least part of the Ice Age New York was covered almost entirely by glaciers, the only exceptions being southern Long Island, all of Staten Island, and the far southeastern corner of the state.

Geological regions. The movement of the glaciers over long time periods left New York with nine distinct geological regions and, within these, 28 subregions. Each possesses its own landform, with distinctive geological structures and patterns of erosion. In the northeast, the Adirondack upland is characterized by the highest and most rugged mountains in the state, reaching 5,344 feet (1,629 metres) in Mt. Marcy and 5,114 feet (1,559 metres) in Algonquin Peak of Mt. McIntyre. Except for some forestry activities, the region is without economic value other than for recreation. A large part of it has been designated as a wilderness preserve by the state.

The St. Lawrence–Champlain Lowlands extend northeastward from Lake Ontario to the ocean along the boundary with Canada. Within this area are three subdivisions: the St. Lawrence marine plain, a flat to gently rolling strip along the St. Lawrence River; the St. Lawrence hills south of this plain; and, farther south, the Champlain Lake Plain.

The routes of settlement and commerce

Another region, the Hudson–Mohawk Lowland, follows the Hudson River from New York City to Albany then turns west along the Mohawk River. The Hudson Valley portion, between the Catskill Mountains on the west and the Taconic Range on the east, is from ten to 20 miles wide; the Mohawk Valley portion is ten to 30 miles wide. These routes provided easy access from both New York City and New England into the hinterland of the state and formed the natural paths for canal, railroad, and highway. Cutting natural pathways through the mountains of central and western New York, these rivers became the state's major avenues of commerce, serving first as the basis of the Erie Canal and later as the route of the New York Central Railroad and of the New York State (now Thomas E. Dewey) Thruway.

To the east of the Hudson River lies the New England Upland, extending eastward into Massachusetts and Connecticut and southward across the Lower Hudson Valley into Pennsylvania.

Two small regions complete the geologic picture in southeastern New York. The Atlantic Coastal Plain, which extends from Massachusetts to Florida, takes in Long Island and Staten Island. A small finger of the eastern Piedmont region juts up from New Jersey for some distance along the west bank of the Hudson.

The Appalachian Highlands

The largest region in New York, the Appalachian Highlands, occupies about one-half of the state, extending westward from the Hudson Valley to the state's southern and western boundaries. Located within it are the Catskill Mountains, the peaks of which reach 2,000 to 4,000 feet, the Finger Lakes hills area, and the Delaware Basin. The Catskills, with their mountains and lakes, are primarily a

recreation area. The Finger Lakes area also provides many opportunities for summer and winter sports, and its valleys provide excellent grasslands for dairying. The Delaware Basin, drained by the Delaware River, is a mixed-farming area.

Lying to the north of the Appalachian Upland and to the west of the Mohawk Valley and extending along the southern shores of the Great Lakes is a plateau-like region known as the Erie–Ontario Lowlands. It is composed of lake plains bordering the Great Lakes that extend between five and 30 miles inland from the lakes. Because of the moderating influence of the lakes on the weather, the region has become an important fruit-growing area. Finally, between the lake lowlands and the western reaches of the Adirondacks and north of Oneida Lake lies the Tug Hill Upland, which is one of the least settled parts of the state because of its poor soil and drainage and its severe climatic conditions.

Waters. Among New York's special geographical features are its two major shorelines: 127 miles bordering the Atlantic, 371 miles on Lakes Erie and Ontario. In addition, it has some 8,000 lakes and nine major rivers. The Hudson and Mohawk rivers have played the most significant roles in the state's history, but the Genesee and Oswego, flowing northward into Lake Ontario, also have been important. The Delaware, Susquehanna, and Allegheny drain the southern and western portions of the state and provide a large part of New York City's water supply. Connecting Long Island Sound with New York Bay and separating Long Island and Manhattan is the tidal estuary known as the East River. The most dramatic of the waterfalls that dot the state is Niagara Falls, a source of much hydroelectric power as well as one of the chief scenic attractions of the Northeast.

Climate. New York's climate proved a great disappointment to the early Dutch settlers. Since Manhattan was actually Mediterranean in latitude, it was rather bewildering to them to find that "it freezes and snows severely in winter." If Manhattan was uncomfortably cold and wet in the winter months, the rest of the state must have been an even greater disappointment.

The mean monthly temperature ranges from a high of 54° F (12.2° C) in New York City's Central Park to a low of 40° F (4.4° C) at Lake Placid in the Adirondacks. Average August temperatures range from 73° F (23° C) in New York City to 62° F (17° C) at Indian Lake in the Adirondacks; in February, from 33° F (0.5° C) on Long Isand to 14° F (−10° C) at Stillwater Reservoir in the Adirondacks. These figures represent the extremes, but substantial differences exist between New York City and Upstate Albany, Buffalo, Rochester, and Syracuse. A tendency to cloudiness across the state produces few completely clear days. New York City has about 100 such days during the average year, Syracuse and Buffalo have 72, Binghamton 68, and Albany 73.

The growing season also varies substantially: Long Island has the longest, with 200 days, while in the extreme north central region there are only 85. Rainfall and melted snow produce a range of precipitation from 32 to 45 inches a year, with the coastal plain receiving the greatest amount of precipitation, while the Erie–Ontario lowland receives the least. Nearly all parts of the state, however, receive sufficient rainfall for adequate crop growing, with the occasional exception being parts of the Erie–Ontario lowland. The Buffalo region receives an unusual amount of snow since it is on the eastern shore of Lake Erie.

The human regions. The cultural and social distinctions among various parts of New York state have tended to decline. Upstate cities, for example, are nearly as ethnically varied as New York City. Certain cultural and social characteristics brought by early settlers remain visible and, to some degree, still influence life-styles. During the Colonial period and for a number of years after the American Revolution, New England was a major source of immigrants; and traces of the New England influence are still present, particularly in the architecture and small-town planning of the north shore of Long Island and in northern Westchester County. The Dutch influence around Albany is now little more than place-names and street names, plus some preserved or rehabilitated Dutch architecture. German and Scottish settlers have left their mark in the Schoharie Valley; parts of the Hudson and Mohawk valleys (German); and in Orange and Ulster counties and in the Cherry Valley area (Scots).

Diversity of ethnic, national, and regional influences

The basic distinction between Upstate and Downstate is normally related to political differences—Upstate, conservative; Downstate, liberal. These political differences are matched by and interact with social ones. Downstate must be divided between New York City and the suburbs, and, within the city, differences among the boroughs are significant. Manhattan, although containing many low-income residents, is the centre for sophisticated life-styles and liberal politics, the home of the "limousine liberals." The outer boroughs are characterized by relatively stable ethnic neighbourhoods and "communities in transition," more conservative in social attitude but oriented to the Democratic Party. The suburbs are dominated by white middle- and high-income families living in detached houses, though in recent years the income spread in the suburbs has increased, and the inner suburbs are beginning to resemble the city's outer boroughs.

The rural Upstate areas must be distinguished from the Upstate cities and their suburbs. Rural New York remains conservative both politically and socially. The city regions vary from relatively sophisticated Rochester, with its heavy concentration of white-collar technical and managerial employees, to the more conservative Syracuse–central New York area. Buffalo, with its emphasis on heavy industry, has a large blue-collar population.

THE PEOPLE OF NEW YORK

The growth of the population. Since the Colonial period, much of New York's growth has resulted from immigration, both from other states and from abroad. Before the American Revolution, the Dutch, English, Scots, and Germans were the primary settlers, to be followed in the first half of the 19th century by New Englanders spreading across developing parts of Upstate and into Westchester and northern Long Island. The influx of European immigrants came first from the northern and central parts of the Continent, later from the southern countries.

Patterns of immigration

In the early 1970s, nearly one out of eight New Yorkers was foreign born, and 33 percent of the population was either foreign born or had one or both parents born abroad. The nations of origin of this stock are, in order, Italy, the Soviet Union, Poland, Germany, Ireland, the United Kingdom, and Canada, while nearly 1,000,000 persons are from a great diversity of other countries.

Related to this concentration of foreign origin is the state's religious composition. About one-third of the population is Roman Catholic, while 15 percent is Jewish.

The growth of the non-white portion of the population in the 20th century is of particular significance. The first large-scale influx of blacks from Southern states occurred during World War I, but it was small compared to that during and following World War II. In 1940, only 4.4 percent of the population was non-white, but by 1970 the percentage had increased to 12.9. The non-white population is concentrated in the state's metropolitan areas and, within those areas, in the central cities. In 1970 the non-white percentage reached 18 percent in the New York City metropolitan area, 29 percent in the city's borough of Manhattan. Other urban areas show smaller percentages, the Buffalo area, at 8.8, having the highest.

The non-whites

Another immigrant group that has had a significant impact on the economy and culture of New York since World War II is the Puerto Ricans. In 1970 over 870,000 Puerto Ricans resided in the state, over 90 percent of them in New York City. Following a heavy immigration of Puerto Ricans in the 1950s and early 1960s, the growing economic strength of Puerto Rico caused a considerable reduction, with those entering the state being largely offset by those returning to Puerto Rico.

Population concentration and movement. Overall, New York's population density in the early 1970s—380 people per square mile, compared to the national average of 58—is exceeded only by New Jersey, Rhode Island, Massachusetts, Connecticut, and Maryland. The distribu-

tion is uneven, with densities much higher in urban areas. In New York City, the density was 26,316; and for the entire New York City metropolitan area, 5,415.

The population of the state's seven metropolitan areas accounted for 87 percent of all New Yorkers, with 64 percent in the New York City metropolitan area. In line with national trends, most of the cities continued to lose population to surrounding suburban areas during the 1960s. The slight increase in New York City was concentrated in the outlying, suburb-like boroughs of Queens and Richmond (Staten Island), whereas both Manhattan and Brooklyn lost residents.

This shifting, far from being random, represents a sorting out of the population, with middle- and higher-income whites moving to the suburbs, leaving low-income groups and blacks within the central cities. Many economic activities, particularly manufacturing and headquarters of corporations, have moved to the suburbs as well. This movement of people and economic activity resulted in the urban crisis that was becoming familiar across the United States in the 1960s and 1970s: the increasing needs of the cities to combat crime and other symptoms of the poverty concentrated within them at the same time that their social and economic resources to do so were being removed. While the total economic strength of the metropolitan areas was growing rapidly, the cities were increasingly unable to participate in the prosperity and seemed likely to slip increasingly further behind.

THE STATE'S ECONOMY

New York's economy possesses characteristics similar to those of the other Northeastern states. The economies of other states—in the South, the Middle West, and the Far West—are growing more rapidly than the Northeast, but this more mature region still possesses great economic strength. New York has a complex network of nearly every form of transportation. Its resources of electrical power for domestic and commercial use are enormous, including conventional coal- and oil-burning plants, hydroelectricity from the Niagara region, and a small but growing nuclear source; however, at least 10,000,000,000 kilowatt-hours must be imported from other states.

Components of the economy. New York is represented in every economic category designated by the federal Bureau of the Census. In comparison to the rest of the United States, New York's economy has a disproportionately large share of the country's economy in the fields of wholesale–retail trade, finance–insurance–real estate, transportation–communications–public utilities, and services; it is under-represented in farming, mining, manufacturing, construction, and government. Even within these latter fields, it is over-represented in manufacturing, for example, in instruments production (three times the national average), wearing apparel, leather and leather products, and printing, publishing, and related industries.

Economic strengths and weaknesses

In spite of these strengths, New York, like the nation as a whole, has a declining proportion of its population engaged in manufacturing. In 1947, more than one-third of its employed population were in manufacturing, while by 1969 the figure had shrunk to about one-fourth. Between 1958 and 1970, actual manufacturing employment in New York increased only 1 percent, whereas the nation had a 25 percent growth. Thus, New York leads the nation in this fundamental economic change, which will find a decreasing portion of the nation's work force engaged in manufacturing. The growth sectors of the economy are nonmanufacturing ones. Employment increases in the 1960s came in finance–insurance–real estate, in services, and in government, approximately doubling wage and salary disbursements in these fields.

Areas of economic specialization. Some economic specialization exists within different parts of the state. Manufacturing, for example, is more highly concentrated in the Upstate metropolitan areas than in the New York City area, whereas such activities as services and finance–insurance–real estate are much larger in the New York City metropolitan area than Upstate. Further specialization exists among the Upstate areas. Buffalo is strong in heavy industry, while Rochester, with a national dominance in the manufacture of photographic and optical equipment, is primarily responsible for the state's strong position in instrument production.

Syracuse is substantially over-represented, relative to both the nation and the rest of New York, in the production of primary metals, machinery, and paper and allied products, as well as in educational employment. The Utica–Rome area specializes in machinery and transportation equipment, while the Albany–Troy–Schenectady area is strong in the field of paper and allied products. Albany, as the state capital, leads in government employment. The state's smallest metropolitan area, Binghamton, is the original site of the International Business Machine Corporation (IBM) and thus has a concentration of employment in the computer and business-machine field.

Incomes. In 1970, per capita income of New York workers was 22 percent above the national figure and about 13 percent above that in neighbouring North Atlantic states. The income growth rate increased significantly during the 1960s, no doubt a reflection of unusually high inflationary trends. This growth rate is about the same in the New York City area and Upstate; in both regions, personal incomes doubled between 1950 and 1970. Figures for 1950 and 1968 show also that the relative positions of the seven metropolitan areas in regard to percent share of the economy was virtually identical: in order, New York City, Buffalo, Rochester, Albany–Schenectady–Troy, Syracuse, Utica–Rome, and Binghamton. Nonmetropolitan areas accounted for about one-third of Upstate New York's economic incomes; just over 3 percent of income in such areas was from farms, in contrast to a figure of just over 0.5 percent statewide.

Economic management. The state plays both regulatory and promotional roles in the economy. The Public Service Commission controls the rates charged by the public utilities, and the Urban Development Corporation encourages the production of housing and other community facilities. The Department of Commerce aids in attracting new economic activity to the state, providing information and assistance to industries seeking to locate there, giving some financial support to local communities interested in developing industrial parks, and offering other incentives to encourage the location of new industries within such areas.

The state's role in the economy

New York tends to have somewhat lower unemployment rates during downturns in the national economy than does the rest of the nation, but it also recovers less rapidly. This behaviour results in large measure from the state's economic mix and its heavy dependence on nonmanufacturing activities.

New York has one of the most highly unionized work forces of any state. The most rapidly growing unionization is in the service sector, particularly among such government workers as teachers, sanitation men, police, and firemen. The nature of labour–management relations varies considerably from industry to industry, with workers in construction and the garment and apparel industries wielding great power. Nearly every session of the state legislature devotes attention to the field of labour relations, particularly public-sector employee relations, in which no satisfactory means of settling disputes has yet emerged. Simple "no-strike" statutes do not work, and the arbitration devices used thus far remain unsatisfactory to both workers and management.

As the country's economy shifts increasingly from manufacturing to nonmanufacturing, the strength of New York in service and related activities will continue to serve it well. Although definite economic problems exist in various parts of the state, primarily in its central cities, the overall economic outlook is good.

Transportation. Underlying the economic activity of New York state is its transportation system. New York's early economic advantages were themselves a product of its location relative to the country's natural transportation routes, and its current position remains heavily dependent on their continuing to tie together the centres of population within and without the state.

The Erie Canal, opened in 1825, tied New York City and its port to Buffalo and the westward-expanding na-

Waterways and railways

tion. The main railroad system followed the route of the canal, with feeder lines jutting north and south into the rest of the state. Following World War II, the limited-access Thomas E. Dewey Thruway stretched from New York to the Pennsylvania state line, passing through Albany, Utica, Syracuse, Rochester, and Buffalo. The basic paths of these main transportation routes are not substantially different from those trod by the state's original settlers.

With the completion in 1918 of the New York State Barge Canal System, which incorporated the old Erie Canal, New York had the country's most extensive inland-waterway system. Although still an important means for moving goods—particularly petroleum products, which provide nearly half the tonnage—the present annual tonnage carried on the system is a considerable drop from earlier decades.

The railways first challenged the supremacy of the canal. Beginning in the mid-19th century with the establishment of the New York Central Railroad, a system was built that tied New York's major cities to Chicago, Boston, Montreal, and other urban centres. Although declining in the number of passengers carried, the railroads remain important handlers of freight. Much of this freight originates at the Port of New York, still the largest port in the United States, handling in 1970 approximately 15 percent of all the nation's imports. Nearly one-half of all passengers to and from the United States pass through this port.

Highways

Central to the highway system are the limited-access highways. The Thruway connects at Albany to the Northway, which extends northward to Canada. In central New York, a major highway runs from the Pennsylvania state line to Canada, passing through Binghamton, Syracuse, and Watertown. At Syracuse, this route intersects with the Thruway, causing the city to remain a transportation hub and accounting, in large part, for its economic viability. Another limited-access expressway extends across the southern tier of the state. On Long Island a set of east–west highways ties the entire island to New York City, New England, and Upstate New York.

The most complex commuting system exists in the New York metropolitan area, with its combination of subways, buses, and railroads. The nearly 800-mile-long New York transit system provides intra-city passenger transport. Commuter railroads serve suburban Long Island, New Jersey, Connecticut, and Westchester County. Many of these transportation networks were brought under the control of a single agency, the New York Metropolitan Transportation Authority, in 1968.

In 1971, New York had 446 airfield facilities, including 20 major municipal airports. The three largest are in the New York City metropolitan area—John F. Kennedy International, La Guardia, and Newark (New Jersey)—with a combined total of nearly 14,000,000 passenger enplanements in 1970.

ADMINISTRATION AND SOCIAL CONDITIONS

Structure of government. New York's constitution prescribes the distribution of powers among the branches of state government as well as the system of local government throughout the state. As is the case in some other states, however, the document is excessively detailed, including provisions that most constitutional scholars consider more appropriately treated in legislative statutes than in a constitution. Because of this detail, articles tend to become outdated rapidly, requiring fairly frequent conventions for constitutional revision. Since the first convention in 1777, eight others have been held at roughly 20- to 25-year intervals, the last in 1967. The present document requires that the question of holding a convention be placed before the voters every 20 years.

The constitutional framework

Shifting constitutional forms. The first constitution established a bicameral legislature and provided for the first popularly elected governor in the United States, but it restricted the suffrage to male property holders. Veto power over legislation was vested in a Council of Revision, comprising the governor and justices of the Supreme Court, while the appointment of nearly all state and local officials lay with the Council of Appointment, comprising the governor and four senators. These councils represented an effort to avoid the autocratic rule that New Yorkers had experienced at the hands of the colonial governors appointed by the British crown. The convention of 1821 abolished the bodies, extended the franchise, and introduced a formal bill of rights.

The convention of 1846, influenced heavily by the populist spirit of Jacksonian democracy, imposed financial limitations on the legislature, which often had extended state credit to private ventures in such areas as railroad and canal building. Many state offices, including the judiciary, were made elective.

The constitution of 1894 remains New York's basic law today, but it has been amended more than 160 times. Its new provisions included a merit civil-service system, limitations on disposal of the state's forest preserve, a commitment to public education, and the first constitutional definition of state–local relations.

Earlier and later conventions produced documents that were rejected by the voters or sets of amendments reflecting the spirit of the times. If not accepted, such amendments often were introduced into law through legislation. The conventions of 1938 and 1967 were concerned primarily with social issues—of welfare, labour, and other rights during the Depression years, of aid to parochial schools and the solution of urban social and economic crises more recently. Whereas the voters accepted most of the 1938 amendments, they soundly rejected the 1967 document.

Contemporary executive, legislature, and judiciary. Emerging from these various constitutional revisions and amendments is a state government characterized by a strong governorship based on power over appointments and budget. The governor normally has the upper hand in contests with the legislature.

He is somewhat restricted in his executive authority, however, by a number of independently appointed or elected officials. The Board of Regents, for example, which presides over the education establishment of the state, is appointed by the legislature. An independently elected comptroller acts as auditor for both state and local governments.

The legislature, found recently by an independent analysis of legislative structure and process to be the second most effective in the country, comprises a Senate of 58 members and an Assembly of 150. Each house has standing committees concerned with broad issues of public policy. Several committees composed of both senators and assemblymen study specific policy issues and make recommendations to the legislature. The state also uses numerous nonlegislative commissions—appointed by the governor, by the legislature, or by both—to investigate such problems as education-aid formulas, state–local relations, the judicial process, welfare administration, and basic governmental organization.

Legislative and paralegislative committees

For judicial purposes, New York is divided into 14 districts. Called the Supreme Court, each district has several elected judges. Four judicial departments act as appeal divisions from the Supreme and inferior courts. The highest court is the Court of Appeals. The governor appoints the judges to the appellate departments from those elected to the Supreme Court, while those serving on the Court of Appeals are elected for 14-year terms. At the state level, the Court of Claims hears cases against the state. A variety of local bodies functions, including county courts and the court system of New York City. Other local courts include family and surrogate courts.

Local government. Much legislative debate revolves around the allocation of state aid to local jurisdictions. The constitution has contained a home-rule provision since 1896, but court interpretations of the provision, which gives the state the power to act in any matter in which there is a state concern, have tended to weaken the home-rule concept and continue Albany's domination of local governments. Moreover, the increasing interdependence of the state and its parts caused by metropolitanization and industrialization inevitably has reduced the autonomy of local jurisdictions.

Local jurisdictions consist of five types of governments; counties (62), towns (well over 900), villages (over 550), cities (62), and special districts (almost 1,000). The entire state is covered by county jurisdictions, which are divided into towns. Urban areas may be incorporated as either cities or villages. When it incorporates as a city, the town jurisdiction is eliminated, but villages remain a part of the town in which they are located and residents pay town as well as village taxes.

Unlike states in which either town or county government is weak, New York has strong jurisdictions of both types. This situation often leads to considerable overlapping in the provision of governmental services outside of cities. Special districts include those for schools (encompassing the entire state) and such urban services as sewers, water, and streetlighting. The largest regional district is the Port of New York Authority, which operates bridges, harbours, and related facilities throughout the New York City metropolitan area, including those in northern New Jersey.

The decision of the United States Supreme Court in 1965 requiring legislative districts to be roughly equal in population brought new life into New York's county government, since town supervisors were no longer able to have dual responsibility as county supervisors as well. Many counties, including the most urban outside New York City, have opted for single-executive systems.

Cities and villages generally are governed by a mayor and a council, with only a few cities, the largest being Rochester, using the city-manager plan. Some of the larger cities have a second legislative body, often called the Board of Estimate. In New York City, the mayor, the president of the City Council, the comptroller, and the five borough presidents serve on this body. In other cities, membership usually includes the mayor, the president of the city council, and one or more high-ranking fiscal officers.

State financial aid to local governments

The state–local governing system of New York places heavy responsibilities on local governments. To help them function, more than one-half of the state budget consists of aid to local government. Nearly one-half of this is for public schools; other major aid categories include welfare, health, highways, and housing and urban renewal.

Current debate indicates that the state will soon be playing a larger and perhaps exclusive role in financing education, perhaps using a statewide property tax for part of the revenue. If the federal government assumes a much larger share in financing welfare, the local governments in New York will become concerned primarily with providing only traditional municipal services. Such changes would go far toward solving the severe financial pressures now felt by the state's major cities.

Taxation. To finance such services, New York's relatively healthy economic base provides the source of the highest per capita taxation in the United States. In fiscal 1970–71 the figure for state and local taxes was $689, compared to second-place Hawaii's $614 and the national average of $460. State impositions include income, sales, business, and excise taxes; local revenues are derived mainly from property and sales taxes. New York City is the sole local jurisdiction imposing an income tax. The broad state base plus the widespread use of local sales taxes allows New York to rely less on local property taxes than do such other large or heavily populated states as California, Massachusetts, and New Jersey.

Political life. The political struggles between Upstate and New York City cannot be reduced entirely to a matter of Democratic-Republican hostility. New York City's politics has been characterized until recently by strong Democratic (Tammany) leaders of the city's five counties, especially New York (Manhattan) and Bronx counties, but the leaders of these organizations often had as much difficulty dealing with Democratic governors in Albany as with Republican ones. The Republican Party, generally controlling the state legislature, if not the governorship, remained strongly conservative until its traditional leadership was threatened around 1900 by Theodore Roosevelt and Charles Evans Hughes.

The four major political parties in state politics

Although both the Democrats and Republicans have strong statewide party organizations, New York is one of the few states in which third and fourth parties have thrived and often played significant roles in elections. Originally confined to New York City, where the Liberal Party has at times endorsed Republican candidates and in 1969 slated and won re-election for Mayor John Lindsay after he was rejected by his former Republican sponsors, the splinter movement produced in the 1960s the statewide Conservative Party, which in 1970 saw its candidate elected to the U.S. Senate.

Since 1920 the governorship has been about equally divided between Democrats and Republicans. In contrast, both houses of the legislature have been almost consistently Republican since 1940. Political behaviour is characterized by generally effective party discipline in the legislature and strong leadership by the governor. The most bitter clashes have been between the two strongest political figures in the state, the governor and the mayor of New York City. Even when they have been of the same party, the conflicting necessities of the state and of the metropolis, usually financial, have brought the two into strident confrontation.

Education. New York spends more money per pupil for public elementary and secondary education than any other state. This public-school system, with compulsory schooling between the ages of seven and 16, had its beginnings in the Colonial period. Schools were established by churches with government support as early as 1638 in New Amsterdam. It was not until 1791, however, that the state's first public school was established. Some state support was granted in 1795 to elementary schools, and in 1812 a permanent system of public schools was established. Parent-paid fees provided a part of the support, however, until all elementary schools became free in 1867. Public secondary schools came even later. During the 1850s a few cities established such schools, and, during the second half of the 19th century, they spread across the state.

The Board of Regents

The University of the State of New York was established in 1784 and its governance placed under a Board of Regents. All private and public institutions of higher education were included in this general organization. In 1904 the legislature placed the entire educational system under the Board of Regents. This board governs all educational activities in the state. It selects the state commissioner of education, approves the establishment of new colleges, licenses entry into the professions, approves new degree programs, and advises the legislature on all educational issues. Standardized exams used in all secondary schools are called regents' exams. Scores on these exams provide a measure for determining school performance and form the basis for the awarding of a wide range of scholarships.

In 1948 the several public institutions of higher education, primarily teachers colleges and two-year agricultural and technical institutions, plus newly established ones, were incorporated into the State University of New York, an institution distinct from the University of the State of New York but a part of that larger entity. Until the creation of the State University, private institutions dominated higher education. Although private institutions still enroll a higher proportion of college students than in most other states, the state system has been the most rapidly growing public institution of higher education in the country since its founding.

The state-university system comprises four general types of institution. Major university centres are located at Stony Brook, Albany, Binghamton, and Buffalo. The teachers colleges and some new campuses have become general colleges concentrating on undergraduate education but providing some graduate training. Two-year state institutions and community colleges are supported in about equal parts by the state, the county, and student fees. Independent but supported by the state and by New York City, City University of New York provides a great variety of programs, ranging from two-year community colleges to graduate instruction.

Operating alongside this dual public system are over 120

Private colleges and universities

private colleges and universities ranging in size from a few hundred students to the 42,000 at New York University. Included in this group are some of the country's best known. Columbia University, founded in 1754 as King's College, is known for the high quality of its graduate instruction and for the national influence of its teachers' college. Cornell, the base for the agriculture, home-economics, veterinary-medicine, and industrial- and labour-relations units of the State University, is a member of the Ivy League, as is Columbia. Fordham is perhaps the best known of the state's many Catholic colleges and universities. The University of Rochester, known for its pioneering role in music and the natural sciences, and Syracuse University, home of the Maxwell Graduate School, the first university unit established for training students for public service, are also well-known private institutions. Other high-ranked institutions include Colgate, Hamilton, Union, St. Lawrence, Bard, Skidmore, Barnard, and Vassar.

Educational issues dominate much of the public debate in the state: public support of parochial schools, state aid to elementary and secondary schools, establishment of tuition at City University of New York, the relation between the state and city universities, and public support for private higher education.

CULTURAL LIFE AND INSTITUTIONS

The milieu of New York City. Much of the style and tone of life in the United States is set in New York City, which remains the artistic, cultural, and economic capital of the nation. The fashion industry is headquartered in its garment district; the chief live theatre in the country exists on and off Broadway; and many television programs originate in New York, where the three major networks have their home offices. The city's museums—particularly the Metropolitan Museum of Art, the Museum of Modern Art, and the American Museum of Natural History—set the pace for similar institutions across the land, and more and more motion pictures are being filmed on its streets.

Many major publishing houses, too, have their headquarters in New York, as do a large number of national magazines. Serving their needs and those of the central offices of many of the country's largest corporations are banks, public-relations and advertising firms, management consultants, and legal firms. This concentration of business and culture gives New York City its leading national position in so many aspects of U.S. life. These features are covered in greater detail in the article NEW YORK (CITY).

Regional museums and institutions

The arts elsewhere in the state. Cultural and related activities are not confined to New York City. Many art museums are located in the state's large and small cities. Among them, the Albright-Knox Art Gallery, in Buffalo, has outstanding collections of contemporary paintings and sculptures. In Rochester are the Memorial Art Gallery, the Rochester Museum of Arts and Sciences, and the George Eastman House of Photography, the last devoted to the history of photography. The Everson Museum of Art in Syracuse is considered an outstanding example of modern architecture, while the city's Canal Museum is the only museum in the U.S. devoted to canal history. Outstanding symphony orchestras include those of Buffalo and Rochester, while the Eastman School of Music at Rochester is internationally known. Fine architectural specimens are scattered across the state, and the performing arts are actively pursued by professional and amateur groups. The cultural and artistic life of the state's many college and university towns often is centred on these institutions, notably in departments of art, music, theatre, and the like.

Across the state, numerous artistic and cultural activities are presented throughout the year. The Saratoga Performing Arts Center in Saratoga Springs is the summer home of the Philadelphia Orchestra and the New York City Ballet. Theatrical performances also are held at this modern cultural centre. The Chautauqua Institution, located on Chautauqua Lake in southwestern New York, was founded in 1874 as a training school for Sunday-school teachers. The name Chautauqua has since been adopted to include a wide range of cultural and educational activities, including concerts, opera, drama, and lectures. Other music and art festivals are the Adirondack Chamber Music Festival at Schroon Lake, the Signal Hill Festival of Music and Dance at Lake Placid, and the Lake George Opera Festival at Glens Falls.

Cooperstown, founded by the father of James Fenimore Cooper (who used the central New York area as the locale for many of his novels) and known as the village of museums, has six, the best known of which is the National Baseball Hall of Fame and Museum. At Hill Cumorah near Palmyra, an annual pageant depicts the founding of the Mormon Church. Also located in this region are the Finger Lakes Wine Museum and the Glass Center in Corning. Dotting the state are historic homes, forts, and battlefields. Over one-third of all the battles of the American Revolution were fought in New York, including the decisive Battle of Saratoga.

Many of these activities have been encouraged in recent years, as New York became the first state in the Union to establish a program for continuing financial support of the arts. The Council on the Arts, which administers the program, had a 1970–71 budget of $18,000,000 for funding to organizations in the fields of the performing arts, visual arts, film and media, and special programs.

Recreation and tourism. The variety of New York's geography provides not only areas of great beauty but also vast opportunities for recreation, relaxation, and a study of the past across the entire state. The cool summers of the Adirondacks, the snowy slopes of the Catskills, the beaches of ocean and lakes, and an almost unlimited repertory of water sports give New York a recreational base certainly as broad as any state in the nation and as yet only partially exploited.

PROSPECTS

New York, now second to California in population, still leads the United States in many fields. The cultural centre of the nation, the headquarters for much of the nation's business community, and the national leader in the move from a manufacturing to service economy, the state has a generally bright future. Problems remain, however. Troubled by urban unrest and with a distribution of resources not matched to needs, the state soon must reorganize its system of local government. At the same time, it faces the prospect that its already high taxes, in the search for a more equitable balance of needs and resources, will probably continue to climb.

BIBLIOGRAPHY. Many state agencies publish annual reports in many areas, but the 16 volumes of the NEW YORK TEMPORARY STATE COMMISSION ON THE CONSTITUTIONAL CONVENTION, *Reports* (1967), are especially valuable. Other general works include the New York State, *Red Book* (annual); the FEDERAL WRITERS' PROJECT, *New York: A Guide to the Empire State* (1940); R.J. RAYBACK (ed.), *Richards Atlas of New York State* (1959); and JOHN H. THOMPSON (ed.), *Geography of New York State* (1966).

Historical works include D.M. ELLIS *et al.*, *A History of New York State*, rev. ed. (1967), containing a large annotated bibliography, and *New York: The Empire State*, 3rd ed. (1969). A host of books cover special aspects of the state's history, among them OSCAR HANDLIN, *Al Smith and His America* (1958); WILLIAM RITCHIE, *The Archaeology of New York State* (1965), which is lavishly illustrated; H.W. HERTZBERG, *The Great Tree and the Longhouse: The Culture of the Iroquois* (1966), one of many works on New York's Indians; LIONEL D. WYLD, *Low Bridge! Folklore and the Erie Canal* (1962); and CARL L. CARMER, *The Tavern Lamps Are Burning: Literary Journeys Through Six Regions and Four Centuries of New York State* (1964); and H.W. THOMPSON, *Body, Boots and Britches* (1940), among many works on folklore and related topics.

For state government, see L.K. CALDWELL, *The Government and Administration of New York* (1954); and ROBERT RIENOW, *Our State and Local Government*, rev. ed. (1965).

(A.K.Ca.)

New York (City)

The most populous metropolis of the Western Hemisphere, New York is, depending on one's point of view,

The four faces of the city

any one of four cities: to social scientists it is a laboratory in which to study the challenges of urban life, from ghastly slum to tycoon luxury; to tourists it is a city of jostling crowds, horn-honking traffic jams, dirty streets, smelly subways—all in dramatic contast to such international symbols as the skyscraper skyline, the United Nations buildings, Wall Street, the Statue of Liberty, the Metropolitan Museum of Art, Times Square, and Broadway theatres; to commuters it is an enervating beehive of world trade and finance, mass media, business administration, fashion, and assorted entrepreneurial activities and manufacture—a place to leave as soon as possible in the evening for the more serene atmosphere of greener suburbia.

But to the nearly 8,000,000 persons who live in this temperate, humid city, most of it raised up on the islands where the Hudson River empties into the Atlantic Ocean, New York is in reality a collection of many neighbourhoods scattered among the city's five boroughs—Manhattan, Brooklyn, the Bronx, Queens, and Richmond (Staten Island)—each ranging in size from a few thousand to more than 100,000 persons and exhibiting its own lifestyle. To move from one neighbourhood to another in the city's 320 square miles (830 square kilometres) may be like passing from one country to another. Such clichés as "New York is not the United States" or "It's a great place to visit, but I wouldn't want to live there" were probably coined by commentators unaware of the importance of this neighbourly aspect of what Walt Whitman called "the hurrying feverish, electric crowds."

This article is divided into the following sections:

For information on related topics, see the articles UNITED STATES OF AMERICA and NEW YORK (STATE).

I. Character and growth of the metropolis

New York has often been called ungovernable, and predictions of its approaching death date back into the 19th century. The prophecies of doom have become increasingly frequent since the late 1960s, as important businesses have left the Empire City for the suburbs; street crimes have multiplied in what the city's press agents have dubbed Fun City; and public services have deteriorated, jobs declined, and the fiscal plight become perilous in the city that displays a beaver, once a symbol of wealth, on its municipal seal.

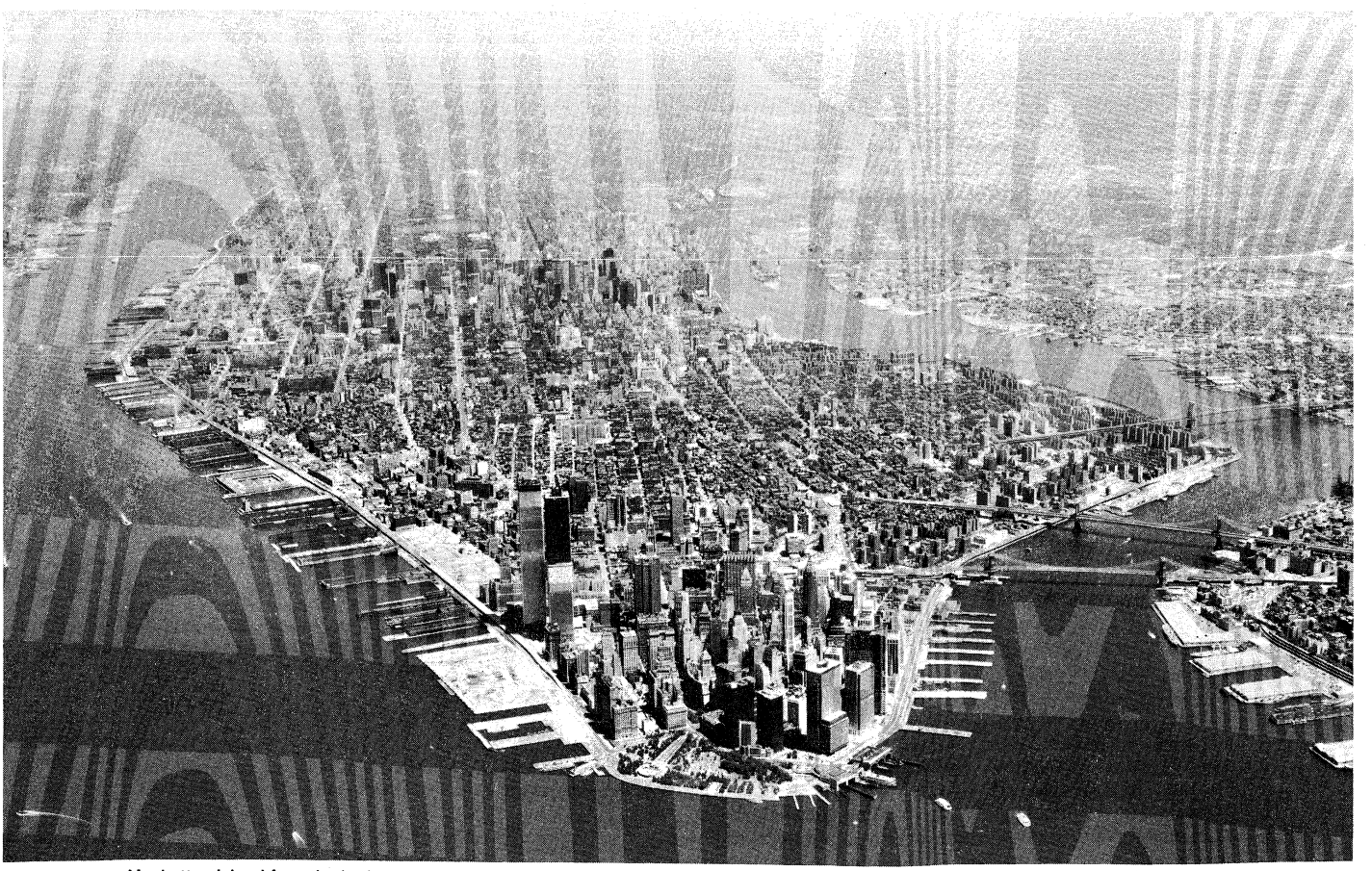
THE PROLIFERATION OF CRISIS

After spending about half of the 20th century reshaping New York with parks, highways, bridges, tunnels, and housing projects, one agency head reflected in his retirement on the difficulty of labelling the city.

> The best of the original Gotham recorders, the O. Henrys, Ring Lardners, Damon Runyons of bygone days, gave us sidelights, vignettes and what in painting is called genre, but it was only a quarter, a corner, phase or facet, certainly not the essence.

The novelist James Fenimore Cooper had sensed this a century earlier, when he wrote: "New York is essentially national in interest, position, pursuits. No one thinks of the place as belonging to a particular state, but to the United States." Periodically, thousands of New Yorkers and sometimes even their mayors, obsessed with the unique character of their fabulous city, have argued that

Thomas Airviews



Manhattan Island from the harbour; on the left is the Hudson River and on the right is the East River. New York's financial district and Battery Park are in the centre foreground. Central Park is at upper centre.

it ought to break away from New York state and set itself up as a separate state of the Union.

Centralization and decentralization. Since the 1960s, two basic thrusts in city government have been in conflict in New York City. Some say that at issue is the very survival of the city or of any huge city. Others contend that the city is going through another of its major transitions from which ultimately it will bring about a blend of the two approaches into a formula for a livable megalopolis. The argument, then, is between advocates of centralization and decentralization.

For more than three decades, beginning with the 1930s, reformers and other opponents of "machine politics" were generally successful in New York City in advancing centralization. The powers of the presidents of the five boroughs were greatly reduced and shifted to the mayor and his appointees. Such operations as education, health, sanitation, parks, recreation, sewers, and roads were directed almost entirely from ever larger central bureaus, almost always in Manhattan. Centralization, its supporters insisted, brought greater efficiency by eliminating replicated layers of bureaucracy in the boroughs and by dropping useless jobs created in those boroughs by district politicians.

During the administration of Mayor John Lindsay, beginning in 1966, the voices of decentralization became steadily stronger. The decentralists said that centralization had isolated the city government on pinnacles too far removed from the people and their problems.

The borough presidents called for the return of their powers so that they could start doing a better job of getting garbage collected, potholes in the streets filled, parks maintained. Sixty-two community-planning boards, their members appointed by the borough presidents, became more active and influential. This was as true in affluent Riverdale, in the Bronx, where beautiful homes and expensive apartment houses overlook the Hudson River, as in the slum of the Brownsville section of Brooklyn, in which burned-out tenements were falling down around newer low-income housing projects. Proposals advanced by the mayor and his City Planning Commission were delayed and sometimes blocked entirely by the pressure of the community-planning boards.

Resurgence of community self-determination

The communities assumed more power in all sorts of activities: picking principals of schools in Harlem, preventing construction of a department store in Manhattan's West Side, saving homes planned for demolition for a high school in the middle class Corona section of Queens, clearing welfare cases out of hotels in Greenwich Village or Brooklyn Heights. Community groups became the watchdogs of city operations on the basic level, pointing out where a police station was undermanned, a traffic light was essential, a hospital was in disrepair, fire equipment was antiquated. The neighbourhood newspaper, almost extinct, was reborn in many parts of the city.

Centralists claimed that the life of the city was being strangled by the whims of community leaders. Decentralists retorted that without healthy communities there could be no healthy city.

The community groups were cemented by fear—the fear of muggers and burglars. Like other large cities, New York was afflicted with many street crimes during the 1960s and 1970s, some of them sadistic. Since New York was the main point of entry and the major marketplace for smuggled narcotics, it became the nation's chief gathering place for drug addicts. Much of the street crime and burglaries were perpetrated by addicts.

Few areas were free of street crime. In the slums, iron fences were fixed inside windows that opened on fire escapes. Double and triple locks, often with bars that fitted from the door into the floor, were used in many apartments in middle class neighbourhoods. And the rich, despite round-the-clock doormen, often installed elaborate burglar systems in the luxury apartments that they owned and for which they paid $1,000 a month for maintenance alone. Purse snatching near supermarkets became so frequent that women stopped carrying handbags or bound the leather straps of the bags around their wrists. Parks and playgrounds became nocturnal hangouts for addicts and pushers. Dogs were bought more for protection than as pets. Skyscraper office buildings hired private police forces and iron gates were drawn across stores at night. Neighbourhoods organized their own volunteer auxiliary police details to patrol at night in cars equipped with two-way radios. Subways, the key arteries of the city, had their own police force.

Historic antecedents of contemporary crises. The growing strength of neighbourhood groups and the fears that bound them were manifestations of an older problem revived. At the core of nearly all of the city's crises—and of its triumphs—have been the masses of impoverished migrants who sailed into its beautiful harbour or spilled out upon its mammoth bus, train, and airplane terminals, their belongings in cardboard cartons, battered suitcases, knotted sheets and pillow cases. Writer and editor E.B. White described New Yorkers as "to a large extent strangers who have pulled up stakes somewhere and come to town, seeking sanctuary or fulfillment or some greater or lesser grail." The migrants, in search of security, prosperity, political power, set up the city's neighbourhoods—its Harlem, Little Italy, Chinatown, Greenwich Village, and scores of communities of tenements, apartment houses, and private homes. They gave the city a variety of flavours ranging from cuisine to speech.

Ethnic diversity and conflict

As far back as 1643, a Catholic missionary, walking the dirt streets of the fledgling Dutch community at the southern tip of Manhattan Island, was astonished to hear 18 languages. In 1870 the census of the turbulent city—it was still limited to Manhattan—showed that of nearly 1,000,000 residents, more than 400,000 were foreign born. In 1920, with the population above 5,600,000, the foreign born totalled just under 2,000,000 and the number of Negroes more than 150,000.

With the sharp curtailment of immigration by federal law in 1921 and 1924, the city entered a period of increasing stability, in spite of the violence of Prohibition, political scandals, the Depression, and World War II. Many of the immigrants and their children prospered and moved from the slums to middle or upper class neighbourhoods. The city was safe enough that New Yorkers during the 1920s, 1930s, and 1940s often slept in the parks or on the beaches on stifling summer nights. In the slums, residents kept their doors open for ventilation, and, in many sections of the city, residents did not lock their doors.

In the 1950s and 1960s, hundreds of thousands of new migrants arrived. The City Planning Commission has estimated that the combined population of blacks and Puerto Ricans in that period rose from 1,000,000 to about 2,300,000. The 1970 census, which counted over 7,900,000 residents and which was probably an undercount, revealed that during the 1960s the city's white population declined by 592,000 and its nonwhites increased by 705,000. Like the migrant waves before them—the Germans and Irish in the 19th century and the Italians and Jews from eastern Europe in the early 20th century—the arrival of these new minorities was followed by the usual disruptions.

Continuities in civic life. That the city has survived more than 130 years of slums, riots, epidemics, crime waves, and corruption, touched off by the many instabilities resulting from waves of immigration, plus vicious commercial competition is a tribute to its adaptability, basic democracy, and the priority given to the art of making a living—a mixture that has produced demagogues and knaves as easily as it has idealists and statesmen. In *Beyond the Melting Pot,* sociologists Nathan Glazer and Daniel Patrick Moynihan observed that "New York became the first great city in history to be ruled by men of the people, not as an isolated phenomenon of the Gracchi or Commune, but as a persisting, established pattern." But, looking upon the same municipal turmoil, the critic Lewis Mumford asserted that "purposeless materialism became the essential principle of the city's life."

THE HISTORY OF NEW YORK

Although the Italian explorer Giovanni da Verrazano spotted the site of New York City in 1524, it is the

English navigator Henry Hudson who is treated as the city's discoverer. On September 3, 1609, he sailed into its harbour and up the deep tidal river that now bears his name in a vain quest for a passage to India on behalf of the Dutch West India Company. Hudson's reports to his employers about the magnificent harbour, sheltered by green hills and limitless potential farmland at a seemingly semitropical degree of latitude, were responsible for the Dutch settlement that set the pattern for the city's drive and independence, a spirit that endured when only corrupted Dutch names remained to remind New Yorkers that it was the Dutch who had founded their city.

The colonizers. In 1614, six years before the Pilgrims landed on Plymouth Rock, a Dutch skipper and his crew were forced to spend a winter on Manhattan Island because their ship had burned; they built there a new seagoing vessel and, prophetically, named the 18-ton ship "Restless." In 1626, Peter Minuit, director general of the Dutch province of New Netherland, which included not only all of what has since become New York City but which also extended into what is now Connecticut, New Jersey, and Long Island, purchased Manhattan from the Indians for 60 guilders worth of gewgaws, the equivalent of $24. On February 2, 1653, New Amsterdam, capital of the province, became a city; its population was 800.

Extent of the Dutch colony

The Dutch West India Company refused to sell land in Manhattan until 1638, but early settlers found ample farmland. They also found out that much of the island was underlain by extremely hard rock and that in the southern portion were substantial stretches of marsh, as well as ponds and creeks. They learned to drain and fill swampy areas and to create land along the riverfronts by dumping refuse. Much of what was to become the borough of the Bronx was rocky ridge. The portion of Long Island that became Brooklyn and Queens was flatter and better suited to farming, though often stony and sandy in consistency. The section that became Staten Island was hilly, with a maximum altitude of 409 feet (125 metres), the highest point in the city.

The Dutch opened the city to foreign commerce, set up trade with the Indians, and laid out the roads that were to become major streets of downtown Manhattan. After a number of bloody clashes with the Indians between 1643 and 1655, interspersed with uncertain truces, the Dutch established the dominance of whites on the island and much of the nearby areas that became part of New York City.

The English city. Dutch supremacy ended on September 8, 1664, when a fleet sent by the Duke of York as part of the English–Dutch war in Europe seized the city easily, since the Dutch were not too fond of their stern ruler, Peter Stuyvesant. The English changed the name to New York. There was a short interruption of English rule in 1673, when a surprise invasion from the Netherlands was successful; the name was changed to New Orange, in honour of the Prince of Orange. But, the following year, in accordance with the Treaty of Westminster between England and the Netherlands, the city reverted to English rule and name. In 1686 New York became the first city in the colonies to receive a royal charter.

Since the Dutch and English got along well in Manhattan, the changes in political office did not interfere greatly with life and trade. English as well as Dutch settlers were agreed that representatives of the British crown were too autocratic. Between 1689 and 1691, Jacob Leisler, preaching rebellion in the city, tried to organize an expedition to attack Canada. He was hanged and beheaded in 1691, but resentment was so strong in the city that the British Parliament cleared his name.

Thus, long before the Revolution, the city had established a strong sense of independence, allied with a business acumen and raffishness that it never lost. Pirates made the city their hangout in the 17th century, and the notorious Capt. William Kidd was regarded as one of the city's leading citizens. He donated the block and tackle with which historic Trinity Church was first built. Merchants helped outfit pirates and shared in their booty.

It was in New York that the first great test of journalistic independence was won and the base of the nation's freedom of the press established in 1734, when John Peter Zenger, publisher and editor of the *New York Weekly-Journal,* charged with libel and jailed for making strong attacks on the government, fought the case and was acquitted in 1735. The first bloodshed in the struggle for independence is believed to have occurred in New York. On January 18, 1775, the Sons of Liberty, a patriotic group, clashed with British soldiers; one of the Sons was killed, and several were wounded.

Precursors of rebellion

Revolution and growth to world status. New York, which was occupied and almost destroyed during the Revolution, became the first capital of the nation (1789–90). George Washington was inaugurated as first president in the city on April 30, 1789. The first session of the state legislature was held there in 1784, and it remained the state capital until 1796. By 1790, with a population of 33,000, it had become the largest city in the nation; by the turn of the century, the population had passed 60,000.

The opening of the Erie Canal in 1825, connecting New York with Buffalo, the Great Lakes, and the opening West, guaranteed the city's pre-eminence as a seaport and world entrepreneur in commerce. When the immigrant waves swept ashore in the city in the 1840s, vast pools of cheap labour, skilled and unskilled, fuelled a new era of rapid growth and social turbulence.

One of the chief beneficiaries of this development was the local Democratic organization known as Tammany Hall. Incorporated in 1789, mainly for social and patriotic activities, Tammany developed political strength under Aaron Burr toward the end of the century and became an important political force after backing Andrew Jackson in his two victorious presidential campaigns (1828 and 1832). With success, Tammany, which had fought for wider suffrage, came under the domination of conservative elements for a time; but with the floods of immigrants in the 1840s, it directed strong appeals to them and built on their votes a massive power base that lasted for nearly a century, making it a national symbol of political bossism. Though Tammany became synonymous with corruption, it also served the vital function of opening up political opportunities to the immigrant poor.

During the Civil War, the city was shaken by its worst riots. For five days in July 1863, many thousands of rioters, mostly impoverished Irish immigrants, infuriated by the new draft law that permitted a draftee to buy his way out of service for $300, swept the city, looting, burning, and killing. Negroes were hanged from streetlights and trees. Warships trained guns on the city as they clashed repeatedly with the police, national guardsmen, and the army. At least 2,000 persons were killed and 8,000 wounded, and all business halted in the face of the armed conflict.

After the Civil War there was a steady clamour in the city for a merger with Brooklyn, Queens, the Bronx, and Staten Island. The strongest resistance was from Brooklyn, a city in its own right, which with good reason feared that the enormous corruption so evident in Tammany Hall under the political bosses William Marcy Tweed and Richard Croker would be extended to Brooklyn through any consolidation. With the opening of the Brooklyn Bridge connecting Brooklyn and Manhattan in 1883, the merger became inevitable. It took place on January 1, 1898, with Manhattan, the smallest of the five boroughs in size, becoming the most powerful. The birth of Greater New York represented a change from city to metropolis, from national to world stature.

Problems of consolidation

The transition to world metropolis in the first two decades of the 20th century was powered by the arrival of millions more immigrants from Italy and eastern Europe. This huge pool of cheap—and often skilled—labour tied together the sprawling city with networks of bridges, tunnels, elevated and subway systems; created its famous garment industry; boomed its printing trades and small manufacturing; drew entrepreneurs from around the world; and made the city a laboratory for unionism and radicalism. During the first decades of the century, the public education system turned out armies of

New York City Metropolitan Area.

extremely efficient white-collar and civil-service workers for the city's increasingly complex but booming economy.

II. The contemporary city

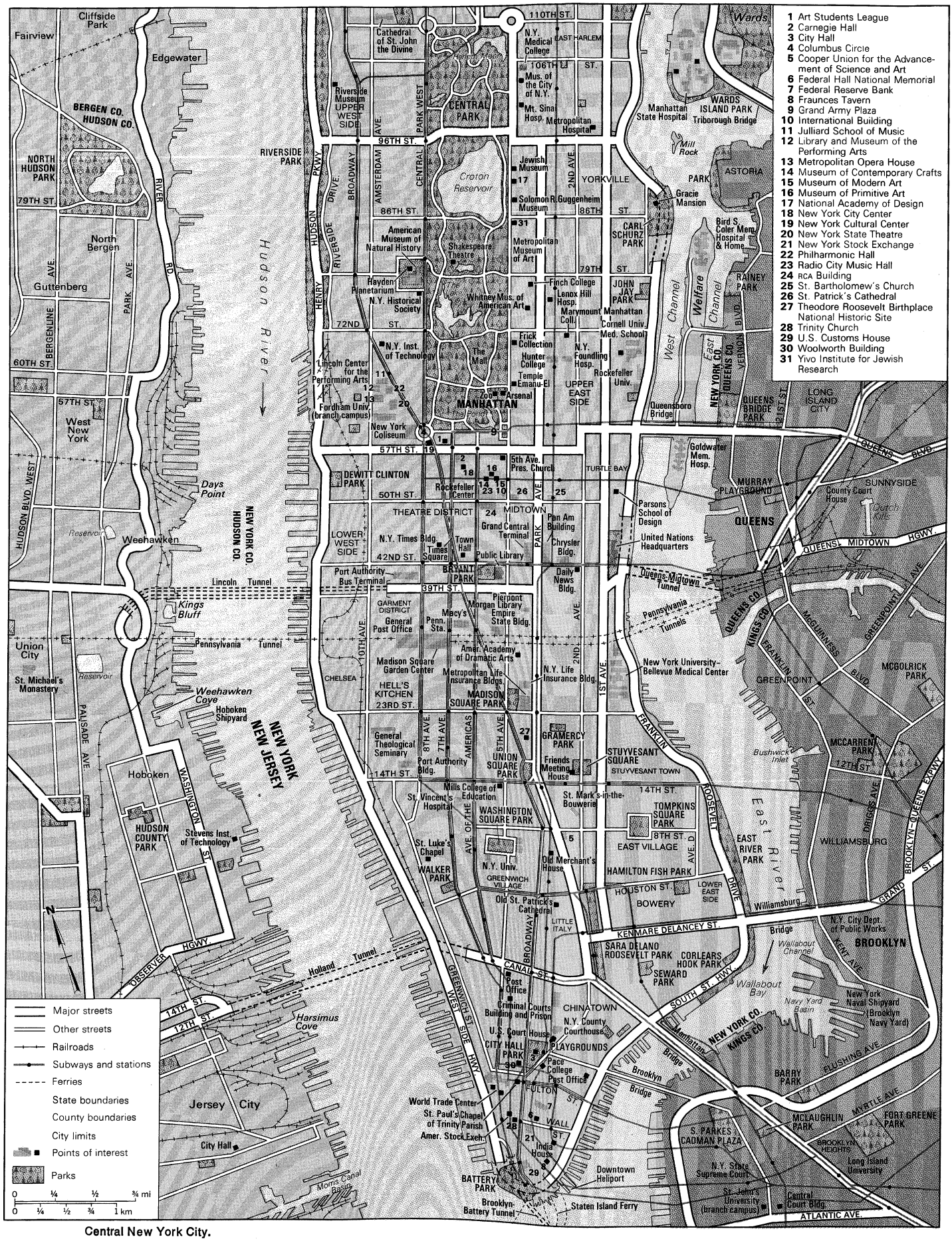
If eventually New York does prove to be ungovernable, its residential areas truly unlivable, and its commercial and cultural centres completely unvisitable, it will not be for want of effort on the part of its leaders or populace. The New York of the 1970s remains one of the most geographically and demographically complex of world cities, its economy one of the most diversified, its political structure among the most chaotic, and its cultural scene unquestionably the richest in the Western Hemisphere.

THE BOROUGHS AND THEIR PEOPLE

Many decades after the city's consolidation, tourists and even most New Yorkers, when they speak of "the city," still mean Manhattan. The boroughs of Brooklyn, Queens, the Bronx, and Richmond—each of them a separate county—remain the subject of jokes among New Yorkers as well as strangers. Brooklyn continues to be taunted with the title of the Thomas Wolfe story *Only the Dead Know Brooklyn*. Queens is known as "the borough of cemeteries." The Bronx is constantly reminded that its greatest claim to fame may be the derisive noise popularized in the bleachers of that borough's Yankee Stadium and known as "the Bronx cheer." And Richmond, closer to the smokestacks of New Jersey than to its political kin, is treated as outlying wilderness. Since most residents of these four boroughs work in Manhattan, the lingering feeling remains that Brooklyn, Queens, the Bronx, and Richmond are mere "bedrooms" of Manhattan.

Stereotypes and realities of the boroughs

Manhattan actually has been surpassed in population by all of the other boroughs except Richmond. Three of these boroughs—Brooklyn, Queens, and the Bronx—



Central New York City.

would rank among the nation's major cities if they were not a part of New York. Scores of thousands of residents of these other boroughs rarely visit Manhattan and know much less about the heart of their city than do the most casual of tourists. Legislators from these other boroughs complain endlessly that they are neglected by the city government in favour of Manhattan.

Only when seen in its entirety can the waterborne character of the city be fully appreciated. The Bronx, before 1895 a part of suburban Westchester County, is the only one of the boroughs that is part of the mainland of the United States, and even that borough is bounded on south, east, and west by water. The other boroughs are either islands or parts of an island. The city's only land boundaries are Westchester County on the north and Nassau County on Long Island to the east. Other than that, it is hemmed in by the Atlantic Ocean, Long Island Sound, the Hudson River, and assorted bays, straits, and rivulets that give the city a waterfront of 578 miles (930 kilometres) for shipping and recreation.

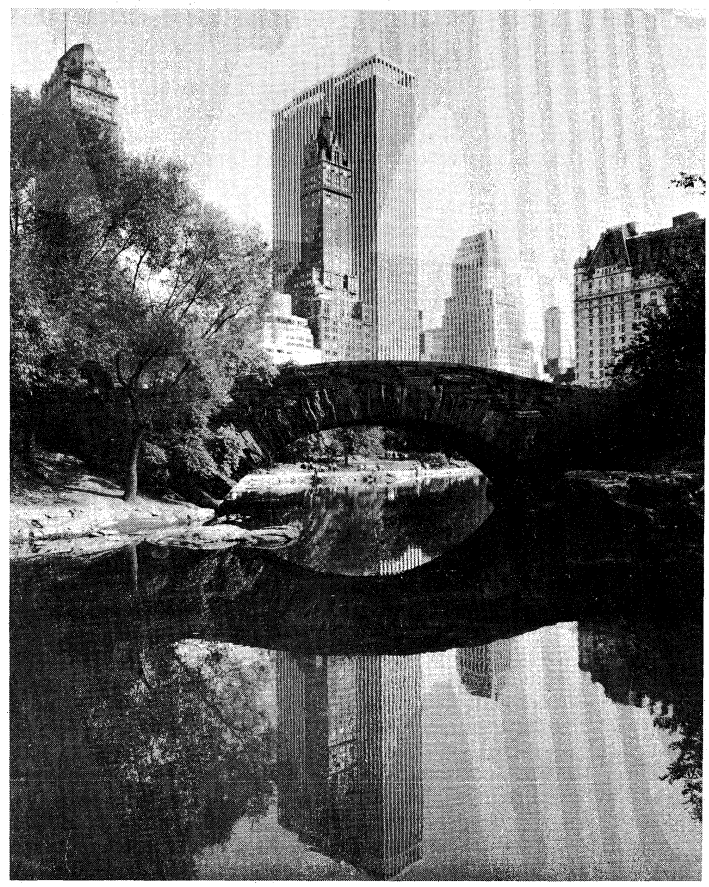
Manhattan. The magnet for tourists, the symbol of the city, Manhattan is probably the most deceptive of the boroughs to outsiders, who generally limit themselves to quick looks at the Times Square theatre district (moving gingerly past the seediness of 42nd Street west of Broadway), the shopping promenade of Fifth Avenue, the munificence of the temples of finance on and near Wall Street, the eccentricities of bohemian life in Greenwich Village, the exotica of Chinatown, or the special flavours of Little Italy and Harlem. At first glance, Manhattan is only the city of skyscrapers, glaring lights, frenzied pace, an island of the strange, the neurotic, the avant-garde. Crammed into its 22.36 square miles (57.91 square kilometres) were some 1,500,000 residents—according to the 1970 census, which undoubtedly missed many thousands because census takers feared to look too carefully in some slums. Its waterfront, formed by the Hudson, East, and Harlem rivers, is 43 miles in length, but only scattered groups of slum children swim in the pollution; and the few fishermen find only scanty catches.

To the residents of the island, each section is a hometown. Those who live in the West 70s, 80s, and 90s—the Upper West Side, though streets run above 200 at the northern tip—know their neighbourhoods as a cosmopolitan mixture of languages, occupations, and income levels. It is the cauldron in which much of the liberal experimentation in the Democratic Party is prepared, and some say it is the origin of much of the chaos of the party. On the Upper East Side, east of Central Park, is a different mixture, generally more affluent, with more unmarried men and women having a first fling in the big city.

The Manhattan that tourists do not see

The Chelsea area of the West 20s, with its tenements, renovated brownstones, and huge cooperatives built by labour unions, has a more sedate pace than the East Village, comprising much of the old Lower East Side in which generations of European immigrants festered. There, poverty, radicalism, crime, and drugs have created the most unstable area in the city. Harlem means more than just tenements, housing projects, and black politics. It means a vibrant street life ranging from sports to stoop seminars, and it is spiced with luxury apartment houses with doormen, inhabited almost entirely by blacks. Yorkville, in the East 80s, retains pockets of Czech, Hungarian, and German cultures in a clash of old tenements and towering luxury apartment houses. The neighbourhood taverns of the Irish proliferate through Inwood at the northernmost part of the island, where the borough of Manhattan spills over the Harlem River to encompass a few-square-block enclave within mainland Bronx. In Inwood lie Manhattan's few remaining forested acres, and on open recreation areas the Irish keep alive their national sports of hurling and Gaelic football—much as courts are maintained for bocciball games in Little Italy many miles to the south. On Morningside Heights around Columbia University, the civilities of the academic world overlook the bleak stretches of Harlem below and to the east and north.

Even fantastic Lower Manhattan, from the Battery, with its ferry slips at the island's tip, to City Hall, began taking on the atmosphere of a neighbourhood. Apartment houses went up in the vicinity of City Hall, and the overwhelming skyscraper jungle around Wall Street, which was home to some 2,900 financial and insurance institutions and six of the nation's 10 largest banks (with assets of $350,000,000,000), exerted international power. Indicative of the new era for Lower Manhattan were plans announced in 1972 to build a $1,200,000,000-development of apartments and office buildings on platforms in the East River bordering the financial district.



Luxury hotels and the General Motors Building surrounding the lower end of Manhattan's Central Park.
Peter Gridley—FPG

Brooklyn. The most populous of the city's boroughs and, next to Manhattan, the best known is Brooklyn, coterminous with Kings County. With adjacent Queens, it comprises the western end of Long Island. Bounded by the Atlantic Ocean, the East River, New York Bay, and Queens, it covers 80.95 square miles (209.7 square kilometres) and has a population of more than 2,600,000 at the start of this decade and 201 miles (323 kilometres) of waterfront, including that subway-Riviera amusement park known as Coney Island.

Brooklyn retains much of the identity that it developed as a separate city. Its downtown area has a bustling shopping district and office buildings. Many factories, some of them quite large, are scattered along its waterfront, and the discontinued Brooklyn Navy Yard, now the New York Naval Shipyard, where many of the nation's greatest warships were built and berthed, is being turned to industry in which the poor are being trained. Though Brooklyn has many private homes, the majority of Brooklynites live in apartment houses, tenements, or mammoth housing projects. The contrast of neighbourhoods in Brooklyn is, in its special way, as dramatic as those of Manhattan, ranging from the choice residential section of Brooklyn Heights, with its magnificent view of New York's harbour and the Manhattan skyline, to the dreadful slum of Brownsville, with many blocks of burned-out or abandoned buildings, dreary tenements, boarded-up stores, and low-income housing projects.

Contrasting life-styles of Brooklyn

The borough has one of the largest black communities in the world, Bedford-Stuyvesant, with some 400,000 persons cramped in a mixture of attractive brownstones, towering housing projects, and slums. There are, however, unmistakable signs that a stable neighbourhood is being built, and the area elected the first black U.S. congresswomen. Just as special in character are Borough Park, with an Orthodox Jewish community and storefront synagogues; Bensonhurst, with Italians who cling to their private homes with gardens to the rear; and Crown Heights, in which blacks and Jews are working to stem decline.

Queens. With its 118.63 square miles (307.3 square kilometres) and nearly 2,000,000 residents, Queens has become the bastion of the city's middle class, a catch basin for those who have fled deteriorating sections of Brooklyn, Manhattan, or the Bronx. Bounded by Long Island Sound, the East River, the Atlantic Ocean, Brooklyn, and Nassau County, it contains the beautiful beaches of the Rockaways and the boating facilities along the sound. Unlike Manhattan, which has suffered a population decline, Queens has boomed, with an increase of some 200,000 since 1960. It has drawn new department stores and shopping centres and is the site of Shea Stadium, home of the baseball Mets and the football Jets. Although the boom has brought with it many large apartment houses and tax-eating sewers, schools, and subways, the majority of its residents live in private homes. Thousands of these homes are owned by blacks, a situation uncommon in the rest of the city.

The borough of old neighbourhoods

In Queens, people usually think of themselves as belonging first to their neighbourhood, only then to the borough, and, finally, to the city. Continuity has become important in Queens. In the Irish area of Woodside, it is not uncommon for parent and children to have been married in the same church. Astoria has a public school in which the children were born in more than 20 countries, as well as the largest Greek community outside Greece. In Corona, couples with grandchildren are living in homes built by their Italian forebears. It is no accident that it is in Queens that the homeowners, white and black, have made the fiercest fight against low-income housing projects and high-rise apartment houses. The borough has some pockets of poverty but no slum in the usual big-city sense. "We don't want to be Manhattanized," is their slogan.

The Bronx. With 41.44 square miles (107.3 square kilometres) and about 1,470,000 residents, the Bronx is bounded by Long Island Sound, the East, Harlem, and Hudson rivers, and Westchester County. This borough, also known as Bronx County, has become the scene of one of the most dramatic urban struggles in the nation. The major effort has been to restrain the spread of slums, with their abandoned and burned-out buildings, closed down stores, and spreading crime. The areas known as South Bronx and Hunt's Point are infested with drug addicts and pushers who lurk in abandoned buildings the windows of which have been covered with metal and the doors of which are bricked up.

In these same areas and in Morrisania to the north, an exciting political battle has been under way between the blacks and the Puerto Ricans. As the population of the latter increased, they acquired greater political control, with disputes centring frequently in the antipoverty agencies funded with government money. The first Puerto Rican borough president in the city's history was elected in this borough and became a U.S. congressman. Puerto Ricans from this borough have also been elected to the state legislature. While slums spread steadily to the north in the borough, undermining even the once affluent Grand Concourse, a 15,000-apartment middle-income development has been built in an effort to halt the flight from run-down areas out of the city.

Richmond. Popularly known as Staten Island, a corruption of the original Dutch name Staaten Landt, Richmond is separated from the rest of the city by New York Bay and from New Jersey by narrow straits. Much of its shore lies along the Atlantic Ocean. It is the least densely populated of the city's boroughs, with some 295,000 persons in its 58 square miles (150 square kilometres). Much of its acreage remains open, almost rural country, with relatively few apartment houses. Richmond, however, may be facing the biggest boom of any of the city's boroughs, mainly because of the opening of the Verrazano-Narrows Bridge, which crosses the neck of the city's harbour to connect the borough with Brooklyn. In the 1960s, the population increased by more than 30 percent, by far the largest percentage increase of any of the city's boroughs.

The rural city: Staten Island

The major struggle on this island is between real-estate operators who want to realize the largest possible profit and civic groups who want to avoid the congestion that has accompanied growth in the other boroughs. Meanwhile, large tracts of woodland and meadow have vanished to become housing tracts for refugees from Brooklyn, Manhattan, and the Bronx. The growth of the borough has increased the strain on the famous ferry that runs between the island and Manhattan, which long has been a favourite ride of tourists and New Yorkers, since it passes the Statue of Liberty and affords a fine view of the Lower Manhattan skyline.

DEMOGRAPHIC CHANGE

There is strong evidence to support the view that, since the turn of the 20th century, demography more than geography has been the key to the many changes in New York. Except for land added by fill, the boundaries of the city and of its five boroughs in the early 1970s remained the same as they were in 1898; yet, over the intervening period the city was swept by political, social, and economic upheavals, demographic in origin, that in retrospect seem nothing short of revolutionary. New York City gave birth to much of the urban crossbreeding of capitalism and Socialism that, during the administrations of Pres. Franklin D. Roosevelt, came to be known as the New Deal. The impact of demography, more than the penal code or police activity, has determined the amount and kinds of crime in the city. Demography, more than educational theory or the quality of teachers, has been responsible for the serious dislocations in public education in New York—and for the many experiments in the field.

Demography has always been, in fact, a critical factor in shaping the city. First, New York was a Dutch city, then English, then Irish, then Irish–Italian–Jewish, then Jewish–Italian, then Jewish–Italian–Negro. Now—assuming that most Puerto Ricans classify themselves as white—it has become white–Negro. As each ethnic or religious group has attained power then moved away and lost power to the succeeding wave of immigrants and migrants, the life-style of the city has undergone transformation. But, no matter how often this has happened, New Yorkers, with short memories of their city's history, have behaved as though each demographic quake was unprecedented. Those who are threatened revile the newcomers and are, in turn, denounced as "the establishment." Physical force, political pressure, and economic muscle are all used by both sides in the periodic demographic struggles.

The past illustrates how attitudes change as demographic tides turn. Today, the city's Jews are, it appears, highly regarded for their ambition, industry, and cleanliness, and their contributions to the city's business and culture are enormous. The best estimate is that there are about 1,900,000 Jews in the city, but this figure could be high or low by more than 100,000. Many outsiders think of the city—mistakenly—as mainly Jewish. Yet, in 1895 an article in *The New York Times,* appraising the first large waves of Jewish immigration from eastern Europe to the Lower East Side of Manhattan, said, "Cleanliness is an unknown quality to these people. They cannot be lifted to a higher plane because they do not want to be." Attacks on the Irish a few decades earlier and on the Italians shortly after the turn of the century were just as savage. As late as the 1920s, opponents of private bathrooms in tenements argued that tenement dwellers would "only put coal in the bathtub." In recent years, descendants of these immigrants have made similar comments about blacks and Puerto Ricans, who have been stamped as uniformly

Historic animosity toward the immigrant

filthy, lazy, and criminal—a wholesale menace to society.

Today, as in earlier decades, there is an element of truth in these charges. The latest waves of immigrants and migrants do live in slums, amid prostitution and crime. They are aggressive in trying to move up. They become a threat as they reach for power. A major difference today, however, is that the tensions between white and black that exist in New York City are national, if not international, in scope as well. At the same time, journalistic standards have risen, and sharper distinctions are drawn between fact and prejudice.

The key unit in the city's demographic pattern is not the borough or other political district but its 2,159 census tracts. The 1970 census showed that, with the rise of blacks and Puerto Ricans and the departure of non-Puerto Rican whites, segregation became intensified at a time when the courts were handing down many decisions for desegregation. More than two-thirds of the census tracts are either 90 percent black or 90 percent white. Of the 5,100,000 non-Puerto Rican whites, some 4,400,000 live in tracts that are more than 90 percent non-Puerto Rican white. Dozens of these tracts, according to a study made by *The New York Times,* have no Puerto Ricans or blacks. This has meant, in effect, that, since 1960, ghetto areas have expanded to twice and three times their size in Brooklyn, the Bronx, and Queens.

A study of the city's Human Resources Administration revealed that over 80 percent of the city's black population lived in its 26 poverty areas in 1970 and that during the 1960s the population of these areas increased from 980,000 to over 1,300,000. During the same period, about 1,000,000 non-Puerto Rican whites left these areas, generally for other parts of the city, and some 750,000 non-Puerto Rican whites fled from the Bedford-Stuyvesant section of Brooklyn alone.

Shifting ethnic patterns

Overall city figures showed that during the 1960s nearly 1,000,000 non-Puerto Rican whites left the city entirely, while the black population grew by nearly 600,000, the Puerto Rican white by nearly 400,000, the Puerto Rican black by over 40,000, and the "other" group (mainly Chinese) by over 120,000 (more than tripling in size).

Apart from statistics, the movements of the ethnic groups within the city showed a strong hostility toward integration by whites, and any integration that did occur was short-term in areas into which blacks were moving and from which whites were fleeing. Whites did not move into black neighbourhoods, though there were some black neighbourhoods that were more prosperous than some white ones. Most whites living in black areas were elderly, poor, and unable to leave.

The demographic changes, however, were more than just colour. The median age of the black population in 1970 was about 10 years lower than that of whites, a difference that would have been even greater if the whites did not include Puerto Ricans, whose median age was believed to be even lower than that of the blacks.

The age figures are particularly significant, for crimes of violence in all ethnic groups are most prevalent among the young. So is social unrest. The lower age—combined with the larger families among blacks and Puerto Ricans—also placed an enormous strain on the public schools because the blacks and Puerto Ricans, for reasons that, it was alleged, ranged from fatherless homes to poverty, found it difficult to keep up with non-Puerto Rican whites. In 1971, tests taken by over 580,000 public-school pupils from the second to the ninth grades in Brooklyn and Queens showed a continued decline in reading achievement as compared to the rest of the nation. Two of every three pupils were from two to nine months below the national average. In poor areas the upper grades were frequently two to three years behind the national average, whereas in middle-income areas the students were as much as two years ahead of the average.

Though the enormity of these demographic shifts cannot be underestimated, they are not unprecedented in New York City. The census figures of 1870 show a situation not very different from those in 1970. The population of the city in 1870—it was then limited to Manhattan—was 940,000, of whom about 420,000 were foreign born. Considering the foreign born plus their New York-born children, the Anglo-Saxon elements that had dominated the city for the first half of the century were outnumbered. The Irish born, numbering over 200,000, were already the strongest political force in the city. Another 150,000 had moved from Germany, while England, the next highest supplier, sent only 24,432. Smaller amounts landed from dozens of other countries. The black population of the city at that time was just over 13,000.

Antagonisms between ethnic groups

The anti-Catholic feeling against the Irish in the city at that time was as bitter in other parts of the nation, often worse. Just as there is animosity in the city today between blacks and Puerto Ricans, so the antipathy between Germans and Irish was so strong that it erupted into bloody clashes. And the middle class was moving as fast as it could from the areas being settled by the Irish, the suburbs of that time being Brooklyn, Queens, the Bronx, and Staten Island—and a great deal of open land remained in Manhattan. A book of that period noted that "Strangers coming to New York are struck with the fact that there are but two classes in the city—the poor and the rich."

Yet, it would be a mistake to assume that all immigrant—or migrant—peoples in New York have faced equal difficulties. The Germans of the last century encountered much less trouble than the Irish. One reason was that most of them were Protestant, the religion of the group entrenched in power. The Italians, though of the same religion as the Irish, often built their own churches rather than go to those attended by Irish Catholics, partly because they wanted to be able to go to confession to a priest who knew their language. Difference of religion was an important reason for hostility toward Jewish immigrants.

There is a general assumption that the blacks and Puerto Ricans in New York are treated more harshly than earlier immigrants. There is undoubtedly a deep intensity to the anti-Negro feeling and, by many of the Negroes, to the anti-white feeling. Yet, in many respects, the blacks and Puerto Ricans are better treated than all earlier immigrants and migrants. None of their predecessors in the frightening city received welfare. During the first two decades of this century, immigrants often froze or starved, living in cellars or even in the streets. There was a time when tuberculosis became known as Jews' disease. As recently as the 1920s and 1930s, it was common in immigrant areas to see the possessions of these newcomers on the street for failure to pay one month's rent on time. Immigrants were expected to learn English or suffer. They were tormented and taunted for their different custom or their accent.

The city has learned—or been forced to learn—to handle immigrants and migrants with more consideration. For more than a decade, signs in such public buildings as police stations have often been in Spanish as well as English, and many policemen study Spanish. In one hospital near Chinatown, signs in the corridor are in Chinese as well as English, the latest indication of a new demographic trend in the city.

In trying to understand New York, the city he had written about for more than 40 years, Meyer Berger, in his book *New York,* wrote "The place wasn't always concrete and it wasn't always crowded. It just grew faster than any other city in history." The hordes of strangers were what made it grow so fast.

ECONOMIC LIFE

Contrary to the impression created by its soaring buildings, transportation crushes, and mass outpourings from office buildings, the economic life of New York City rests not on a few vast corporations but on a multitude of small businesses and manufacturing establishments. This diversity in the world's financial capital gives the city the flexibility and strength to withstand recession or depression with less suffering than Detroit, for instance, with its dependence on huge automobile plants, or Southern California, supported so heavily by government contracts in the aerospace and defense industries.

Components of the economy. Endless beehives of small factories populate the old loft buildings of Lower,

or downtown, Manhattan, lining the narrow cobblestoned streets built for horse and wagon in the long stretch of the island below glittering Midtown. In many glistening skyscrapers are thousands of small offices staffed by only a few persons. The world knows of the city's fabulous garment centre, but it is in little factories or storefronts of Brooklyn, Queens, and the Bronx or in ancient loft buildings of the Lower East Side that much of the subcontracting work is turned out for the kingpins in mid-Manhattan.

Patterns of industry and employment

According to studies by the New York City Economic Development Administration, the average manufacturer in the city in 1970 had only 29 employees, and the average business had 17. There are between 190,000 and 200,000 employers in New York, not counting the city itself, which is the largest employer, with a payroll ranging between 300,000 and 400,000. All told, about 4,000,000 persons worked in the city in the early 1970s including those on city, state, and federal payrolls.

White-collar workers made up the vast majority of the employment rolls, with slightly more than 800,000 in manufacturing and more than 3,000,000 in nonmanufacturing jobs. Of those in nonmanufacturing, nearly 800,000 were in service businesses; about 750,000 in wholesale or retail trade; more than 470,000 in finance, insurance, and real estate; about 330,000 in transport and public utilities; and more than 100,000 in contract construction. Governmental payrolls had more than 550,000.

Though small business and manufacturing make up the bulk of the city's employment, its biggest economic growth is in banking. The city's banking institutions handle some $80,000,000,000 a year. The glass-and-steel skyscraper headquarters of the mammoth banks are most conspicuous, but modest street-corner banks have sprung up all over the city.

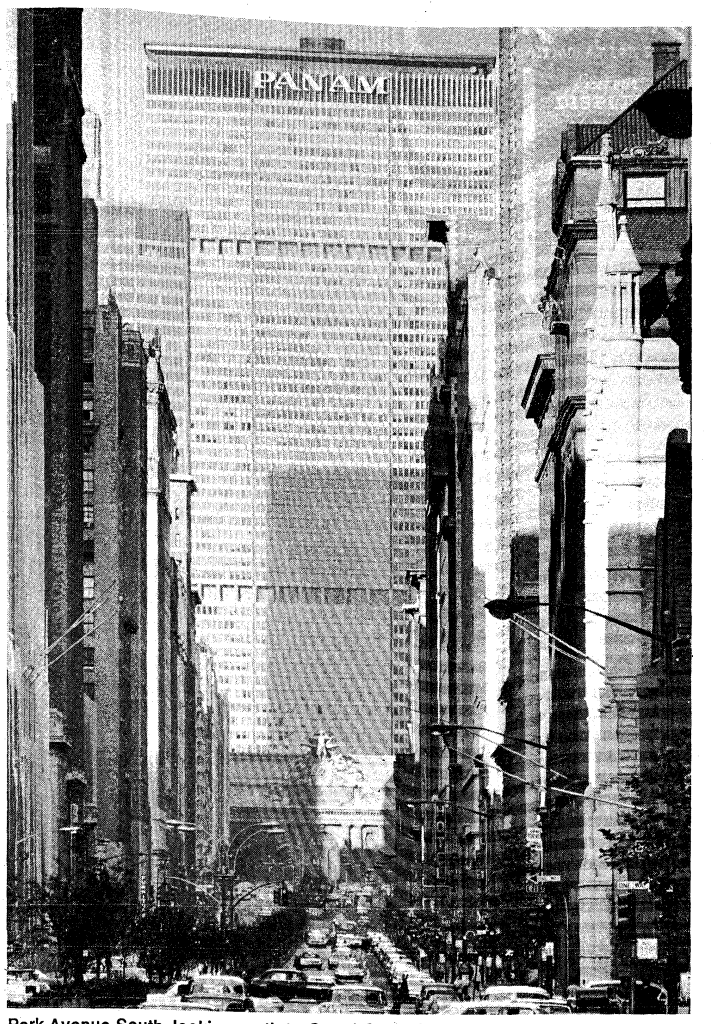
The city's wholesale business runs also to some $80,000,000,000 a year; its retail trade, to about $13,000,000,000; department store sales, $660,000,000. With the headquarters of the nation's television and radio networks, the city is the heart of the mass media, and mid-Manhattan holds the main offices of most of the nation's major advertising agencies. Increasing computerization has made the city's white-collar workers more than ever the mainstay of employment.

New York's major industry is, as it has been for many years, apparel. In spite of the fact that many clothing manufacturers have moved to the South, this field employed more than 200,000 persons in the early 1970s. Printing and publishing, despite the death of newspapers in the city, continued to grow and employed more than 120,000. The headquarters of most of the nation's major publishing houses were clustered in the Midtown area.

Crisis in economic activity. Although the departure of a corporation headquarters for the suburbs of New Jersey, Connecticut, or Westchester County attracts a great deal of attention, losses have been greater in manufacturing. The decline in manufacturing was caused more by automation and mergers than by exodus from the city. City officials are particularly concerned about this drop, since it is in this sector of the economy that the relatively unskilled workers, especially the ethnic minorities, can get jobs or be trained for them.

One of the basic problems the city faces in struggling to keep manufacturing companies is how to find space in which they can expand without excessive cost increases. In manufacturing, city rents often account for 20 percent or more of the cost. Thus, some companies find it more expedient to leave the city entirely rather than pay the increased rent that expanded quarters in the city would add to the cost of operation.

To induce expanding manufacturing companies to remain, the city has been acquiring large areas of blighted or idle land for industrial parks in outlying parts of Brooklyn, the Bronx, Queens, and Staten Island. The city can offer this land to manufacturers at a rental that is better than they could obtain from private owners, often with an option to buy. The success of this program would keep more jobs in the city and bring into the treasury tax money that was lost in earlier exoduses of business.



Park Avenue South, looking north to Grand Central Terminal and the Pan Am Building.
Jordan Wilson—Pix

Shipping, which for generations was one of the city's most flourishing industries, has been in steady decline in recent years. Considerable shipping has shifted to the New Jersey side of the harbour, where new facilities for containerization have made the loading and unloading of ships less costly. Wildcat strikes and waterfront crime have prompted other shippers to use ports a considerable distance from the city. Port facilities have become inefficient from disuse. In an effort to increase freight shipping to and from the city, plans have been worked out to resuscitate the huge terminals along the Brooklyn waterfront with a complex of warehouse and railroad facilities. Transoceanic luxury liners, once so common in the city's port, have decreased considerably, outdated by international air travel. It is conceded that nothing can save the once-famous docks along the Hudson known as Hell's Kitchen. The city plans to renovate this area by building "Convention City," with new housing and office buildings replacing the ancient tenements and marginal business operations.

Declining port activities

While jobs lost in shipping have been replaced, to a large extent, by gains in aviation, the city's survival rests on its ability to arrest and reverse the flight of business and industry for reasons other than technological change. One of the main reasons some corporations have left New York is the spread of crime. Those who feel threatened are not relieved to learn that crime is growing even faster in other cities and has spread to suburbia.

Street muggings in the early 1970s made a substantial number of New Yorkers fearful of walking along many streets at night. The hijacking of freight-carrying trucks, particularly those going to and coming from airports, became a matter of serious concern. Pilferage in offices forced enormous security expenses on corporations for

private guards and crime-prevention devices. And immigrant labour is no longer a guarantee of cheap workers, for the city is highly unionized. Finally, rising taxes to cope with soaring welfare costs—in 1971 about one in six New Yorkers was receiving public assistance—have become irksome to businessmen.

These rising costs have militated against such advantages as good transportation, enormous pools of skilled labour, a long tradition of organizational talents, and a city the people of which are geared to deal imaginatively with projects of almost any size or complexity. As the nuisances and even dangers of living in New York have mounted, the appeal of the city for the gifted has begun to wane. In some cases, young executives have declined promotion if it means living and working in New York City. At the same time, the huge exodus of middle class and upper middle class New Yorkers to the suburbs has increased substantially the excellent labour pool available outside the city.

Loss of civic appeal

Another development that some see as a threat to the city's economic health is the enormous advance in communications that may make highly centralized business unnecessary, perhaps inefficient. In recent years, the New York Stock Exchange, facing special taxes, has considered moving to New Jersey from Wall Street, where it was founded in 1792 when two dozen traders, gathered under a buttonwood tree, made a pact setting forth rules for trade in securities.

While statistics on economics and crime are used to argue that the city is or is not growing, there is no question that fear is contagious and breeds in businessmen and employees the thought that they ought to leave the city. To combat this psychology, in 1971, a group of business, industrial, financial, and civic leaders in the city formed the Association for a Better New York with the slogan "Where the corporate action is." Their campaign stressed that new skyscrapers were being built, that office rents in the city were competitive with those of the suburbs and other major cities, and that new businesses were moving into the city from points as far away as California. But, regardless of the future, there was no doubt that the city was struggling with one of the worst of its many crises.

POLITICS AND GOVERNMENT

"You can't fight city hall." This is said out of the side of the mouth by New Yorkers or with a fatalistic shrug or both, or else, "Go fight city hall," spoken with a mixture of resignation and anger. These clichés, trademarks of New York's political life for generations, are not entirely true, and probably they never have been—certainly not during the 20th century. What New Yorkers mean when they make these observations is that the gap between government and the people is greater than it should be or that government is not as responsive as it should be. What this boils down to is that New Yorkers have a strong feeling that they have a right at least to air their complaints before their local government, if not to get action in their favour.

Backgrounds of civic politics. There have been periods in the city's history—particularly in this century—when the avenues to power were clogged or changing so rapidly that a time lag developed during which the public felt particularly frustrated while it groped for openings. Many New Yorkers think this is the situation today, whereas others believe the government is more responsive than ever. The sense of frustration has happened in the past as much when reform groups established political organizations as when the traditional political "machines" tossed out the reformers. Thus, during the latter part of the 19th century, when the Tammany scandal under the political chief called Boss Croker became so outrageous that voters elected reformers, one term of high principled reform proved so irritating to the voters that in the very next campaign the old-line politicians, running their candidates on a slogan of "To hell with reform," regained control. On the other hand, Fiorello Henry La Guardia, during the 1930s, and, more recently, John Lindsay, although making repeated attacks on "clubhouse politics," were elected precisely because clubhouse politics had become lethargic. And once in office, they proceeded to win re-election by utilizing many of the methods that had made political machines successful in the first place. Their secret was to find ways to help the people "fight city hall"—or at least think they were fighting.

Machines and reformers in city hall

Charters of New York City government over the years have often been less important than the manner in which the government, which they established, actually functioned. This often depended on individuals. During many years, the organizational talent and style of a political "boss" have been dominant, for evil or good. The nefarious Boss Tweed used power to corrupt in the post-Civil War era, but Boss Charles Francis Murphy, during the early decades of the 20th century, had the vision to pick Alfred E. Smith from the midst of ward politics and back him for governor. By 1928, Smith had risen to become the Democratic presidential candidate.

Sometimes, it has been the personality and ability of the mayor that made democracy stumble or work in the highly complex and fluid society that is New York. Mayor James "Gentleman Jimmy" Walker, probably as popular as any mayor the city ever had, was forced to resign under fire in the early 1930s because he was too tolerant of the dishonesty of political cronies. La Guardia, in the later 1930s, was not only the mayor but the political ruler as well.

No matter whether it is a political boss or an elected mayor who rules, those New Yorkers who care make the necessary adjustment in learning how to fight or use city hall, for New Yorkers are never allowed to forget, in their politically turbulent city, that as long as they can vote, they have a voice in city hall—if they remember that in organization there is strength. On this fact, bosses and mayors never disagree. This was the essence of the only book written by Edward J. Flynn, who rose from district leader in the Bronx, to borough leader, and finally to close adviser to Pres. Franklin D. Roosevelt and Democratic national chairman. In explaining why he wrote *You're the Boss,* Flynn said,

> I am a practical politician. I know the facts of political life. I know that political machines, far from being anachronisms, are as modern as the combustion engine and as indispensable. . . . I know that wherever the majority of voters work actively inside a political machine, you have a machine that represents the voters. It is as simple as that.

Volunteers for Lindsay, operating out of storefront headquarters in 1965 and 1969, demonstrated this fact just as clearly as the Tammany block captains did during the 1920s.

Party organization and roles. Two facts are basic in the city's political life and its government: it is predominantly Democratic, and it is strongly independent. Registration in 1969 was over 3,000,000, two-thirds of it Democratic, one-fifth of it Republican. In the 1972 Presidential election, Richard Nixon was narrowly beaten by George McGovern, though Nixon swept the state.

The Liberal and Conservative parties

Many other voters are accounted for by the Liberal Party, founded in the 1930s; its leadership and adherents tend to represent an even stronger liberal position than does the Democratic Party. Although more often than not it lent support to Democratic candidates for city, state, or national office, it has backed liberal Republicans such as Lindsay in 1965 or, as in 1969, has fielded a candidate of its own. In this instance it was Lindsay again, who, having been repudiated by the city's Republican leadership, defeated both the Democratic and Republican candidates. Since the mid-1960s the statewide Conservative Party also has made inroads into registrations, tending to pick up support from normally Republican voters or from ethnic white communities concerned with rising crime, taxes, and welfare rolls. Both mayoralty elections demonstrated another New York City tradition: to win the city, a non-Democrat must split the vote of his opponents. Lindsay did not become a Democrat until 1971, when he opened his drive for the presidency.

The independent vote in New York City, with the ceaseless internal bickering that it generates in reform Demo-

cratic clubs, has had state and national repercussions. It has weakened the Democratic Party not only in the city but also in the state and nation. No longer do the Democrats, by piling up huge majorities in the city, guarantee the election of Democratic governors and United States senators. In the early 1970s, Republicans held the state governorship and split the Senate seats with the Conservatives, and all three men probably owed their election to the Democratic infighting in New York City. This weakness has been reflected on the national scene as each of the city groups has sought the support of national leaders or has fought them. The multiplicity of Democratic candidates for the presidency in 1972 was at least partly the result of the fragmentation that had its origins in New York City political arguments.

Infighting and outcries. Ironically, factionalism in New York City was fomented by the biggest Democratic vote getter in history, Franklin D. Roosevelt. When he sought the nomination for his first term, in 1932, Tammany Hall fought him bitterly, aiding the Democratic machines in the other boroughs, except Flynn's Bronx organization.

Once elected, Roosevelt decided to destroy Tammany's power. He supported the formation in the city of the American Labor Party, which drew largely upon unions. When it became apparent that it was being manipulated by Communist-oriented politicians, it was split, and the anti-Communist group that formed the Liberal Party significantly weakened Tammany. This condition stirred up other dissidents in the party, and the discipline was shattered. When Robert F. Kennedy was elected United States senator from New York in 1964, many Democrats believed he would unify the party in the city, but his assassination in 1968 was followed by even greater factionalism. Lindsay's switch to the Democratic Party raised hopes among some Democratic leaders that he might reconcile warring elements.

Political power of ethnic communities and the poor

Despite the splits in the party, the Democratic organization in the city has an enormous potentiality because, as always, it has a virtual monopoly on the poor. The organization that once built its strength on the Irish, then Italian, then Jewish immigrants now has almost exclusive power among the blacks and Puerto Ricans and has sponsored the only members of these groups to be elected to office.

One reason for this success in the slums is that the Democratic Party has always believed strongly in the "balanced" ticket—that is, a ticket on which racial or religious blocs are represented. It is understood, for example, that the Democratic candidate for borough president of Manhattan will be a black and that any citywide ticket will have a Catholic, a Jew, and a Protestant. This balance has been attacked by many as a form of demagoguery, whereas to others it is democracy at work, a way of showing that the Democratic Party believes in helping everyone have a foot in the door of City Hall—or of the governor's mansion in Albany.

Not since mobs of Irish immigrants were riled up in the 19th century to descend on City Hall have New Yorkers become so persistent and resourceful in forcing officials to listen to their complaints. In the early 1970s, the poor —this time the blacks and Puerto Ricans—were trying once more to push their way into office. They were developing their own leaders, some more opportunistic and ruthless than others, and they had the perfect medium of the poor—television, which does not require literacy. The poor and their leaders had learned that television likes demonstrations, the noisier the better. And, since New York City is not only the headquarters of the major television stations and a major outlet for public television, demonstrations can be fun. All subways lead to City Hall—and wooden barricades and mounted police are always waiting.

The rest of the city has learned that the way to win the attention of the mayor and other officials is to copy the television antics of the poor. They, too, picket, hold sitins, and shout in angry chorus at hearings on bills and appropriations. Unions, tenant groups, parent–teacher associations, civic organizations, civil-rights disputants —all come equipped with signs, bullhorns, public-relations panoply. They all know how to fight city hall.

Structure of city government. The structure of this city government under such constant pressure—the windows of the mayor's office are bulletproof—has three major components: the mayor, the Board of Estimate, and the City Council. Elected for four years, the mayor appoints his deputy mayor and agency heads and holds considerable additional patronage; he also prepares the budget. Thus, under the city charter, his power is considerable. Whether he is in fact powerful has varied with the individual mayor and the other political figures. In general, beginning with La Guardia, the mayors have been stronger than party leaders and usually led their own party.

Board of Estimate, City Council, and the mayor

The City Council is the main legislative body. It introduces and passes all laws and can override a mayoral veto by a two-thirds vote. It is made up of the president of the council, elected on a citywide basis, and 37 councilmen, elected for four years. Of these, 27 are elected from districts within the boroughs, two at large (and of different parties) from each borough.

The Board of Estimate, the main power of which is to act on the budget prepared by the mayor, is made up of the mayor, the council president, the comptroller, and the five borough presidents. Thus, the mayor, comptroller, and council president, through a system of unequal allocation of votes, can outvote the five borough presidents combined.

Friction between city and borough leadership

The real struggle for power by the borough presidents lies on the administrative side of the city's government. The Lindsay regime, which repeatedly voiced support for "community control," also set up an array of superagencies, the heads of which Lindsay appointed to run the city administratively. These superagencies operated in the fields of human resources, environmental protection, health and services, finance, housing and development, economic development, municipal services, and parks and recreation. In addition, a quasi-public organization called Health and Hospitals Corporation, with 10 of its 16 directors named by the mayor, was established. Plans were outlined by borough presidents and the governor to decentralize authority in the city by the establishment of neighbourhood councils and increased power for the borough presidents. And all of this stems from the old battle cry of New Yorkers, "Go fight city hall."

PUBLIC SERVICES

New York City, with a real estate assessed valuation of several tens of billions of dollars, spends more on its residents than any other city in the United States—more, in fact, than any state. And the more it spends, the larger its army of public employees becomes, and the louder is the clamour of New Yorkers about poor service or lack of service. From birth to death, New Yorkers are encased in billions of dollars worth of city services, taking them so much for granted that only when they falter are the people aware of them. One of the fascinating oddities of New York life is the strong strain of Socialism in the most capitalistic of cities. The city's services are burdened not only by size but also by the complexities of the metropolis. For example, in 1971 the city spent about $500,000 in the hopeless task of trying to remove the proliferating graffiti—often painted in letters a foot high —from subway stations and trains. In the parks, saplings are often chained to the ground immediately after they are planted to thwart tree thieves.

During the early 1970s, the city's annual budgets hovered close to $8,000,000,000, slightly lower than that for the entire state. Any one of the city's major functions—welfare, housing, sanitation, health, protection against crime or fire, recreation, water, roads, transportation—is in itself a mammoth operation. Even the air that is breathed involves an annual city expenditure of several million in pollution tests and prevention.

Despite endless complaints about municipal shortcomings, the city does far more for its residents—certainly spends vastly more money on them—than it used to. In the 1920s, for instance, snow removal was left mainly

Riverbend, new housing in Harlem.
Norman McGrath

to sun and rain. In the 1970s the failure to remove snow in most areas within a week—or to make a start in that direction—sparked demonstrations against the mayor.

The welfare rolls. But some services have suffered in the last few years, mainly because the soaring cost of welfare has forced the city to skimp on other services to which it gives lower priorities. Annual welfare costs passed $1,200,000,000, not including over $160,000,000 on staff and $410,000,000 on Medicaid. The number of welfare cases at the end of 1971 passed 1,200,000 and was rising at the rate of nearly 10,000 a month. The city estimated that soon one out of every six New Yorkers would be receiving some form of welfare assistance. Though the federal government paid about 40 percent of the welfare costs and the state assumed 30 percent, the remaining 30 percent was still a serious strain on the city treasury, particularly since the tax base was undermined by the departure of middle class wage earners for the suburbs. To cope with this the city imposed an income tax on commuters who work in the city.

Persons on welfare who are under 21 years of age comprise about 58 percent of the roll; those over 65, over 6 percent; adults caring for others, nearly 20 percent; disabled adults, over 10 percent. Only 2.8 percent of those on the rolls were considered employable. The enormous welfare costs created a widespread feeling among New Yorkers that there is considerable fraud in the area, caused by laxity in administration. The city's Human Resources Administration, which supervises welfare through its Department of Social Services, conceded that the system used to guard against the clients receiving duplicate checks was not effective. It estimated that fraud amounted to about $5,000,000 a year. The state, however, contended this was just a small fraction of the true total and that the number of employable persons on the welfare rolls was far greater. To some extent these disagreements reflected personal as well as traditional animosities between the city government and Albany.

When it was revealed in 1971 that several thousand welfare families had been placed in hotels, often at costs exceeding $1,000 a month, the condition illustrated more than laxity by city officials. It dramatized the dire shortage of apartments for the poor at a time when so many were on welfare. Lower middle class areas, when evacuated by the whites, soon became dreadful slums inhabited by blacks and Puerto Ricans. Apartments and often entire tenements were burned out, often by drug addicts.

Public housing

As a better solution than welfare hotels, the city had sought to increase construction of low-income housing. New York is already far ahead of other cities in these projects, which are now built almost entirely with federal funds. At the end of 1971, New York had 192 projects with some 157,000 apartments housing more than 500,000 persons. To run the buildings, the city's Housing Authority had 11,000 employees, including its own police force. Rents ranged from $18 to $22 a room. The apartments were mainly for the working poor, but the percentage of those on welfare in projects had been rising steadily for 10 years.

Education. The drama of school disorders and busing for racial balance have overshadowed the size and complexity of the city's educational system. At a cost of $1,500,000,000—with federal and state governments donating large shares—the system included almost 1,150,000 pupils, from pre-kindergarten through high school, some 70,000 teachers and 41,000 administrative personnel, and over 950 permanent buildings and 260 temporary buildings. More than 400,000 lunches are served daily to the students, of which about 75 percent are free.

Public education in New York is a vivid example of the desire for community power in the city. Under a state law of 1969, the city's school system was apportioned into 31 school districts, each with an elected board. The powers of these local boards include administrative duties in elementary and junior–intermediate schools, selection of the community superintendent, assignment of teachers, and some aspects of school construction and repair. The chancellor, the chief executive of the school system, is chosen by the citywide Board of Education, the members of which are appointed by the mayor.

The educational system includes schools for the handicapped and convalescent, trade schools, day schools for the socially maladjusted, and experimental programs with longer school days and individual instruction. Special programs reach back into the ninth grade to prepare underprivileged individuals for college. More than 120,000 pupils are learning English as a second language.

For high school students who want to specialize, there are the High School of Art and Design, the High School for Music and the Arts, the High School of Fashion Industries, the School for the Performing Arts, the New York School of Printing, and the Food and Maritime Trades School.

The city university system

New York is the only city in the United States with a large public-university system. Costs for public higher education in 1972-73, for example, were $441,100,000. The higher-education program has become increasingly important since the introduction of "open admission" in 1970, which allows any high school graduate into either a senior or a community college, depending on grades. Before open admission, the standards for entrance into senior colleges were very high. The new program was adopted to enable more blacks and Puerto Ricans to get into city colleges. More than 100,000 students are enrolled in the senior colleges, 51,000 in community colleges, and 15,000 in graduate programs.

The private universities offer, in addition to the usual courses, special areas of study. Columbia University, for instance, has its Russian studies and school of journalism; Fordham University its medieval studies; New York University its courses in art scholarship. Then there is Juilliard, with its world-famous music, and, more recently, theatre and dance; and the Rockefeller University of Biological Science.

Sanitation and water supply. For generations, New York City has been known as a dirty city. This is the result of ever present slums and of relative indifference of New Yorkers. In recent years, however, residents have become aroused by concern about environmental pollution and are more demanding. But it is still a dirty city in spite of the fact that some 17,000 men are assigned to sanitation work—street cleaning, refuse collection, and waste disposal. The city uses a wide variety of mechanical equipment and spends nearly $20,000,000 a year on automotive and plant maintenance. On an average day the city collects 26,000 tons of garbage. To some extent, the city is handicapped by lack of funds.

A prime example of the city's acceptance of a service until something goes wrong is its water system, which is piped in from many parts of the state, from as much as 300 miles away. On an average day the city uses 1,300,000,000 gallons. At times the various city agencies, each in its desire to perform an efficient service, clash. One agency, for example, decided to introduce electricity for heating in public housing as being cheaper and cleaner at the same time another agency was trying to hold down electric consumption because of the power shortage.

Police and fire protection. One of the most glaring examples of how service in the city is curtailed by the fiscal crisis is the police department. At a time when the most agonizing problem to most New Yorkers is street crime, the police department has been forced to allow its force to diminish by not replacing the men who retire or those who leave. At the end of 1971, the police force numbered 33,015, about 1,500 fewer than the year before. During that year, instances of homicide, rape, robbery, aggravated assault, and motor-vehicle theft rose, and 10 policemen were killed.

The statistic of most concern was that of crimes against the person. The city was third among the nation's top 10 most populous cities in major crimes. To cope with street crime, the department set up special units and task forces and decided virtually to ignore gambling offenses, an area in which police corruption has been the cause of many scandals over the years. New Yorkers are notoriously indifferent to laws against gambling, and strong pressure has begun to build for legalized gambling in the wake of legalized off-track betting. The police have been hampered further by their diversion to cope with all sorts of civic demonstrations.

The city's fire department, highly respected for its ability, has encountered two major difficulties in recent years. False alarms have risen to the point at which, for every three legitimate alarms turned in, there have been two false alarms. In addition, firemen have repeatedly found themselves under attack when trying to fight fires in ghetto areas. In New York, firemen have to know how to fight fires in skyscraper or subway, in ancient tenement or luxury apartment. In 1970, the department fought an average of 350 fires a day.

Subways and street jams

Transportation. The subways, though now part of a state-operated system, are still largely subsidized by the city, in addition to the revenue from the 4,400,000 passengers on an average day. The subways make up over 40 percent of all mass transit in the metropolitan area. About 90 percent of persons using the subways are residents of the city. The three subway divisions of the city have 237 miles of track. During rush hour, trains often run on a one-minute headway. In spite of the long lines of bumper-to-bumper traffic on highways into New York during rush hours, only about 5 percent of those who drive to Manhattan come from outside the island and only 2 percent from the suburbs.

Health and recreational services. Through its Health and Hospitals Corporation, the city operates 18 municipal hospitals. In addition, the city has over 80 voluntary hospitals (privately operated but nonprofit) and some 35 proprietary hospitals (private and for profit). Supplementing these medical institutions are six state hospitals and 22 health centres. The city, because of its connection with medical colleges and hospitals and its enormous variety of cases, long has been a magnificent training ground for doctors. Among the most famous training grounds for interns as well as experienced doctors are New York University–Bellevue Medical Center, New York Hospital–Cornell Medical Center, Columbia Presbyterian Medical Center, and Yeshiva–Albert Einstein, Mt. Sinai, and Roosevelt hospitals.

Some 37,000 acres of the city—more than twice the area of Manhattan—are devoted to parkland. Parks range in size from tiny triangular green patches amid heavy traffic to 526-acre Prospect Park in Brooklyn, 1,257-acre Flushing Meadows–Corona Park in Queens, 840-acre Central Park in Manhattan, 2,117-acre Pehlham Bay Park in The Bronx, and 1,454-acre Great Kills Park in Richmond. Included in the city's park areas are such intensively used public beaches as Coney Island in Brooklyn; the Rockaway beaches in Queens; Orchard Beach in The Bronx; and South Beach of Richmond. Erosion and official reports of pollution at some of these beaches have reduced usage.

CULTURAL LIFE AND INSTITUTIONS

The cultural milieu

Though many New Yorkers have come to think of the Fun City label pinned on their city as a bad joke, there may still be a great deal of truth in the nickname. For those who seek pleasure in almost any form, New York in the 1970s was, as it had been for more than a century, "the place to go." Whether it is theatre, music, ballet, painting, or literature or whether it is baseball, football, basketball, track, hockey, boxing, horse racing, soccer, cricket, or rugby or whether it is live, filmed, or printed erotica—whatever the taste of man or woman in culture, the arts, sports, or sex, New York has it in abundance. The city is the nation's greatest culture mart—and a major fleshpot.

Tied in with the culture and the arts are the middlemen —impresarios, agents, and assorted hucksters. Interlocked with and contributing to the talent pool required for these endeavours are great museums, a mammoth library system, and such world-famous schools as the Juilliard School of Music and the Art Students League. The overall result is an industry, ranging from genius to faker, that gives the city a flavour that attracts and churns up talent from all over the world.

There is no doubt that, as a purveyor of pleasure, the city has been hurt by the widely reported crime in its streets. Tourist business has declined, holding down attendance at theatres, sports arenas, and concert halls. The old New York tradition of staying up late after a show or musical event has been curtailed by fear. Restaurants close earlier. Taxis are scarce in the early morning hours. The city that was once wide awake around the clock now slinks home about 1 AM. New York still has dozens of

night spots, but the carefree atmosphere has been dampened. In some ways, however, New York is more exciting than ever as a cultural and artistic test tube.

The performing arts: drama, music, dance, and film. The most vivid paradox of simultaneous decline and growth is in the city's theatre. The decline of the Broadway theatre began in the late 1920s when the movies learned to talk. Since most plays of that period were mediocre or worse, they offered little competition to the "talkies." The Depression of the 1930s added to Broadway's woes, turning the theatres along 42nd Street into burlesque houses and then cinema palaces. World War II, with so much money to be spent, gave legitimate theatre a boost, but afterward it was hurt by television, rising production costs that could not be met by increased ticket prices, and the exodus of large chunks of the middle class audience to the suburbs. In the early 1970s, the 35 major theatres remaining in the Times Square area represented about one-half the number of the 1920s.

Perhaps more important, competition to the New York theatre was springing up in other parts of the country. What had once been fleeting summer theatres in the suburbs began to take root on a year-round basis. Regional theatres developed in Minneapolis, Los Angeles, Dallas, Washington, D.C., New Haven, and other cities. Theatres in universities grew in number, size, and professionalism.

Continuing vitality of theatrical activity

Yet, there are indications that the legitimate theatre in New York as a whole—not just the Times Square area—is very much alive. Off-Broadway and off-off-Broadway theatres by the scores have sprung up since the 1950s in lofts, cellars, garages, churches, and old restaurants—by the 1970s more than 150 of them. These new little theatres, having fewer than 300 seats, are a fertile field for experimental plays and a training ground for actors, directors, playwrights, and technicians who could no longer be absorbed by the Broadway theatre. They are particularly important for blacks and Puerto Ricans forming their own companies. A detailed study of the theatre in 1971 by scholars, professionals, and civic leaders established that the number of plays staged in the city almost certainly exceeded that in any city at any time. *Variety*, the show business weekly, reported that the 1970–71 season along Broadway had the fourth highest gross income in history, and road companies earned almost as much in spite of the nationwide economic recession.

Box-office statistics tell only a part of what the theatre means to New York City. It is impossible to convert into dollars the impact of curtain time in Times Square, with well-dressed crowds spilling out of taxis, thronging across sidewalks and into narrow lobbies, dashing back and forth from theatre to tavern during intermission, chattering excitedly at the final curtain, and chasing for cabs after the last echo of applause. There is a spontaneity, sophistication, and animation to legitimate theatre that movie premieres, even at their most elaborate in Hollywood, cannot capture with spotlights and velvet ropes.

Municipal subsidies for the arts

The city, aware that the theatre means more than money it brings in at the box office—or even at the hotels and restaurants—has taken several steps to bolster it. Four new theatres opened in Times Square skyscrapers in 1972 because of a change in building ordinances allowing builders a bonus in floor space if they include a theatre in the structure. Some experts in the field believe that in the future all Broadway theatres will be in skyscrapers instead of separate buildings. In addition, the city has given financial aid to theatrical work. It sponsors Shakespearean productions that for years have been shown in the parks without charge. It bought an old library that has been converted into five little theatres. The New York City Center, with its ballets and plays, has been able to maintain high standards at moderate ticket prices because the building is tax-exempt and, like the converted library, rented from the city for $1.00 a year. Despite the city's enormous financial difficulties, it spends over $1,000,000 a year on the performing arts, including theatrical productions in the streets of ghetto areas and operas and concerts in the parks by the Metropolitan Opera and the New York Philharmonic.

The most spectacular of the city's institutions of art and culture is the Lincoln Center for the Performing Arts, which has since been copied by Los Angeles and Washington, D.C. Its five buildings, built around a lagoon and fountain, are the home of the Metropolitan Opera, the New York City Opera, the New York Philharmonic, the New York City Ballet, a repertory theatre company, and a library for the arts. The performances at this centre, which has become one of the leading tourist attractions in the city, drew about 3,500,000 persons in 1971. It has been host to theatrical, musical, and dance groups from all over the world. Like most operatic and musical organizations, the Metropolitan Opera and the New York Philharmonic operate at deficits that are met by private contributions.

Concertgoers still think the best concert hall in the city is Carnegie Hall, which gets many of the major musical organizations of the world. Town Hall, the auditorium at Hunter College, and the Brooklyn Academy of Music are in constant use by major artists and chamber groups.

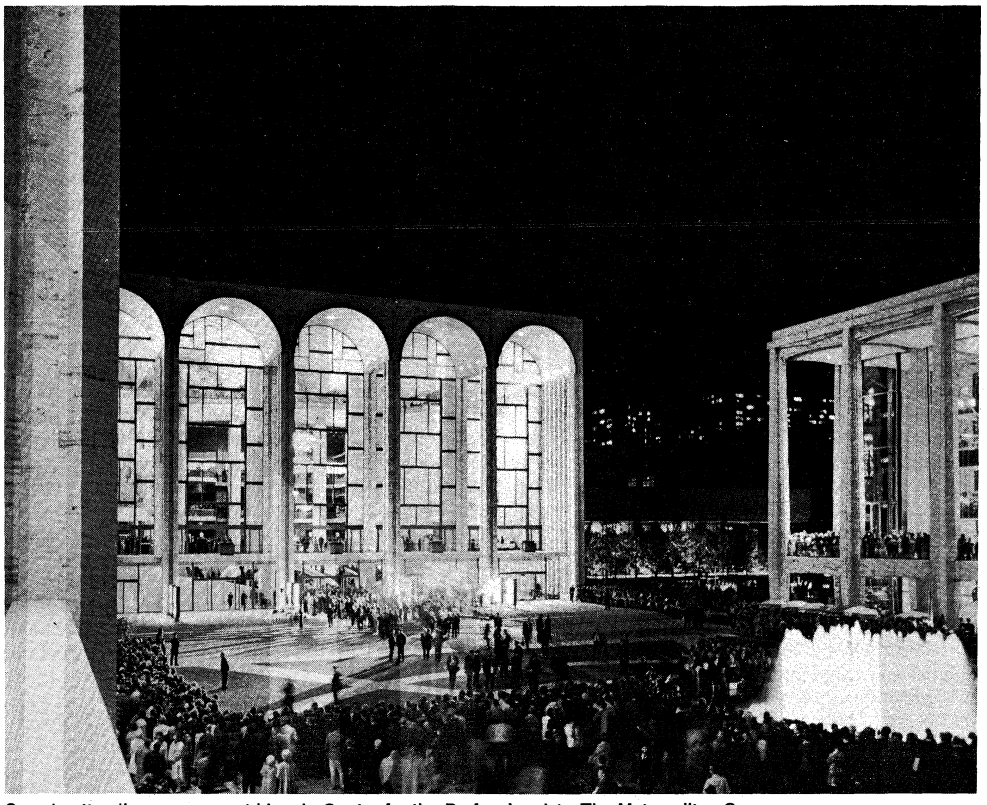
The selection of movies available in New York City is a major attraction. At any time, movie houses in Manhattan alone are showing hundreds of films—first runs, revivals, and anything in between. It is a treasury for those who seek old or new foreign films. The city is also growing as a film-making centre because directors and producers have learned that using the city as a setting gives any movie dramatic impact. This development is a reminder that, before movies went to Hollywood, they were made mainly in New York City.

Libraries, museums, and galleries. New York's library system, with nearly 200 branches, is one of the city's glories. As a result of the city's economy drive, the libraries are not open for as many hours as they once were, and book circulation has fallen. This decline may also reflect the impact of television on reading habits in the city. The circulation figures do not, however, represent the vast amount of research done in the main library, guarded at 42nd Street and Fifth Avenue by stone lions that are among the city's best known sights. Nor do the figures for the city's system reflect the value of special libraries such as the Pierpont Morgan Library, the Low Memorial Library at Columbia, or the collection of books and records at the Lincoln Center Library.

Museums are another important aspect of the city's cultural life. There the visitor can see paintings by old or modern masters, a replica of a whale diving from the ceiling, or a display of planetary mysteries. By the millions, New Yorkers and visitors troop through museum corridors, singly or in school groups, sometimes listening to explanations on earphones or to guides. Most famous are the Metropolitan Museum of Art and the American Museum of Natural History and the adjacent Hayden Planetarium, all situated on parkland. Other art museums are the Museum of Modern Art (with its major collections of photography and films as well), the Frick Collection, the Solomon R. Guggenheim Museum, the John Hay Whitney Collection, and the Brooklyn Museum.

Though the museums are private, the city pays about $18,000,000 a year to help some of them. The general pattern, however, is for museums to finance deficits by private contributions. In general, attendance at the city's museums has declined, and, in recent years, the Museum of Natural History and the Metropolitan have introduced a system of discretionary payment by visitors. Lesser-known institutions with special appeal for scholars as well as visitors are the Museum of the City of New York, the New York Historical Society, the Jewish Museum, the Museum of the American Indian, Heye Foundation, the Museum of Contemporary Crafts, the Studio Museum of Harlem, and the research centre for students of Jewish life known as Yivo. In addition, at least 70 private art galleries help keep the nation's artistic life centred in New York. Among recent developments in the arts are the special arrangements for artists to live and work in loft buildings in the vicinity that has come to be known as Soho—"south of Houston Street."

Outdoor attractions. Deep in the affections of city dwellers and visitors of all ages are the zoos and botanical



Crowds attending programs at Lincoln Center for the Performing Arts. The Metropolitan Opera House is at centre and Philharmonic Hall at right.
By courtesy of the Lincoln Center for the Performing Arts, Inc.; photograph, Bob Serating

gardens. The city's major zoo is in the Bronx, but, because of its location, a much smaller zoo in Central Park has been photographed more often for movies. Other zoos are in Brooklyn, Queens, and Staten Island. The botanical gardens in the Bronx and Brooklyn are a meeting place for botanists as well as the less scholarly flower lovers. Attractive but not as well-known is the botanical garden in Queens.

The major-league sports events at Yankee Stadium, Shea Stadium, or Madison Square Garden Center may get the headlines and television coverage, but for New Yorkers the parks and beaches get much more recreational use. On an average summer weekend about 1,000,000 persons are at Brooklyn's Coney Island and another 750,000 swim at the Rockaway Beach in Queens, both on the southern shore of Long Island. Baseball can be seen at its most exciting among the Puerto Rican teams and the show-business teams in Central Park, in which rowboaters may while away the summer hours or skaters the gray winter days. At these events and among the countless thousands of picnickers in the city's parks is discernible a warmth and neighbourliness, an underlying dream of the human melting pot never realized but ever sought. It is perhaps fitting that New York City is the home of the United Nations.

PROBLEMS AND PROSPECTS

For all its troubles, New York remains the most exciting city in the nation, an endless drama that brings together the cultures and the peoples of the world. It is a metropolis of limitless paradox and surprise. When a 12-day subway strike hit New York in 1966, thousands of motorists gave lifts to strangers, and the city adjusted in dozens of ways to its paralysis. During an East Coast power blackout of 1965, the city of lawlessness had a drop in crime.

Still nothing compares with the subway crush, the ticker-tape parades for the nation's heroes, the glitter of an opening night on Broadway or at the Metropolitan Opera, the promenade of the young singles and couples along Second and Third avenues, the view from the Empire State Building, the sight of a luxury oceanliner gliding down the Hudson into New York Bay. If New York is dying, as many say, many others insisted that it is far more alive on its deathbed than the places to which so many New Yorkers have fled in the suburbs.

BIBLIOGRAPHY. The FEDERAL WRITERS' PROJECT, *New York City Guide* (1939), is a basic book about the city. The colour, history, and many of the social problems of New York have been caught in many works of fiction, notably CLAUDE BROWN, *Manchild in the Promised Land* (1965); and RALPH ELLISON, *Invisible Man* (1952), both concerned with the black experience of the city; STEPHEN CRANE, *Maggie and Other Stories* (1960); and HENRY ROTH, *Call It Sleep* (1935), recalling the life of the Jewish immigrant. WASHINGTON IRVING's classic *Knickerbocker's History of New York* (1809, many later editions), casts a facetious eye over the city's earlier centuries, especially the Dutch years. Works of nonfiction that offer a myriad of glimpses are MEYER BERGER, *New York* (1960), selections from his decades of chronicling for *The New York Times*; PHILIP HONE, *The Diary of Philip Hone, 1828–1851* (1889); and JAMES D. MCCABE, *Lights and Shadows of New York Life* (1872, reprinted 1970), both of which are impressions of 19th-century New York; ALEXANDER KLEIN (ed.), *The Empire City* (1955), a collection of essays and articles on the city; and LINCOLN STEFFENS, *The Autobiography of Lincoln Steffens*, 2 vol. (1931, reprinted 1968).

Problems of New York's political life are covered in WALLACE S. SAYRE and HERBERT KAUFMAN, *Governing New York City* (1960); and in EDWARD J. FLYNN, *You're the Boss* (1947, reprinted 1970). Tammany Hall is the subject of EDWARD N. COSTIKYAN, *Behind Closed Doors* (1966); GUSTAVUS MYERS, *The History of Tammany Hall*, 2nd ed. rev. (1917); and WILLIAM J. RIORDAN, *Plunkitt of Tammany Hall* (1948).

Aspects of New York's sociological and related problems may be found in HERBERT ASBURY, *Gangs of New York* (1937), a look at the structure of gangland; NATHAN GLAZER and DANIEL P. MOYNIHAN, *Beyond the Melting Pot: The Negroes, Puerto Ricans, Jews, Italians and Irish of New York City*, 2nd ed. (1970); JANE JACOBS, *The Life and Death of Great American Cities* (1961), a study of the nature and process of urban decay; OSCAR LEWIS, *La Vida* (1966), on the Puerto Rican community; and JACOB A. RIIS, *How the Other Half Lives* (1901, reprinted 1971), an early and classic study of poverty on the Lower East Side by a muckraking journalist. JOSEPH MITCHELL, *The Bottom of the Harbor* (1959), is an intimate look into the city's harbour and waterfront life; while ROBERT MOSES, *Public Works: A Dangerous*

Trade (1970), reviews the author's years as a builder of the city's physical plants.

Aspects of New York's cultural life appear throughout such books as BROOKS ATKINSON, *Broadway* (1970); IRA GLACKENS, *William Glackens and the Ashcan Group* (1957); MOSS HART, *Act One* (1959); and NORVAL WHITE and ELLIOT WILENSKY (eds.), *AIA Guide to New York City* (1967).

(Mu.S.)

New Zealand

New Zealand, an independent member of the Commonwealth, is situated in the South Pacific over 1,000 miles southeast of Australia, its nearest neighbour. Its population in the early 1970s was almost 2,900,000. The land area is just over 100,000 square miles (260,000 square kilometres)—about twice the size of England. About two-thirds of the land is economically useful, the remainder being mountainous. The country comprises two main islands—the North and South islands—and a number of little islands, some of them hundreds of miles from the main group. New Zealand administers the South Pacific islands of Niue and the Tokelaus, and a section of the Antarctic continent. The country's isolation has played an important part in the development of its social, cultural, and economic characteristics.

Originally a British colony, New Zealand has been technically sovereign only since 1947, although in practice it has been self-determining in many internal and external policy matters since 1901. Small, isolated, and of little political or strategic importance on the one hand, it enjoys a high standard of living on the other. Western-type society has been made to incorporate the original Maori inhabitants of the islands. It played a small, but significant, part in the United Nations immediately after World War II, but in the following two decades showed comparatively little interest in this area. Subsequently, however, New Zealand has become more conscious of its neighbours in the South Pacific, an underdeveloped region in which more developed countries such as Australia and New Zealand have an important contribution to make. This greater awareness of the existence of its smaller neighbours parallels greater interest in race relations and minority problems at home.

Economically, the country is very dependent on the export of its pastoral products, and therefore sensitive to conditions in Europe, the United States, and Japan, its main markets. The entry of Great Britain into the European Economic Community in 1973 will undoubtedly affect New Zealand adversely because much of its meat and dairy produce had been exported to Britain. Some success has attended efforts to expand markets in the developing countries and in Australia, Japan, and the United States, but this has been a slow process. Efforts are also being made to diversify the economy with tourism and greater industrial activity. High wages, a dearth of resources, and the small local market, however, are formidable barriers to efficient manufacturing. (For related information see NEW ZEALAND, HISTORY OF.)

THE LANDSCAPE

Relief. Although New Zealand is small, its geological history is as complex as that of a continent. There must have been land in the area from the oldest Paleozoic times (570,000,000 years ago) or earlier, although by the end of the Tertiary Period (2,500,000 years ago) little of the country remained above sea level. This was followed by periods of gentle warping and changing, alternating with great mountain-building activity during which the "backbone" of the North and South islands was created.

The role of volcanic activity

The Pleistocene Epoch (2,500,000 to 10,000 years ago) saw the beginning of volcanic activity which still continues in the North Island. The major North Island mountains are all volcanic in origin, and all but one are still active. Active volcanoes in a wide belt of thermal activity, including hot springs and geysers, indicate that earth-changing forces are still at work. New Zealand is part of the zone of structural weakness known as the circum-Pacific Mobile Belt; earthquakes are common, although less frequent and severe than in Japan or Chile.

Both the North and South islands are roughly bisected, by mountains in the South and ranges of hills in the North. Swift, snow-fed rivers drain from the hills, although only in the east of the South Island have extensive alluvial plains been built up. The alluvial Canterbury Plains contrast sharply with the precipitous slopes and narrow coastal strip on the west coast of the South Island. The Southern Alps are a 300-mile-long chain of fold mountains containing New Zealand's highest mountain —Mt. Cook (12,349 feet [3,764 metres])—and 16 other peaks over 10,000 feet, as well as an extensive glacier system with associated lakes.

In the north of the South Island, the Alps break up into steep upswelling ridges. On their western face there are mineral deposits, and to the east they continue into two parallel ranges, terminating in a series of sounds. To the south, the Alps break up into rugged, dissected country of difficult access and magnificent scenery, particularly toward the western tip of the island. On its eastern boundary, this wilderness borders a high central plateau with an almost continental climate.

The North Island is much less precipitous than the South and has a more benign climate and greater economic potential. The central region consists of a volcanic plateau rising abruptly from the southern shores of New Zealand's largest lake, Taupo, itself an ancient volcanic crater. To the east, ranges form a backdrop to rolling country in which pockets of great fertility are associated with the river systems. To the south more ranges run right to the sea. On the western and eastern slopes of these ranges the land is generally poor, although the western downland region is fertile enough until it fades into a coastal plain dominated by sand dunes. To the west of the central plateau, fairly mountainous country merges into the undulating farmlands of Taranaki, where the mild climate favours dairy farming even on the slopes of Mt. Egmont, an isolated, extinct volcano.

The northern shores of Lake Taupo bound a large area of high economic activity. Again, the soil, of volcanic origin, is poor, but very suitable for forestry. Even further north there are river terraces sufficiently fertile for widespread dairy and mixed farming. The hub of this area is Auckland, astride an isthmus with a deep harbour on the east and a shallow harbour to the west. The region north of Auckland becomes gradually subtropical in character, marked generally by poor, leached soil and numerous, deep-encroaching inlets of the sea bordered by mangrove swamps.

Hydrography

The mountainous country of both islands is cut by many rivers which are swift, dangerous, and a barrier to communication. The longest is the Waikato, in the North Island, and the swiftest, the Clutha, in the South. Many of the rivers arise from or drain into one or other of the numerous lakes associated with the mountain chains. Many of these lakes have been used as reservoirs for hydroelectric schemes, and artificial lakes of considerable magnitude have also been created for hydroelectric purposes.

The main rock formations of New Zealand are sedimentary and volcanic. The sedimentary rocks, found over three-quarters of the country, give some indication of the country's sojourns beneath the sea, when various strata were first laid down, then lifted up and tilted by mountain-building forces. The soils based upon these formations are mostly clays and generally poor, their continuing productivity depending upon expert farm management and the well-researched use of artificial fertilizers on a large scale. Pockets of fertile alluvial soil in river basins or along river terraces form the orchard and market-gardening regions of the country.

Climate. New Zealand's climate is determined by its latitude, its isolation, and its physical characteristics. Across such vast oceans as surround it, weather from elsewhere has little influence; there are no extremes of temperature.

Because of the high mountain chains which lie across the path of the prevailing westerly winds, the contrast in climate from west to east is sharper than that from north to south. Mountain ranges are also responsible for the

semicontinental climate of central Otago. New Zealand experiences strong equinoctial winds, partly modified on the coast by daytime sea breezes.

Rainfall is highest in areas dominated by mountains exposed to the prevailing winds. Although mean annual rainfall varies from 13 inches (330 millimetres) in central Otago to 300 inches (7,600 millimetres) in the Southern Alps, for the whole country it is typical of temperate zone countries (25–60 inches [635 to 1,525 millimetres]), usually spread reliably throughout the year.

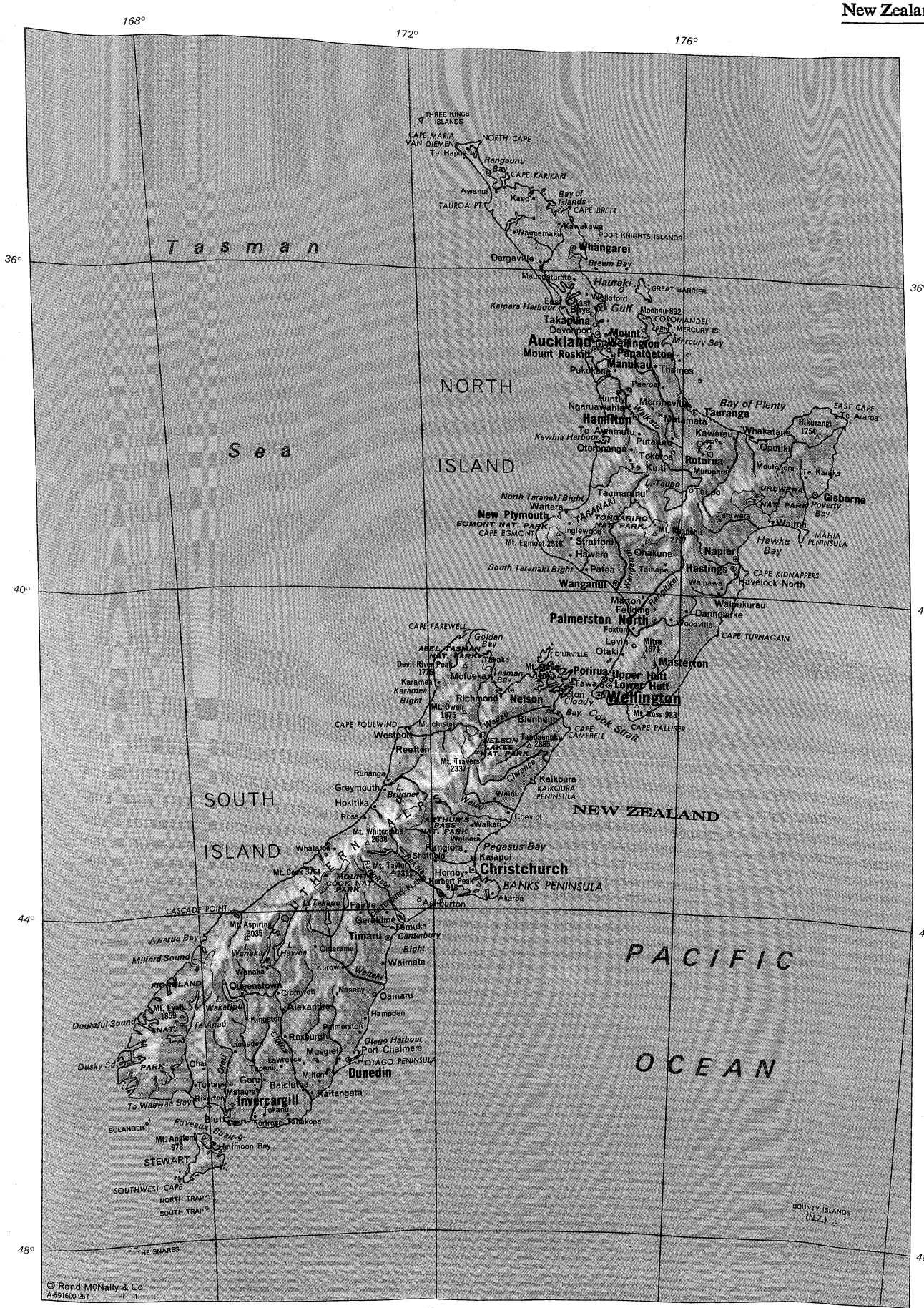
Temperature and humidity

Mean temperatures at sea level decrease from 59° F (15° C) in the far north to 54° F (12° C) at Cook Strait, to 48° F (9° C) in the far south, and there are few extremes. Snow is common only in mountainous regions, but frost is frequent in inland valleys in winter. Humidity ranges from 70 to 80 percent on the coast, generally being 10 percent lower inland. In the lee of the Southern Alps, humidity can become very low. At such times northwesterly winds subject the Canterbury Plains to an unpleasant hot, dry spell. Humidity becomes really oppressive only in Northland, although the temperature there rarely reaches 84° F (29° C). New Zealand is a sunny country, with many areas receiving at least 2,000 hours of sunshine a year, much of it occurring in winter.

Vegetation, soils, and animals. The indigenous vegetation of New Zealand consisted of mixed evergreen forest covering perhaps two-thirds of the country. Prolonged isolation has encouraged the development of species unknown to the rest of the world. Today, dense "bush" remains only in areas unsuitable for settlement and in parks and reserves. On the west coast of the South Island this mixed forest still yields most of the native timber used by industry. Along the mountain chain running the length of the country, beech is the predominant forest tree.

European settlement made such inroads on the natural forest that erosion in high-country areas became a serious problem. The State Forest Service was established to repair the damage by forest-management techniques and reafforestation with exotic trees. Experimental areas on the volcanic plateau were planted with radiata pine, an introduction from California. This conifer has adapted to New Zealand conditions so well that it is now the staple plantation tree, growing to maturity in 25 years

MAP INDEX


168°
172°
176°
180°
36°
40°
44°
48°
Tasman Sea
PACIFIC OCEAN
NORTH ISLAND
SOUTH ISLAND
NEW ZEALAND
THREE KINGS ISLANDS
CAPE MARIA VAN DIEMEN
NORTH CAPE
Te Hapua
Rangaunu Bay
CAPE KARIKARI
Awanui
Kaeo
Bay of Islands
TAURORA PT.
CAPE BRETT
Kawakawa
Waimamaku
POOR KNIGHTS ISLANDS
Whangarei
Dargaville
Bream Bay
Mangawhai
Hauraki Gulf
GREAT BARRIER
Warkworth
Wellsford
East Coast Bays
Kaipara Harbour
Takapuna
Devonport
Auckland
Mount Wellington
Mount Roskill
Papakura
Manukau
Pukekohe
Thames
Moehau 892
COROMANDEL PEN.
MERCURY IS.
Mercury Bay
Paeroa
Huntly
Ngaruawahia
Morrinsville
Waikato
Hamilton
Matamata
Bay of Plenty
Tauranga
EAST CAPE
Te Araroa
Hikurangi 1754
Te Awamutu
Kawerau
Whakatane
Kawhia Harbour
Putaruru
Opotiki
Otorohanga
Tokoroa
Rotorua
Te Kuiti
Murupara
Matata
Te Karaka
L. Taupo
Taupo
North Taranaki Bight
Taumarunui
URUWERA NAT. PARK
Gisborne
Poverty Bay
Waitara
TARANAKI
TONGARIRO NAT. PARK
Tarawera
New Plymouth
EGMONT NAT. PARK
CAPE EGMONT
Mt. Egmont 2518
Inglewood
Stratford
Mt. Ruapehu 2797
Waitoa
MAHIA PENINSULA
Hawke Bay
Napier
Hawera
Ohakune
South Taranaki Bight
Patea
Taihape
Hastings
CAPE KIDNAPPERS
Havelock North
Wanganui
Waipawa
Marton
Feilding
Waipukurau
Danneverke
Palmerston North
Foxton
Woodville
CAPE FAREWELL
Golden Bay
CAPE TURNAGAIN
Levin
Mitre 1571
ABEL TASMAN NAT. PARK
Otaki
Masterton
Devil-River Peak 1775
Mt. D'URVILLE
Porirua
Upper Hutt
Lower Hutt
Wellington
Motueka
Tasman Bay
Tawa
Karamea
Karamea Bight
Richmond
Nelson
Mt. Owen 1875
Cloudy Bk.
Mt. Koss 983
Wairau
Blenheim
Cook Strait
CAPE FOULWIND
Murchison
CAPE CAMPBELL
CAPE PALLISER
Westport
Tasmanenuku 2885
NELSON LAKES NAT. PARK
Reefton
Mt. Travers 2337
Clarence
Runanga
Kaikoura
KAIKOURA PENINSULA
Greymouth
L. Brunner
Waiau
Hokitika
Ross
Cheviot
Waiau
Waikari
ARTHUR'S PASS NAT. PARK
Mt. Whitcombe 2639
Waipara
Whataroa
Rangiora
Pegasus Bay
Kaiapoi
Sheffield
Mt. Cook 3764
Mt. Taylor 2321
Hornby
Christchurch
Herbert Peak 910
BANKS PENINSULA
Akaroa
Ashburton
CASCADE POINT
L. Takaro
Fairlie
Geraldine
Temuka
Mt. Aspiring 9939
Timaru
Canterbury Bight
Awarua Bay
Milford Sound
L. Wanaka
L. Hawea
Omarama
Waimate
Kurow
Waitaki
Whanka
FIORDLAND
Queenstown
Cromwell
Naseby
Oamaru
Mt. Lyon 1859
Wakatipu
Alexandra
Hampden
Kingston
Doubtful Sound
Te Anau
NAT. PARK
Palmerston
Roxburgh
Lumsden
Otago Harbour
Port Chalmers
Mosgiel
Clutha
Lawrence
Dusky Sd
Ohai
Tapanui
Milton
OTAGO PENINSULA
Dunedin
Gore
Balclutha
Tuatapere
Mataura
Kartangata
Te Waewae Bay
Riverton
Invercargill
Tokanui
Bluff
Foveaux Strait
Fortrose
Tahakopa
SOLANDER
Mt. Anglem 978
Halfmoon Bay
STEWART
SOUTHWEST CAPE
NORTH TRAP
SOUTH TRAP
BOUNTY ISLANDS (N.Z.)
THE SNARES
© Rand McNally & Co.
A-691600-257
NEW ZEALAND
Size of symbol indicates relative size of town • ○ ◉ ▣
Elevations in metres
0 25 50 100 200 km
0 25 50 100 150 mi

and having a high rate of natural regeneration. Large areas of the central plateau, together with other marginal or subagricultural land in north Auckland and Nelson, are now planted with this species.

European broadleafs are widely used ornamentally, and willows and poplars are frequently planted to hold eroding hillsides. Gorse has acclimatized so readily that it is now a menace, spreading over good and bad land alike, its only virtue being as a nursery for regenerating bush.

Because of New Zealand's isolation there was no higher animal life in the country when the Maori arrived (at dates variously estimated from AD 970 to several centuries later). There were two species of lizards: the gecko, which is born directly instead of being hatched from an egg, and the tuatara, a "beak-headed" reptile extinct elsewhere for 100,000,000 years. There were also a few primitive frogs, and two species of bat. These are all extant, although confined to outlying islands and isolated parts of the country.

In addition to their domestic animals, Europeans also brought other species with them. The red deer, introduced for sport, and the Australian opossum (for skins) have both multiplied beyond imagination and do untold damage when browsing in the high-country bush. Attempts are being made to exterminate the opossum, although this will be a lengthy process and may be no more successful than the campaign against rabbits: the last rabbit has not been eliminated, but their numbers are rigidly controlled.

In the absence of predatory animals, New Zealand was a paradise for birds, the most interesting of which are flightless. The moa was a very large bird, easily exterminated by the Maori. The kiwi, another flightless species, is extant, though only in secluded bush areas. The weka and the notornis or takahe (barely rescued from extinction) probably became flightless after arrival. The pukeko, a swamphen relative of the weka, is even now in the act of losing the use of its wings.

Some birds, such as the huia, the saddleback, and the native thrush, are peculiar to New Zealand, but many others, such as the tui, the fantail, and the bellbird are closely related to Australian birds. New Zealand also has its share of migrants: the gannet from Australia, and the skua, penguin, shag, and royal albatross from the Antarctic.

Because New Zealand is the meeting place of warm and cool currents, a great variety of fish is found in its waters. Tropical species such as tuna, marlin, and some big-game sharks are attracted by the warm currents which are locally populated by snapper, trevally, and kahawai. The Antarctic cold currents, on the other hand, bring the blue and red cod and the hake, while some fish, tolerant to a considerable range of water temperature, are found off the entire coastline—tarakihi, groper, and bass. Flounder and sole abound on tidal mudflats, and crayfish are prolific in rocky areas off the coastline.

THE PEOPLE

Ethnic structure. New Zealand was first settled by the Maori, a Polynesian race, at a time and in a manner both of which are still disputed. Perhaps they arrived about AD 1300, in great migratory waves, the result of unknown pressures in Polynesia; but they may well have arrived haphazardly when their canoes, on regular interisland expeditions, were blown off course. Whenever and however they arrived, it is certain they remained isolated in New Zealand until the arrival of European explorers, the first of whom was the Dutchman Abel Tasman (1642).

Maori culture

Their culture, influenced by isolation and, perhaps, the temperate climate, became significantly different from that of their Polynesian relatives. They were sophisticated artists and shrewd politicians, skilled botanists, and ardent warriors. Because they were also intelligent, resourceful, and flexible, they fared better in the difficult encounter with white civilization than the original inhabitants of many other colonies.

The European settlers, who arrived in increasing numbers from 1840 onward, were mainly upper working class migrants determined to establish a society free from the bonds of privilege that had made life in Britain unendurable. They arrived to make a home which they were not unwilling to share with the original inhabitants, although as the rate of immigration increased, the European demand for Maori land led to an inevitable confrontation. With the loss, by enforced purchase, of Maori land there went a loss of psychological identity. The Maori birth rate declined and extinction appeared inevitable.

Fortunately for New Zealand, a Maori renascence occurred about the turn of the 20th century, great leaders were elected to Parliament, and the people took heart in the acknowledged worth of their contribution to the new society. The contemporary Maori is, nevertheless, still recovering from the psychological effects of European occupation. Despite legal equality with those of European origin, many Maori still feel unable to take full advantage of the opportunities of a European-type society without seriously compromising their traditional values.

New Zealand thus has two dominant ethnic groups: people of European origin, comprising some 91 percent of the population, and Maori, the indigenous inhabitants of the country, comprising 8 percent (in 1966). Pacific Islanders account for a further 1 percent. The Maori and Pacific Islander proportions are rising—for demographic reasons in the case of the former and through a combination of demography and migration in the latter.

Despite New Zealand's biracial structure, there is little of the racial tension common in other parts of the world. This is not to say that there is no discrimination (on both sides), but increasing efforts are now being made by both the Maori people and a growing body of white New Zealanders to make this a thing of the past.

New Zealand also has a rapidly growing community of Pacific Islanders, people from Samoa, the Cook Islands, Niue, and the Tokelaus, including a large number of young men and women, who are attracted by the benefits of a Western society. They generally congregate in low-cost housing areas where their different cultural habits often alienate their neighbours, whereas their children provide an acute educational problem. The Islanders present a growing problem, particularly as their rate of natural increase is much greater than that of the European or Maori populations.

Apart from that part of the population which is of British origin, there are throughout the country vestiges of settlement by immigrants of European extraction —Dalmations in North Auckland, Chinese market gardeners, Greek and Italian fishermen, Danish farmers, and a small group of Indians. There are also groups of central Europeans who arrived between World Wars I and II, and a large body of assisted Dutch immigrants from after World War II.

Language patterns

New Zealand is predominantly an English-speaking country. Children of European migrant families quickly lose their native tongue in favour of English, particularly when their parents are anxious to establish relationships with the existing community. Only Greeks, Chinese, and Indians make determined efforts to keep their children bilingual. Many Maori are bilingual, but often unsatisfactorily, speaking neither English nor Maori very well. Great efforts are being made to improve the English of Maori children, and Maori is offered in many secondary schools as an optional second language.

New Zealand is nominally Christian, four-fifths of the population adhering to the Anglican, Presbyterian, Roman Catholic, and Methodist denominations (in that order). At the last census only 10 percent either refused to state, or denied a religious affiliation. Minor Protestant sects, the Eastern Orthodox churches, Jewish congregations, and Maori adaptations of Christianity (the Ratana and Ringatu churches) account for the rest. Probably only about 20 percent of the population are active churchgoers. There is no established religion, but Anglican cathedrals are generally used for state occasions.

Patterns of settlement. Because New Zealand is small and homogeneous in population there are no sharply differentiated social or political regions, just as there are

no great differences in climate, terrain, or soil type. The North (roughly the northern half of the North Island), however, is popularly regarded as being more enterprising, while the South—either the whole of the South Island or its lower half—is traditionally regarded as being conservative and dour. While the west coast is romantically nostalgic for its rollicking gold-rush days, the east coast conjures up the not entirely ill-founded picture of sheep barons on their extensive stations. When Europeans first arrived there were practically no Maori in the South Island, and this is still the case today, they and their Pacific Islander cousins being predominantly in the North.

Trends in urban and rural settlement

The New Zealand countryside is thinly populated. There are, however, many small towns of up to 10,000 population and over 20 provincial cities of over 20,000. Very small towns or villages are becoming more and more deserted as people drift to the bigger towns and cities.

The main urban areas are Auckland (approximately 650,000), the centre of the North and the main industrial complex; Wellington, in the centre (approximately 150,000), the political and commercial capital; Christchurch (approximately 280,000), in the middle of the South Island and the second-largest industrial area; and finally, still farther south, Dunedin, with a population of over 100,000. Although New Zealand is notable for the strength and affluence of its rural sector, the great majority of people live in cities, and urban concentration is proceeding apace. There is also a marked difference in the rates of growth of the two main islands—the North now having over two-thirds of the total population, in sharp contrast to the earlier years of systematic settlement.

New Zealand, Area and Population

	area		population	
	sq mi	sq km	1966 census	1971 census
Statistical areas*				
North Island				
Central Auckland	2,150	5,569	614,000	698,000
East Coast	4,200	10,878	47,000	47,000
Hawke's Bay	4,260	11,033	125,000	133,000
Northland	4,880	12,639	94,000	96,000
South Auckland-Bay of Plenty	14,187	36,744	389,000	422,000
Taranaki	3,750	9,713	101,000	101,000
Wellington	10,870	28,153	524,000	553,000
South Island				
Canterbury	16,769	43,432	376,000	399,000
Marlborough	4,220	10,930	29,000	32,000
Nelson	6,910	17,897	67,000	69,000
Otago	14,070	36,441	183,000	183,000
Southland	11,460	29,681	103,000	106,000
Westland	6,010	15,566	24,000	23,000
Total New Zealand	103,736	268,676†	2,677,000†	2,863,000†

*The statistical areas listed have no administrative significance but provide a convenient way of presenting the statistical data. †Figures do not add to total given because of rounding.
Source: Official government figures.

Demographic trends. In recent years births have been stable at around 22 per 1,000 population and deaths at nine, making the annual rate of natural increase about 1.3 percent. For the Maori, the rate of natural increase is over twice as great, because of a higher birth rate and a lower death rate, both largely a consequence of the more youthful structure of the Maori population. The Pacific Islander (immigrant) population exhibits this characteristic to an even greater extent.

Since World War II New Zealand has generally had an annual excess of arrivals over departures of 10,000 to 20,000, contributing about ½ percent to overall population growth. The majority of immigrants have come from Great Britain (or The Netherlands for a decade or so after World War II), but more recently Pacific Islanders have become a major element of the inflow.

Following an economic recession in 1967 there were three years of net outflow or zero inflow, but net immigration has since resumed, although not at as high a rate as formerly. It is clear that both immigration and emigration are sensitive to the rate of growth of the New Zealand economy and its employment opportunities, as well as to conditions overseas.

THE ECONOMY

New Zealand is exceptional in combining high living standards with continued reliance on primary production for a substantial proportion of national income and an overwhelming proportion of exports. In terms of national income per head, the country is among the first ten in the international tables, following immediately after Australia. Small population and great distance from the world's main industrial commercial centres, however, are considerable economic disadvantages.

Natural resources. Most minerals, metallic and nonmetallic, occur in New Zealand but few are found in commercial quantities. The exceptions are gold, which in the early years of organized settlement was a major export; coal, which is still mined to a considerable extent; iron sands, which are exploited both for export and to supply a local steel mill; and, most recently, natural gas, which shows promise of being ultimately the most significant of all. In addition to mining, construction materials, with which the country is well endowed, are quarried. Current interest in mineral exploration may well unearth other exploitable deposits.

Apart from gold's brief heyday, biological resources have always been more significant than minerals. Domestic animals introduced from the Old World have thrived in New Zealand, and the pastoral industry is still the country's mainstay. Forestry has always been important, but the emphasis has swung from felling the long-cycle original forest for timber, to reafforestation with pine trees for both timber and pulp—a flourishing paper industry being based on the latter. Wild animals are of little economic significance, and fish, despite their abundance, scarcely more.

Use of hydroelectric power

The country has great hydroelectric potential, which has been exploited to such an extent that hydro stations supply almost all its needs at low cost. Although costs are rising, further hydroelectric development in preference to fuel-, thermal-, or nuclear-power stations is likely. A notable feature of the New Zealand electricity grid is the direct current cable linking the two main islands, enabling the South's surplus energy to be used by the North's concentration of industry and people. The ready availability of energy sources, and of hydroelectricity in particular, has played a major part in New Zealand's development to date, supporting high standards of domestic comfort and latterly facilitating industrial expansion, including the construction of an aluminum smelter.

Economic activity. *Industry.* Primary industry accounts for roughly 18 percent of national income, agriculture contributing some 90 percent of this, and forestry and mining nearly all the rest in about equal proportions. Productivity is high, in agriculture especially, so that the sector accounts for a comparatively small proportion of total employment. Agricultural produce accounts for some 84 percent of total exports, and forest products for another 6 percent. Farming in New Zealand is highly capitalized, but still mainly an owner-occupier activity with corporate or cooperative marketing.

New Zealand has had relatively large-scale manufacturing units for a considerable period in the primary processing area. The country's long supply line (except from Australia), however, has encouraged the growth of a variety of industries producing for the local market. Even before World War II roughly a fifth of the work force was engaged in manufacturing, and its contribution to national product was of the same order. Manufacturing in the early 1970s accounted for over one-quarter of the national product.

Financial services and foreign trade. Banking was established early in New Zealand, and two British, two Australian, and one (the largest) state-owned bank cover the country with a branch network. All commercial bank branches are linked by a computerized check-clearing system which is being extended to embrace other means of transferring money. Traditionally, banks have played a large part in the financing of external trade.

Foreign trade has always been of great importance in New Zealand, and few countries have a higher ratio of exports to domestic production. Exports of dairy products are handled by a central cooperative agency, meat by individual processors (some cooperative), and wool by auction. Imports are increasingly obtained directly by manufacturers, but there are still major importing wholesalers.

In the last few decades, as industrialization has proceeded, there has been a steady trend toward less fabricated imports and a switch away from consumer goods to capital goods. This trend may be expected to continue, but there is now more emphasis on diversifying exports, a task difficult in a country heavily dependent on livestock production but essential because of British entry into the European Common Market.

Transportation. In spite of the rugged nature of the country, most of the inhabited areas of New Zealand are readily accessible; the road system is good even in rural districts, and modern freeways have been built around the main cities. Bus services link most centres. The difficult country makes for slow journeys, although the distances involved are seldom great.

A railway network, owned and operated by the state, comprises a main trunk line spanning both islands via roll-on ferries and branch lines linking most large towns and many smaller ones. Narrow tunnels limit the gauge of line, and until recently precluded the introduction of fast expresses. Rail travel is notoriously slow, in spite of the widespread use of diesel locomotives, and trains carry a decreasing number of passengers. The rail service is, however, efficient for large-scale movement of goods over considerable distances. Regulations generally forbid cartage of goods by road over routes that compete with the rail.

The difficult terrain has greatly encouraged air travel in New Zealand; most provincial towns have airports, and all major centres are linked by a good air service. The internal airline, New Zealand National Airways, is a public corporation, but smaller private operators run goods and tourist services. Air New Zealand and several foreign airlines provide New Zealand with international flights. There are international air terminals at Auckland, Christchurch, and a smaller terminal at Wellington.

Economic policy. State intervention in the economy has a long history in New Zealand. Many state trading enterprises thus had their origin long before the first Labour government was elected (in 1935), and the National Party has not tried to dismantle the social security system introduced by that government. State intervention is facilitated by the absence of provincial or state governments, and, since 1950, by the unicameral parliament. State influence has steadily increased through attempts—mostly unsuccessful—to stabilize the economy.

Encouragement of local manufactures and exports

In the 1930s deliberate efforts were made to encourage industry, and World War II, with its attendant disruption of imports, greatly boosted local manufacturing. Further impetus was lent by the government, which after the war continued import licensing. By the end of the 1960s the scope of import substitution was virtually exhausted, and policy switched to the encouragement of exports of manufactures other than primary produce and forest products. Stimulated by export incentives and devaluation, and spurred on by a short-term recession of local demand, the expansion of manufactured exports of a nontraditional sort was spectacular.

There is little variation in developed countries in the proportion of taxation to national income, and, despite common views to the contrary, New Zealanders are not heavily taxed by world standards. Direct taxation (taxation of personal income) is relied upon to an unusual extent, and is steeply progressive up to the relatively low maximum of 50 percent. Because indirect taxation (taxation of goods) is politically anathema, the cost of living has been less directly affected by taxation than in many countries. Strong wage inflation has, however, more than compensated for this.

The penchant for centralization, noticeable in New Zealand government circles, has its analogue in well-developed employer associations and a reasonably closely knit Federation of Labour.

Contemporary economic policy is directed mainly to constraining the growth of consumption sufficiently to enable diversification of production and exports in order best to meet the difficulties likely when access to the British market is substantially reduced.

New Zealand's essential economic problem is that no means have been discovered to raise productivity growth consistently above 2 percent a year without threatening external insolvency. Yet all sections of the community, in their demands on the national product, act as if real growth were much higher than this. With rising import prices, continuing weakness in export prices, and the possibility of acute difficulty in securing market access for some primary products in the mid-'70s, the economic picture is somewhat cloudy. Problems are compounded by the newly evident tendency for skills and initiative to emigrate when economic conditions at home deteriorate.

ADMINISTRATION AND SOCIAL CONDITIONS

Government. *Constitutional practice.* New Zealand's form of government reflects its historical association with Britain. Sovereign power is held by an elected single-chamber parliament (the House of Representatives), the decisions of which must command a majority of its members. A two-party system operates; the party holding a majority of the 80 European and four Maori seats is automatically the government. The leader of the governing party is the prime minister, who, with his ministers responsible for different aspects of government, forms a policy-making Cabinet within the government. Most of the legislation introduced to parliament is initiated by the government, and more particularly, by Cabinet. Parliament may veto legislation, but, since the majority party forms the government, this seldom happens.

The two-party system

Bills passed by parliament are sent to the governor general for the royal assent and become law whenever the bill specifies. The governor general—technically the Queen's representative in New Zealand—performs a variety of formal constitutional functions, historical relics from New Zealand's colonial days. In effect, apart from his social duties, he is the servant of the majority party in the House of Representatives, exercising real power only in the unlikely event of the House of Representatives having no party with a clear majority, or if the majority party has no accepted leader. The governor general is appointed by the Queen, in consultation with the New Zealand government, for a five-year period. A high commissioner represents the British government.

The party system. There is universal suffrage over the age of 20. Triennial elections are generally fought on a party affiliation basis, each electorate returning only one of the candidates presented, on a majority-vote basis. New Zealand's two-party system probably reflects the lack of deep ideological divisions in the country; conflict between groups and parties is usually economic, a field in which compromise is generally possible. The National Party bases its support upon a rural-urban coalition of interest and is primarily a conservative welfare party. The Labour Party is mainly supported by trade unions and urban working people; its policy on welfare is scarcely less conservative than that of its opponents.

The two-party system works well in New Zealand; it produces strong government because great power is given to Cabinet, which is ultimately responsible to the people, and which cannot escape responsibility for its decisions. The system also has the advantage of ensuring popular control of its political leaders. Retention of parliamentary power depends upon political competence, and a critical role of the opposition is to focus attention on controversial issues in an attempt to undermine the government's following. Strong political consciousness, however, is not a characteristic of New Zealanders, who are generally content with their triennial opportunity to unseat the governing party.

Departmental organization. The business of government is carried out by 40 government departments, each being responsible for one or more aspects of administra-

tion. Since New Zealand is a welfare state, the fields of government activity are wide and varied. Each department is ruled by a permanent head who is responsible to his minister for the administration of his department. Heads of departments and their officials do not change with a change of government, thus ensuring a continuity of administration.

Work of the parliamentary commissioner (ombudsman)

The work of government administration in New Zealand tends to be bureaucratic in the extreme, and as a check upon possible administrative injustices, the office of parliamentary commissioner (ombudsman) was set up in 1962. As an impartial authority, to whom complaints against decisions of government departments may be brought, he has wide powers of investigation and a final responsibility to the prime minister and parliament. The standard of the civil service is high, but the use of the ombudsman by the general public has confirmed the need for an appeal against bureaucratic authority.

In addition to government departments there are also many government corporations—commercial ventures of national importance in which the government is the sole or a major shareholder. These include the New Zealand Broadcasting Corporation, the New Zealand National Airways Corporation, and Air New Zealand.

Local government. Local government, which has very limited power in all but peculiarly local matters, is directly empowered by parliamentary statute. Local authorities are thus dependent upon the central government, but are at the same time legally independent bodies. The definition of their function and powers is under constant revision as adjustments are made to changing conditions.

Local bodies perform general purpose duties such as those of borough and town council, or else consist of ad hoc authorities with specialized functions such as harbour and hospital boards. Every local authority activity is controlled by an elected board of local members, whose work is generally honorary. The platform for election is sometimes based on party affiliation, although this does not noticeably affect the working of the boards.

Legal system. New Zealand derives from the common law of Britain certain statutes passed before 1947 by the British Parliament. New Zealand law usually follows the precedents of English law, and the British Privy Council acts as the final court of appeal for New Zealand. The law is administered by the Justice Department through its general courts and upheld by the police force, which is also maintained by the government. The hierarchy of courts dealing with civil and criminal cases includes magistrates' courts, the Supreme Court, and the Court of Appeal, in that order. The jury system operates, and legal aid is provided for those who lack financial means.

Lesser courts

In addition to the major courts, there are several specialized courts such as a Compensation Court which deals with claims for workers' compensation. Children's courts deal with most juvenile offenses. As far as possible these courts do not use a courtroom, their proceedings are not public, and convictions need not be recorded. The aim of the penal system is not so much to punish as to reform and rehabilitate. Attempts are made at many levels to use detention periods constructively, and include trade training, group therapy, and creative relaxation. Six prerelease hostels are in operation, and postrelease care is provided in many cases.

Education. Education in New Zealand is free and compulsory between the ages of six and 15, but almost all children start school at five. Many children attend preschool classes, which are financially assisted by the Education Department, after which they proceed to a state primary school. State secondary schools offer a broad general academic education with some emphasis on technical training. Students are prepared for an attainment examination (the school certificate), and for tertiary studies, which are provided free to students fulfilling certain academic prerequisites. There is also a small number of independent private schools taking fee-paying students; these schools receive considerable financial assistance from the state and are also inspected by the Education Department. There are six universities and one agricultural college, as well as several technical institutes. The New Zealand Technical Correspondence Institute provides courses for those who live outside a main city.

Since the 1960s, Maori education has received great attention, and in an effort to improve the attainment level of Maori students, extra financial assistance has been provided through the Maori Education Foundation. The difficulty is more social than academic, and its solution may well be within the Maori community itself.

Health, welfare, and housing. The control of public health is vested in the Health Department, which supervises preventive medicine and education as well as medical and hospital services. Doctors and chemists are usually private practitioners, but under the medical benefits scheme the state remits a large proportion of the cost to the individual. Hospitalization in general, maternity, and mental public hospitals is free, as is any specialist treatment required. The cost of many pharmaceuticals is also met by the state.

Many welfare services are provided by the state, but some are insufficient to meet the demand. Increased longevity has produced a geriatric problem with which the combined resources of church and state hardly begin to deal. An increase in illegitimate births has reopened children's homes and orphanages, while the problems of the unmarried mother have not yet been realistically faced.

The government's social security scheme embraces a system of noncontributory civil pensions, in addition to the medical benefits mentioned above. These include universal superannuation (a taxable grant payable to all citizens over 65) and, subject to a means test, benefits for the aged, widows, orphans, invalids, miners, sickness, unemployment, and emergencies. A family benefit is also paid, irrespective of income, to every mother for each dependent child.

Most New Zealanders attempt to own their own homes, usually bungalow-type (one-level) houses with three bedrooms, set in a garden. In the main cities, lack of space is modifying this concept; gardens are smaller and many elderly people are investing in apartments in high-rise buildings. Mortgage finance at a low interest rate is made available by the state to assist house purchase by those whose income is beneath a certain figure. Low-income rental housing is also provided by the state, and generally follows the traditional, land-consuming pattern. These houses are frequently concentrated in one area, producing a considerable sociological problem, in that there is little occupational range in the neighbourhood and no architectural diversification. The state also subsidizes a pensioners' accommodation built by local authorities, and gives financial assistance to Maori in buying and maintaining their homes. It has also built urban hostels for Maori boys and girls away from home for the first time.

CULTURAL LIFE

The cultural milieu in New Zealand is conditioned by a society which is egalitarian in the extreme. The introduction of a welfare state in the 1930s was a radical affirmation of the country's belief in the right of everyone to a good life, and showed a direct concern for social justice. The less attractive aspect of this thinking appears in a strong tendency toward conformity to an imaginary national norm. There is little tolerance for the eccentric in any field, and New Zealanders appear to be content with the dull uniformity to be found not only in thought and speech patterns, but also in the monotony of domestic architecture, the lack of individuality in status goods, and a ready acquiescence to authority. Contemporary New Zealanders appear hesitant to use the pragmatic initiative so obvious in the first century of European settlement, although since the stimulus was always economic rather than cultural, new signs of life may perhaps emerge with the difficulties that will attend British membership of the European Common Market.

Cultural conformity

A new element is being introduced to the social climate by the emergence of the Maori people, determined to make their contribution as a forceful, enlivening minority. Articulate Maori leaders are concerned to encourage their people, to improve the standard of Maori education,

and to increase Maori self-respect and confidence in their own culture. The activities of the Maori Council, the appointment of the first Maori bishop to a European see, and the introduction of the Maori language in schools are indications that the Maori, by insisting on making his contribution in his own way, is endeavouring to combine the best of both worlds. Perhaps, in some elements of the Maori way of life, New Zealand society may find a solution to the problems of materialism characteristic of Western countries.

The first century of artistic endeavour in New Zealand was one of experimentation and exploration in terms of the classical European tradition. From this, artists and writers struggled painfully toward an acceptance of the unique quality of the country, although this very acceptance embodied a homesickness for the cultural climate of the Northern Hemisphere, exemplified in the life and works of the writer Katherine Mansfield. The Depression of the 1930s, however, saw the appearance of a group of lively and competent poets concerned with themselves as New Zealanders, but the great promise and vitality of many of these writers has not been fulfilled. Contemporary New Zealand possesses a considerable number of writers and artists who, while very much of their society, are generally concerned with personal analysis and introspection.

Maori artistic expression, except for decorative design, was completely oral before the arrival of Europeans, and much has been lost. New songs and chants are being written in Maori, however, as well as some poetry and prose. An impressive number of Maori are now writing in English, a fact which may have considerable influence on the future of New Zealand literature.

Although most New Zealanders are generally suspicious of cultural activities, there is a lively amateur dramatic tradition throughout the country. There is also popular interest and support for highland pipe and brass bands. Pottery is a creative activity that has been consistently popular, and New Zealand's best potters are well able to compete with the world's masters of this art.

The small size of the population poses problems for cultural development. Scattered widely across two islands, it is capable of supporting only three small professional theatre companies and one opera and ballet company, and it is doubtful whether even these can survive without government support. The government subsidizes the activities of chamber music societies, individual writers, artists, and musicians, and assists art galleries and museums.

The national orchestra is completely supported by the government via the New Zealand Broadcasting Corporation, which controls broadcasting and a single-channel, black-and-white television system of a reasonably high standard; a second television channel is planned. The New Zealand Broadcasting Corporation also publishes *The New Zealand Listener*, New Zealand's only weekly magazine concerned with cultural discussion.

Newspapers in New Zealand provide a high standard of reporting with substantial coverage of world news, although this is largely dependent on foreign agencies. There is little cultural content in the daily paper, but there is also little sensationalism. The mildly conservative political colour of most papers suggests the moderate attitudes of their readers.

Sport is the main leisure time occupation of most of the population. There is widespread participation in most major sports, particularly rugby football, over which an almost religious and certainly patriotic fervour is aroused, and horse racing. The climate and the variety of terrain allow for year-round activity in sports of all types.

THE OUTLOOK

New Zealand is a small but affluent nation with a dominant European heritage, yet remote from its cultural forebears, and at the same time a multiracial society. Racial self-consciousness has always been an element in the relationship between Maori and European, but with urbanization and the consequent alienation from tribal roots, the contemporary Maori must solve his identity problem as well as make his proper contribution to the community. The Maori effort on both counts is constructive, forward-looking, and fundamentally practical; nevertheless, the social tensions endemic in such a situation are likely to become exacerbated in a world increasingly dominated by protectionist trade concepts, especially in the case of the agricultural products on which New Zealand depends. Despite some signs of the centrifugal tendencies in society which are afflicting other nations of the Western world, the New Zealand of the 1970s remains a strong society, although the initiative which characterized the early settlers seems to have been replaced by a certain complacency and lack of drive.

BIBLIOGRAPHY. A.H. MCLINTOCK (ed.), *An Encyclopaedia of New Zealand*, 3 vol. (1966), a compendium of articles on all aspects of New Zealand, and *A Descriptive Atlas of New Zealand* (1959), the definitive atlas; *The New Zealand Official Year Book* (annual), facts about the country and its people; G.H. BROWN and H. KEITH, *An Introduction to New Zealand Painting, 1839–1967* (1969), an illustrated historical survey of New Zealand painting; PETER H. BUCK, *The Coming of the Maori*, 2nd ed. (1950), an historical–anthropological study by a great Maori scholar and leader; J.B. CONDLIFFE, *The Welfare State in New Zealand* (1959), an economic analysis of welfare policies; A. CURNOW (comp.), *The Penguin Book of New Zealand Verse* (1960), an anthology of representative verse; K.B. CUMBERLAND and J.W. FOX, *New Zealand: A Regional View*, 2nd ed. (1964); W.H. OLIVER, *The Story of New Zealand*, 2nd ed. (1963), an historical essay by a leading academic; J.W. and M.A. ROWE, *New Zealand* (1967), a general treatment of modern New Zealand; M. SHADBOLT (ed.), *The Shell Guide to New Zealand* (1968), a photographic and textual description; C. WESTRATE, *Portrait of a Modern Mixed Economy: New Zealand* (1959), a study of contemporary economic institutions.

(J.W.R./M.A.Ro.)

New Zealand, History of

European settlement in New Zealand began in the late 18th century; its Polynesian history goes back to the early Christian centuries. In the 19th century the country became predominantly British in the face of determined resistance from the indigenous Maori; subsequently the two races achieved considerable harmony. On land bought or wrested from the Maori, the settlers established a society dependent upon agrarian resources and overseas markets.

DISCOVERY

Little is known about the time or the manner of arrival of the first inhabitants except that they came over the sea from the north. Chronologies based upon Maori oral traditions are uncertain. Archaeology indicates a much longer Polynesian occupancy than had earlier been supposed; other studies point to accidental drift voyages as the means of settlement. No more is certain than that the Maori represent the southernmost point reached by Polynesian expansion in the Pacific.

The Polynesian period has been divided into an early "Archaic" and a later "Classic Maori" phase, but the transition from early to late is unexplored. In the South Island, if not elsewhere, the first Polynesians found moas (flightless birds) in immense numbers on tussock grasslands, and these became their major food supply. The "Classic Maori" encountered later by Europeans had only faint memories of the moa and was an agriculturalist. The 18th-century Maori population was most dense in the warmer northern parts of the country, where the Maori variant of Polynesian culture had reached its high point, particularly in the arts of war, canoe construction, building, weaving, and agriculture. Estimates of the extent of the Maori population vary widely; it was probably not more than a few hundred thousand. Numbers were to drop sharply in the first century of European influence, from the late 18th to the late 19th century.

The first European to arrive in New Zealand was a Dutch sailor, Abel Janszoon Tasman, who sighted the coast of Westland in December 1642. His sole attempt to land brought only a clash with a South Island tribe in which several of his men were killed. After his voyage the western coast of New Zealand became a line upon Euro-

pean charts and was thought of as the possible western edge of a great southern continent.

Captain Cook's explorations

In 1769–70 James Cook, the British naval officer and explorer, completed Tasman's work by circumnavigating the two major islands and charting them with a remarkable degree of accuracy. His first contact with the Maori was violent, but harmonious relations were established later. On this and on later voyages, Cook, with the explorer and naturalist Joseph Banks, made the first systematic observations of Maori life and culture. Cook's journal, published as *A Voyage Towards the South Pole and Round the World* (1777), brought the knowledge of a new land to Europeans. He stressed the intelligence of the natives, and the suitability of the country for colonization, and soon colonists, as well as other discoverers, followed Cook to the country he had made known.

EARLY EUROPEAN SETTLEMENT

Apart from convicts escaping from Australia and sailors seeking asylum with Maori tribes, the first European New Zealanders sought profits—from sealskins, timber, New Zealand flax, and whale oil. Early New Zealand was an offshoot of Australian enterprise in whaling and other activities; Sydney, New South Wales, founded as a convict settlement in 1788, became a base for South Pacific whaling; and Kororareka (now called Russell), in the far north of New Zealand, became a stopping place for American, British, and French deep-sea whalers. Around both islands Australian firms set up tiny settlements of land-based bay whalers. Traders supplying whalers drew Maori into their economic activity, buying provisions, supplying trade goods, implements, muskets, and rum. Initially the Maoris welcomed the newcomers; while the tribes were secure, the European was a useful dependent.

Maori went overseas, some as far as England. A northern chief, Hongi Hika, amassed presents in England, which he exchanged in Australia for muskets; back in New Zealand he waged devastating war on hereditary enemies. The musket travelled south; a series of tribal wars, spreading from north to south, displaced populations and disturbed landholdings, especially in the Waikato, Taranaki, and Cook Strait areas. Europeans were soon to found colonies in these unsettled regions. Missionaries quickly followed the traders. Between 1814 and 1838 Anglicans, Wesleyans, and Roman Catholics set up stations. Success was initially slow, but by the mid-19th century most Maori adhered, for varying reasons, to some form of Christianity.

European colonies

All of these newcomers had a profound effect upon Maori life. Warfare and disease reduced numbers; new values, pursuits, and beliefs modified tribal structure. Christianity cut across the sanctions and prohibitions that had supplied social cohesion. A capitalist economy, to which Maori were introduced both by traders offering new inducements (for instance, the brief demand for flax) and missionaries bringing new agricultural techniques, affected the whole material basis of life. At first in the north, and later over the whole country, a process of adjustment began, which has not yet been completed. By the late 1830s, chiefly through the Australian link, New Zealand had been joined to Europe. Settlers numbered at least some hundreds, and there were certain to be more. Colonization schemes were afoot in Great Britain, and Australian graziers were buying land from the Maori. These circumstances determined British policy.

ANNEXATION AND FURTHER SETTLEMENT

In 1838 the British government decided upon at least partial annexation. It commissioned William Hobson, naval officer, as lieutenant governor and consul to the Maori chiefs in July 1839; he became governor in 1841. Hobson, in the event, annexed the whole country, the North Island by the right of cession from the Maori chiefs and the South Island by the right of discovery. From June 1839 to May 1841 New Zealand was legally part of New South Wales. Before declaring the annexation of New Zealand, Hobson went through a process of discussion with the northern chiefs from which emerged the so-called Treaty of Waitangi (February 1840). Under this instrument, Maori ceded sovereignty to the crown in return for protection and guaranteed possession of their lands; they also agreed to sell land only to the crown. Hobson promised an investigation into past "sales" of land to private individuals to ensure fair dealing. This treaty imposed a strong moral obligation upon the British government to act as guardian to the Maori.

The Treaty of Waitangi

Even before annexation had been proclaimed, the first organized planting of an English colony was under way. The New Zealand Association, founded in 1837 to colonize on the principles laid down by Edward Gibbon Wakefield, sent a survey ship, the "Tory," in 1839. The agents on board were to buy land in both islands around Cook Strait. The company moved hastily because its founders were aware that British annexation was likely and would entail a crown monopoly of land sales and a consequent increase in price. "Purchases" were effected in great haste before Hobson could bring to an end such private transactions. Little effort was made to seek out the true Maori owners; this would have been difficult anyway, as Maori ownership was communal and titles had been disturbed by the musket warfare of the preceding quarter century. The company, combining skillful propaganda with outright trickery and brutality, enforced its claim to the land upon which New Plymouth, Wanganui, and Wellington in the North Island and Nelson in the South Island were founded in the 1840s. Later, through the crown, it secured other areas in the South Island where Otago (1848) and Canterbury (1850) were planted by separate associations. Meanwhile, Hobson moved the seat of government south from the Bay of Islands, bringing Auckland into existence (1841).

In the early 1840s settlement and government began to alarm the Maori. In the Cook Strait area a formidable chief, Te Rauparaha, obstructed settlement. Near the Bay of Islands there was open warfare, and Kororareka was repeatedly raided. Neither Hobson (died 1842) nor his successor, Robert Fitzroy, was able to overcome the Maori. George (afterward Sir George) Grey, who became governor in 1845, had money and troops and the will to use them. His victories brought a peace that lasted from 1847 until 1860. Hone Heke, the northern leader, was thoroughly defeated (1846), and in the south a likely uprising was prevented. Racial strife had been accompanied by economic distress. In the mid-1840s the nascent economy was depressed until the Australian gold rushes of the 1850s offered a market for foodstuffs to the New Zealand farmer, settler, and Maori alike.

War with the Maori

By the end of the 1840s racial and economic trouble gave way to political agitation. The leading settlements, apart from Auckland, began to campaign for representative government in place of Grey's personal rule. He, while refusing to give way, helped to draft the New Zealand Constitution Act (1852), which was designed to meet all the settler demands. He sought not to prevent the introduction of self-government but to delay it until he had determined both native and land policy. He wished to begin the rapid assimilation of the Maori (with whom his relations were excellent) to the British pattern. He also wished to bring in a land policy that would safeguard the small farmer against the great owner. He believed he had secured these goals by the time of his departure at the end of 1853.

Responsible government. When the Constitution Act came into operation, New Zealand was divided into six provinces—Auckland, New Plymouth (Taranaki), Wellington, Nelson, Canterbury, and Otago—each with a superintendent and a Provincial council. The central government (General Assembly) consisted of a governor, a Legislative Council nominated by the crown, and a House of Representatives elected upon a low property franchise for a five-year term. This General Assembly did not meet until 1854; it then embarked upon a quarrel with the acting governor, Colonel Robert Henry Wynyard, that was not ended until the achievement of full responsible government—*i.e.*, a system under which the governor could act in domestic matters only upon the advice of ministers enjoying the confidence of the elected chamber. Henry Sewell and James FitzGerald, of Canter-

bury, led the representatives in this struggle, against the opposition of E.G. Wakefield, who, having first moved the resolution for responsible government, then secretly opposed it while serving as extra-official adviser to the acting governor. The Colonial Office conceded responsible government in 1856. The next governor, Thomas Gore Browne, reserved Maori affairs to the competence of the governor alone.

Provincial institutions

For most purposes, during the 1850s New Zealand was administered not by central but by provincial institutions. These authorities (nine in number by the time of their abolition in 1876) directly affected the settler through their administration of land and control of immigration and public works. The native department, directly under the governor, bought land from the Maori; the provincial governments settled it, regulated immigration, and built roads and bridges. Until the wars of the 1860s the central legislature was less important, though its ultimate authority remained.

Each province disposed of a revenue arising from land sales, and upon this revenue depended its strength. Canterbury and Otago, with hardly any Maori, grew wealthy, spending their money upon communications, immigration, and education. Other provinces were either less fortunate or less wise and enjoyed smaller success. In the North Island, numerous and anxious Maori held on to desirable land. Here most of the land available for settlement had been taken up by the end of the 1850s, a good deal of it by speculators, and some of it was given away to attract immigrants. The island remained largely without roads until the 1870s, so impecunious were its governments. But by that time the major obstacle to settlement had been removed—the continuing power of the tribes. This was the result of a decade of war.

Racial conflict. In the 1850s race relations deteriorated. The settler population and the demand for land, especially pastoral land, increased. Many Maori, fearing for their future, became reluctant to sell more land. In the Taranaki Province, where the land shortage was acute, both settlers and those Maori willing to sell were opposed by Wiremu Kingi (Te Rangitake), chief of Te Atiawa. In the Waikato, where good land was coveted by settlers and speculators, an elderly chief, Te Wherowhero, was elected "king" in 1857, largely by the Waikato and Maniopoto tribes, and reigned as King Potatau I. This "king" movement and also the unrest in the Taranaki headed by Wiremu Kingi (the two movements remained distinct though related) were opposed to further land sales.

Land sales

The likelihood of conflict was not reduced by any particular wisdom in government policy. Gore Browne was guided in native policy by the head of the Native Land Purchase Department, Donald McLean, who, responsive to settler demands, increased pressure upon potential sellers. Grey's caution and his recognition that a chief could veto sales proposed by any section of his tribe were forgotten. McLean sowed a rich harvest of distrust. Christopher Richmond, the member of Cabinet in charge of native affairs, was also a Taranaki representative, fully responsive to the needs of his settler neighbours. The central ministry, theoretically unconcerned with native policy, could not, despite the promise of protection made to the Maori in the Treaty of Waitangi, neglect a matter so vital to the colony's future. In 1859 the representative of the crown unwittingly supplied the occasion for the outbreak of civil strife.

Gore Browne accepted an offer to sell from a Taranaki subchief, Te Teira, and ignored the veto imposed by the paramount chief, Wiremu Kingi. Early in 1860 troops were used to dislodge Kingi from the land in question, the Waitara block. A decade of fighting began. In 1861 Grey was sent back for a second term as governor in the hope that he would again prove to be a peacemaker. In fact he accelerated the extension of conflict. Fearing that Auckland was menaced by the followers of the Maori king, he took defensive measures which could easily be interpreted as acts of aggression, and the fighting subsequently spread from Taranaki to the Waikato. Imperial troops, colonial militia, Maori allies (for not all the tribes supported the Maori nationalist movement) had no easy task, but their victory could not be postponed for long. By the mid-1860s Maori resistance in the Taranaki and Waikato was ended. But the "king" tribes were by no means crushed, and the fear that they would embark upon war again haunted the colony for many years.

In the later 1860s the fighting was of a different character, in which religion acted as a last, desperate stiffener of Maori resistance. Pai Marire (Hauhauism), an amalgam of Christian and primitive beliefs, was the first of many cults in which the Maori, rejecting the religion of settler and missionary, put their own imprint upon Christianity. Toward the end of the decade, Te Kooti organized resistance on the east coast. He was the founder of another religious cult as well as a guerilla of some note; his adaptation of Christianity, Rin-gatu, still had thousands of followers in the mid-20th century. Te Kooti was never finally defeated, but by the early 1870s he was forced to retreat into the "king country" (the centre of the island), where he devoted the rest of his life to religious leadership.

An uneasy peace settled upon the colony in 1870. Casualties had not been high, but the loss of life was serious for the tribes concerned. Especially in those areas in which the Maori king retained some authority, defeat led to a period of withdrawal from settler society. Resentment was deepened by a punitive policy of land confiscation adopted by the victors, a policy improper in its nature and made worse in some places by undiscriminating application to "guilty" and "innocent" tribes alike. The Maori future looked black. By the 1862 Native Land Act private land transactions between settler and Maori had been legalized, and during the next 40 years the Maori lost most of their best land. Many years were to elapse before Maori numbers, morale, and confidence could revive over the whole country.

The Native Land Act

DEVELOPMENT OF THE COLONY

Economic growth in the North Island had been considerably retarded by the wars. Meanwhile, the South Island, especially Canterbury and Otago, had grown increasingly prosperous. Pastoral farming expanded steadily, and the discovery of gold, first in Otago and then on the west coast, led to a sudden boom in production and trade. Population rose as diggers poured in; economic life quickened as gold brought prosperity, less to the digger than to bankers, merchants, land sellers, and farmers supplying provisions. The South Island share of the European population jumped from about 40 percent to 60 percent during the 1860s. The North Island did not recover its previous lead until the 20th century.

Discovery of gold

Attempts by other provinces to emulate the development of Canterbury and Otago normally ended in embarrassment (in one case in bankruptcy) as money was recklessly borrowed and spent. To preserve the colony's reputation, the central government in 1867 banned further provincial overseas borrowing. About this time depression struck the greater part of the country, especially the South Island, where the first alluvial gold had by then been worked out. The South Island was thus looking for a stimulus, while the ending of the wars now made further development possible in the North Island. It was widely agreed that only the central government could adequately revitalize the economy.

In 1870 a development policy was provided by Julius Vogel, the colonial treasurer, who was convinced (not altogether accurately) that New Zealand was bursting with potential resources needing no more than the stimulus of capital and labour for their exploitation. He borrowed overseas capital for public works on an unprecedented scale and swelled the labour force with assisted immigrants.

Julius Vogel's development policy

Not all Vogel's schemes were wisely conceived; the prosperity of the mid-1870s was more an investment boom than a solid growth of productivity. But the colony ended the decade with a doubled population (about 500,000) and the beginnings of efficient internal and external communications. Roads, bridges, railways, and telegraph systems had been built, and overseas shipping services improved.

Private lending agencies contributed to the boom, and in a heady atmosphere land values and interest rates climbed alarmingly. The public debt greatly increased, and many of the men who had acquired land were in desperate financial straits. Falling overseas prices for farm products (chiefly wool and wheat), a declining gold output, a cautious note in government finance, and widespread unemployment marked the 1880s. Emigrant ships discharged their passengers at ports where unemployment was already rife. There had been growth in the 1870s, but it was succeeded by a depression that lasted until 1895.

Economic depression

Vogel abolished the provincial governments in 1876. They had earned his enmity by refusing to allow their lands to be used as security for public works and by blocking a forest conservation scheme. Essentially, they became outmoded when, in the early 1870s, the initiative in development passed to the central government. Provincial governments had been set up to colonize their districts; when the centre assumed this function they lost their raison d'être. Abolition came fairly painlessly; it was an affront more to local pride than to local prosperity. Only in Otago was there a strong attempt to resist change. Thereafter, provincial interests were long pursued by the respective delegates in the General Assembly, whose achievements were in no way diminished by the lack of particularist (provincial) institutions.

The governments of the 1880s, though led by men of some ability and imagination, such as Sir Robert Stout and Sir Harry Atkinson, did not deal effectively with the depression. The time-honoured remedy, spending loan money on development, was not fully given up until 1887. The basic problem was to find productive work for the country's labour force; closer land settlement was the remedy suggested in the 1880s and applied in the 1890s. Great areas, especially in the South Island, had fallen to large owners; these "monopolists" were attacked by the radicals, though probably the pastoral industry could not have been established under any other system. William Rolleston, minister of lands from 1882 to 1884, first proposed that the state should help men to become small farmers as state tenants; John McKenzie and the Liberal government applied this remedy with vigour in the 1890s. But closer settlement and intensive farming did not succeed until small farmers had a product to export and gained a good price for that product. Refrigeration and rising world prices provided the answer. It became possible in the 1880s to send to Great Britain refrigerated cargoes of butter, cheese, and meat; this encouraged the spread of small-scale intensive farming.

The Liberal era (1891–1912). The energetic Liberal government led by John Ballance, which took office in 1891, accelerated the process of change. It opened more land (much of it bought from the Maori), established farmers on perpetual state leaseholds, provided credit for land purchase and improvements, and built roads. So came into existence great dairying and meat producing areas, especially in the North Island. Dairy, meat, and also wool prices rose in about 1895 and stayed generally high until about 1920.

This economic stimulus was not limited to farmers. Urban distress had been serious in the 1880s, for many recent immigrants had been townsmen who had stayed in New Zealand towns on arrival. The ultimate cure for their distress was for the towns to share in the farmers' high prices. Urban New Zealand depended upon the prosperity of the country. But other remedies were considered, and some of them were applied. In the 1880s there was serious discussion of insurance against sickness, poverty, and old age; the Old Age Pensions Act of 1898 was the first measure of social security. Tariff protection to foster industrial employment was halfheartedly applied in the late 1880s. Revelations of oppression in industry led, in the 1890s, to a labour code to protect workers.

But the chief Liberal industrial policy, formulated by William Pember Reeves, minister of labour from 1892 to 1896, was to encourage trade unions and to introduce, in the Industrial Conciliation and Arbitration Act of 1894, a conciliation and compulsory arbitration system intended to end industrial unrest and give the unions the means of protecting their members. The growth of unions was stimulated by the fact that only through them could the workers use the system. Reeves's act, amended and occasionally suspended but still essentially his own handiwork, has never been repealed. It has enabled the worker in good times to resist wage cuts and to press for increases; but it has not managed to prevent cuts and unemployment when falling overseas prices have brought depression to New Zealand. It has not been strikingly radical in effect; employers and governments have used it to break strikes, such as that of miners at Waihi in 1912. It has built up the power of those majority elements in the unions that prefer coming to terms with capitalism to any effort to destroy it. Some occupations, such as transport, cargo handling, meat processing, and mining, have fostered unions keen to relinquish arbitration for direct action, but they have been in a minority and seldom, in the long run, successful. Farmers and governments have usually acted with severity in disputes affecting the movement of exports.

The growth of trade unions

The Liberal era, from 1891 to 1912, transformed political life. Previously politics had not been marked by neat party divisions. Local advantage had determined political behaviour in the development period during and after the 1870s; men had argued over the scope and details of policies and had advanced the claims of locality and province for a proper share of largess. Acute depression ended development and with it the politics of local advantage. In 1890 the Liberals began to act as a more or less unified party. Their 20 years in office, the success of their land and labour policies, and the formidable qualities of leadership discovered in Richard John Seddon, premier from 1893 to his death in 1906, welded the Liberals into a fairly coherent parliamentary and popular party.

R.J. Seddon as premier

Seddon was a portent of a new age. In 1893 this energetic goldfields trader-turned-politician provided a sharp contrast to the gentlemanly premiers who had preceded him. But his crudeness assisted rather than hindered the attainment of popularity none of them had known. He was devoted to political success and skilled in the manipulation of the means of success—parliamentary procedure, patronage, and party organization. By the time of his death he had established a kind of elective despotism over the country.

THE 20TH CENTURY

Seddon's successors, in his own and in other parties, were of the same stamp—men of the people devoted to a political career. Politics ceased to be a duty of the well-to-do amateur. The Liberal government, under Sir Joseph Ward, survived Seddon by six years. In 1912 it fell before a new party, Reform, led by a dairy farmer, William Ferguson Massey, prime minister until 1925. Based on prospering farmers and townsmen, especially of the North Island, and closely connected with their professional organizations, it was more narrowly sectional than the Liberals had been. Except for views borrowed from the Liberals, it had little positive policy. Reform made much of a promise to enable the state leaseholder to buy the freehold of his farm at original valuation; this promise was an emotional rallying cry for conservatives fearing land nationalization and complete socialism. Only a small minority of farmers were state tenants, and not all bought the freehold when the Reform government gave them the chance.

While the Liberals lost support in rural areas, they were further weakened by urban left-wing defections, which eventually led to a separate Labour Party. Four Labour members were returned in 1911. The initiative, on the right and on the left, was passing to other parties, and the Liberals were gradually eclipsed. The period before World War I was one of discontent and anxiety. Prosperity, though still considerable, had somewhat declined. The farmers were disturbed by what they took to be the threat of socialism, detected in the radicalism of a Liberal minority, but chiefly in the rebirth of direct action in some trade unions. This change in temper arose from labour dissatisfaction with wage levels achieved under arbitration and from the growth of syndicalist and socialist ideas. After 1906 the Arbitration Court refused to grant further

increases of real wages. Discontent flared up in the strikes of 1912–13, the biggest occurring on the waterfront when the farmers' government, headed by Massey, vigorously repressed a strike movement that had slight overtones of social revolution.

World War I and the interwar years. New Zealand supported Great Britain in World Wars I and II, chiefly by sending men overseas and producing food and wool. In World War II the Japanese brought danger close to New Zealand's shores; in the earlier conflict the peace of the Pacific was seldom disturbed.

New Zealanders served in the Dardanelles campaign at Gallipoli and subsequently in France. In these battles they began to learn that they were a distinctive branch of the British people. The infantry "digger" of 1914–18 came to symbolize those qualities that the country most respects —courage, endurance, and resourcefulness. New Zealand lost many of its young men, a serious loss for succeeding decades.

At home the war brought prosperity, as export markets were assured and prices good. Domestic unity was hardly shaken by the antiwar feeling of a handful of left-wingers. Massey remained prime minister, but in the wartime coalition government (1915–19) Ward and the Liberals carried great weight. Reform stayed in office until 1928, led after Massey's death in 1925 by Joseph Gordon Coates. The party survived the first postwar depression, but not that of the mid-1920s. Led by Ward, the Liberals, under the new name of United Party, were victorious in 1928; they thus had to face the deepening depression of 1929–30. After Ward's death (1930) and at the height of the depression, Reform and United formed a new coalition (1931) under the premiership of George Forbes, which lasted until the election of 1935 brought in a Labour government with a large majority.

Some postwar developments were of great importance. In external affairs, Massey led a delegation to the peace conference, signed the Treaty of Versailles, and so committed New Zealand to membership in the League of Nations. New Zealand thus began to acquire the status of a sovereign state, though Massey denied this consequence. The Liberals, especially Seddon, had already taken steps toward autonomy within the empire. At the series of colonial and imperial conferences from 1887 onward, New Zealand had followed Canada and Australia in asserting its right to a voice in certain foreign policy issues. Seddon argued vehemently against British reluctance to acquire more Pacific islands while permitting German influence to grow in Samoa. New Zealand legislation to restrict Asian immigration was sharply and obstinately at variance with British policy. Western Samoa, which New Zealand had captured from the Germans in 1914 and over which it was granted a mandate in 1920, also provided occasions for British and New Zealand differences.

Sovereign status

Reform leaders professed little love for the principle of autonomy, which, in the 1920s, came to dominate Commonwealth relations. New Zealand took part in the conferences leading to the Statute of Westminster in 1931 but did not adopt the statute until 1947. But the substance of autonomy had been enjoyed long before.

The major domestic achievement of the Reform administration was a system of export marketing agencies in which authority was shared by producer and state. These failed in their short-term objective—to sustain farmers' returns while prices fell—but they laid the foundations of a collectivist marketing structure that has continued to expand.

The leading domestic phenomenon was the rise of Labour. The New Zealand Labour Party was established in 1916; in the 1920s it came to dominate working class urban electorates. But complete success eluded it. It had little to attract the rural voter until 1928, when it offered easier credit, and then the revitalized Liberals (United) offered an attractive alternative quite free from the taint of extreme socialism that still clung to Labour. The United Party went back to the old remedy of massive expenditure of borrowed money. In fact, the prospect of electoral success had by this time caused the Labour Party to translate its socialism into a series of welfare and credit proposals. The formation of the coalition between Reform and United in 1931 made Labour the official opposition. Industrial labour, notably sections influenced by semi-syndicalist ideas, was a restive ally for the party. But unemployment, the suspension of the compulsory arbitration system in 1932, and a nationwide series of wage cuts drove the two wings of the Labour movement together. Trade unionists had now learned that their welfare depended upon political power. Politicians could abolish arbitration, their main defense against poverty, and other politicians would be necessary for its restoration. The new solidarity played a large part in the Labour victory of 1935.

Labour in power

J.G. Coates was the most energetic and least conservative coalition minister. His attempts to counter depression concentrated upon the rehabilitation of the farmer as a step toward the revival of the whole country. In order to increase export receipts, he devalued the New Zealand pound from £110 to £125 per £100 sterling; he protected the farmer against foreclosure, and he set up a credit agency, the Mortgage Corporation. He also established the Reserve Bank of New Zealand. When overseas prices began to recover in 1934, the country was financially strong.

But he had done little for the multitude of unemployed. Unenviable conditions in towns and relief camps led to outbreaks of rioting and violence, to widespread discontent, and to the Labour victory. Successful in the towns, Labour also won in many rural areas, especially in the dairying districts. Prices for dairy exports were slowest to recover, and the dairy farmer was drawn by Labour promises of a guaranteed price for dairy produce and of cheap and plentiful credit. The victory was particularly notable in terms of seats, for a right-wing third party (the Democrat Party) split the conservative vote to Labour's advantage. A contest in 1938 without the Democrat Party, however, had the same result, and the opposition (the successors of Reform and United, now renamed the National Party) was rendered temporarily ineffective.

The new ministers, among whom the most notable were Peter Fraser as minister of health and education, and Walter Nash as minister of finance, showed great energy; genially led by Michael Joseph Savage, they had good fortune to govern a country to which prosperity was fast returning. The farmer was enjoying increased earnings; the worker, increased wages and shorter hours. Jobs were multiplied by a massive public works and housing program; attempts were made to stimulate secondary industry and to diversify the economy to make it less vulnerable to overseas conditions. The education system was revitalized. In 1938 the Social Security Act provided a state medical service, extended the pension system, and increased benefits. The expansion of industry was accelerated after the outbreak of World War II in 1939.

World War II and after. The alacrity with which New Zealand went to war in 1939 showed that dominion autonomy had not weakened the country's ties to sentiment with Great Britain. At first the war resembled that of 1914; troops were sent to Egypt to train for the European conflict. There they were directly involved by the enemy advance in North Africa and the Balkans and saw action in Greece, Crete, North Africa, and Italy before the final Axis collapse. After 1941 New Zealand was directly threatened by Japan, and New Zealand forces also were engaged in the Pacific. They were not withdrawn from the European fighting, and well before the end of the war the double strain upon the country's manpower, together with the demands of home production, enforced a reduction of commitments in the Pacific.

The Pacific theatre was dominated by the U.S., whose forces, after the loss of Singapore (1942), provided New Zealand's sole defense. U.S. troops were stationed in New Zealand, and New Zealand forces fought under U.S. command. The fact that disaster was averted by U.S. and not by British troops imposed a certain strain on New Zealand's loyalties. For generations security had been guaranteed by British power; now it was conferred by a foreign, though friendly, power. External relations in the postwar period reflected this new situation, chiefly

through the ANZUS pact (1951), a defensive alliance between Australia, New Zealand and the United States.

At home the total economy was mobilized in the war effort. Controls, already considerable by 1939, were extended to cover every aspect of economic life. Conscription and direction sent manpower either into the armed forces or to essential occupations; heavy taxation, war loans, bulk purchase, and controlled marketing kept the economy in a firm grip. These devices also served to keep inflation in check; together with price control and wage restraint they amounted to a complete policy of economic stabilization. These controls were applied by a Labour government that remained in power until 1949. Savage died early in the war. Fraser, his successor as prime minister, inherited a large share of the tasks of war administration and peacetime reconstruction. In economic affairs, the leading minister was Nash. No full coalition was formed, but for a few months in 1942 a War Cabinet, on which some National M.P.'s had seats, among them Coates (died 1943) and Sidney (afterward Sir Sidney) Holland existed alongside the normal Cabinet.

Holland led the revival of the National Party, which was marked in the elections of 1943 and 1946 and culminated in victory in 1949. Discontent with controls and with the rising cost of living were among the factors that caused opinion to swing away from Labour. Subsequently, the two parties remained fairly evenly balanced. The National government benefitted from its vigorous handling of a serious waterfront dispute in 1951, but in later elections its majority narrowed until Labour returned to office in 1957. But in 1960 the National Party, led by Keith Holyoake, was returned to power, which it maintained into the 1970s. Holyoake resigned in 1972 and was replaced by Deputy Prime Minister John Marshall. By the time of the 1966 election the two parties agreed that the economy had to be controlled, secondary industry encouraged, and the welfare state maintained. But the nation's prosperity was precarious, and this led, especially in rural areas, to emergence of a Social Credit movement, which in 1966 won its first parliamentary seat, though it lost it in 1969.

International affairs

The most striking postwar development took place in international affairs. Not only did the U.S. come to supplant the U.K. in New Zealand's thinking about military security but also New Zealand began to play a relatively independent role in world politics. This latter development, in fact, began before World War II, when the Labour government's attitude to the League of Nations was coloured by an idealism that clashed with prewar British policy, especially over the Ethiopian issue. This independent spirit was carried by Fraser to the formation of the United Nations. During the war Fraser had insisted upon an independent voice in the councils of the Allied powers, especially where the fate of New Zealand troops was concerned. At the formation of the United Nations at San Francisco in 1945 he became a notable spokesman for the small powers and made a large impression upon the deliberations of the Trusteeship Council. None of these developments weakened New Zealand's close affinity with Great Britain or its loyalty to the Commonwealth of Nations. Independence and close identity were found to be compatible.

Geography and insecurity shaped New Zealand's postwar foreign policy. With Australia, New Zealand claimed a voice in the settlement of the South Pacific through the Canberra Agreement of 1944. This regional concern also appeared in its role in the South Pacific Commission and in the transfer of authority in Western Samoa, successfully completed when that country became independent in 1962. New Zealand also became deeply involved in Southeast Asia. From 1951, through the Colombo Plan, it provided assistance to many Southeast Asian countries. More militantly, New Zealanders fought in Malaya, Korea, and Vietnam; further, New Zealand became a member of the Southeast Asia Treaty Organization (SEATO) in 1954 and supported U.S. initiatives in that region. This reflected the fear felt at the growth of Communist power in Asia. Thus the independent spirit of the immediate postwar years was modified to a greater dependence on Western powers during the 1950s and 1960s. In the later 1960s the Vietnam war led to a vigorous and continuing public debate on foreign affairs.

Social problems

Two major social problems confronted New Zealand in the second half of the 20th century: to find a productive occupation for a growing population and to preserve harmonious relations between the two races. The country was still an exporter of primary products and an importer of manufactured goods and industrial raw materials. But farming absorbed a decreasing share of the labour force, and traditional farm exports and their markets proved less reliable. To maintain full employment and to strengthen exports, the country began to diversify at home, to explore new products, and to seek new markets. The threat of British entry into the European Economic Community (Common Market) was an incentive to economic reconstruction.

The rapid increase of the Maori population, and its steady urbanization, led governments, especially from the 1950s on, to attempt a "crash program" to raise educational and, eventually, socio-economic levels. These lagged considerably behind the New Zealand average, and this disparity provided ample ground for concern over the possibility of deepening racial tension, in spite of the full legal equality and frequent social harmony that continued to exist. Economic, racial, and foreign policy were the most vital issues in public affairs during the 1960s and early 1970s.

BIBLIOGRAPHY

General works: Of the many short histories, K. SINCLAIR, *A History of New Zealand*, 2nd ed. (1970); and W.H. OLIVER, *The Story of New Zealand*, 2nd ed. (1963), are the most useful.

Early history: Aspects of the culture of the pre-European Maori are explored in two works by PETER H. BUCK: *The Coming of the Maori*, 2nd ed. (1950), and *Vikings of the Sunrise* (1954). The 19th-century compilation of SIR GEORGE GREY, *Polynesian Mythology and Ancient Traditional History of the Maori* (1855; new ed., 1956 and 1970), remains invaluable. European discovery and exploration may be briefly studied in J.C. BEAGLEHOLE, *The Discovery of New Zealand*, 2nd ed. (1961); and W.G. MCCLYMONT, *The Exploration of New Zealand*, 2nd ed. (1959). Early contact between Maori and European is the subject of HARRISON M. WRIGHT, *New Zealand, 1769–1840* (1959). Annexation and early constitutional development are covered by I.M. WARDS in *The Shadow of the Land: A Study of British Policy and Racial Conflict in New Zealand, 1832–1852* (1968). Early settlement and the role of the New Zealand Company are vividly described by J.O. MILLER in *Early Victorian New Zealand: A Study of Racial Tension and Social Attitudes, 1839–1852* (1958).

Development since settlement: The interrelated themes of race relations, political development, and imperial relationships in the period from 1840 to 1870 are explored in J. RUTHERFORD, *Sir George Grey, K.C.B., 1812–1898* (1961); K. SINCLAIR, *The Origins of the Maori Wars* (1957); H.G. MILLER, *Race Conflict in New Zealand, 1814–1865* (1966); J.E. GORST, *The Maori King*, ed. by K. SINCLAIR (1959); B.J. DALTON, *War and Politics in New Zealand, 1855–1870* (1968); and W.P. MORRELL, *The Provincial System in New Zealand, 1852–76*, 2nd ed. (1964). A collection of documents covering this period is G.H. SCHOLEFIELD (ed.), *The Richmond-Atkinson Papers*, 2 vol. (1960). Subsequent political history may be examined through the following books: R.M. BURDON, *The Life and Times of Sir Julius Vogel* (1948) and *King Dick: A Biography of Richard John Seddon* (1955); K. SINCLAIR, *William Pember Reeves* (1965); P.J. O'FARRELL, *Henry Holland, Militant Socialist* (1964); B.M. BROWN, *The Rise of New Zealand Labour: A History of the New Zealand Labour Party from 1916–1940* (1962); R.M. CHAPMAN (ed.), *Ends and Means in New Zealand Politics* (1961); R.S. MILNE, *Political Parties in New Zealand* (1966); K.J. SCOTT, *The New Zealand Constitution* (1962); and R.M. CHAPMAN *et al.*, *New Zealand Politics and Action: The 1960 General Election* (1962). A general economic history with emphasis on the later period of development is M.F.L. PRICHARD, *An Economic History of New Zealand* (1970).

(W.H.O.)

Ney, Michel

Michel Ney was the most famous of Napoleon's marshals, a military hero of legendary courage. His death by

a firing squad in 1815 made him a popular symbol of resistance to the Bourbon restoration as well.

Ney was born January 10, 1769, at Sarrelouis (present-day Saarlouis) in eastern France, the son of a barrel cooper and blacksmith. Apprenticed to a local lawyer, he ran away in 1788 to join a hussar regiment. His opportunity came with the revolutionary wars, in which he fought from the early engagements at Valmy and Jemappes in 1792 to the final battle of the First Republic at Hohenlinden in 1800.

First military service

H. Roger-Viollet

Ney, lithograph by François Le Villain, early 19th century, after a portrait by Maurin.

The early campaigns revealed two contrasting features of Ney's character: his great courage under fire and his strong aversion to promotion. Willing to hurl himself into battle at critical moments to inspire his troops by his personal example, he was unwilling to accept higher rank, and when his name was put forward he protested to his military and political superiors. In every instance he was overruled: it was as general of a division that he fought in Moreau's Army of the Rhine at Hohenlinden.

A year before that battle, Napoleon, under whom Ney had never served, had emerged as master of France. In May 1801, Ney was summoned to be presented to the First Consul at the Tuileries, where Napoleon and Joséphine had surrounded themselves with the ceremony and splendour of a court. The Army of the Rhine had been disbanded, and Ney had bought a modest farm in Lorraine. His first encounter with Bonaparte was formal and unremarkable, for the First Consul, regarding Moreau as a military rival and political opponent, viewed the close associates of that general with suspicion. Joséphine, however, took him up and found him a wife, Aglaé Auguié, one of her maids of honour and daughter of a high civil servant. They were married in the chapel of the Auguié château near Versailles. Ney, with his influential new connections, became, at 33, part of the social and military world of the Consulate.

When peace with England broke down and Bonaparte was assembling armies along the Channel coast, Ney asked for employment and was given command of the VI Army Corps. Early in 1804, when police uncovered a plot by émigré Royalists to kidnap or murder Napoleon and restore the Bourbons to the throne, Ney's republican friend, General Moreau, was said to be involved, and with other alleged conspirators was put on public trial. Napoleon commuted Moreau's two-year sentence for banishment. On May 19, 1804, the day after Napoleon had had himself proclaimed hereditary emperor of the French, he revived the ancient military rank of marshal, and 14 generals, including Ney, were gazetted marshals of the empire.

When Napoleon led his armies in swift marches into the heart of the Continent, after a new European coalition of Russia, Austria, and England had been formed against France, the first victory was won by Ney at Elchingen in October 1805—for which he was created duke of Elchingen in 1808—and less than two months later Napoleon defeated the Russo-Austrian armies at Austerlitz. Ney was active in the defeat of Prussia at Jena in 1806, and of the Russians at Eylau and Friedland in 1807. When he was sent to Spain in 1808, his fame for personal bravery remained undimmed, but at the same time he was also known as a touchy and temperamental commander whom the general staff found difficult to fit into a tactical pattern. His impulsiveness sometimes verged on insubordination when his orders did not come from the Emperor himself. Since Napoleon directed the Spanish operations by remote control, Ney quarrelled with all those set above him, and, early in 1811, he was sent home in near-disgrace.

The Russian campaign of 1812 re-established his position. On the morning after the somewhat inconclusive battle at Borodino, Napoleon made him Prince de la Moskowa. On the retreat from Moscow, Ney was in command of the rear guard, a position in which he was exposed to Russian artillery fire and to numerous Cossack attacks. He rose to heights of courage, resourcefulness, and inspired improvisation that seemed miraculous to the men he led. "He is the bravest of the brave," said Napoleon when Ney, for weeks given up as lost, joined the main body of the frozen and shrunken Grand Army.

In the European campaigns of 1813, Ney had to fight against former friends. Moreau had returned from exile in America to serve as Tsar Alexander I's military adviser and was killed by a French cannonball outside Dresden. Ney had the mortification of being defeated at Dennewitz by the crown prince of Sweden, Charles XIV John, who as Jean Bernadotte had served as a sergeant in the revolutionary armies, as had Ney. At Leipzig, Ney was wounded and had to be sent home. The defeated army fought its way back across Germany into France, where, deaf to all appeals for peace, Napoleon launched a new campaign. Ney, commanding in eastern France, organized the kind of partisan warfare he had learned in the revolutionary wars. Napoleon concentrated his remaining forces at Fontainbleau to fight the allies in Paris, but Ney, speaking for himself and other marshals, told him that the army would not march. "The army will obey *me*," said Napoleon. "Sire," replied the Bravest of the Brave, "the army will obey its generals." Napoleon was forced to abdicate. Ney retained his rank and titles and took an oath of fidelity to the Bourbon dynasty.

Ney's political shift

On March 1, 1815, Napoleon reappeared in France. Ney, ordered to take command in the district of Besançon, told the king "that man deserves to be brought back to Paris in an iron cage." Ney, however, found that the population in his military district was intensely hostile to the Bourbons. Therefore, after receiving messages from Napoleon, he announced his decision to join the Emperor and was deliriously cheered by his soldiers and the populace. The king fled from Paris, and Napoleon re-entered the Tuileries. Ney spent the period mostly in disgruntled retirement at his country estate. He saw little of Napoleon until three days before Waterloo, when he was summoned and asked to serve. He was put in charge of the left wing against the English, Napoleon taking the right wing against the Prussians, whom he defeated at Ligny. Ney fought the English in the drawn battle of Quatre-Bras. His conduct at Waterloo has remained a matter of controversy. When at nightfall the French fled from the field, Ney, his face blackened by smoke and holding a broken sword in his hand, shouted to a colleague: "If they catch us now, they'll hang us," a remark of prophetic accuracy.

After the second return of the Bourbons, Ney made a halfhearted attempt to flee the country, but was recognized and arrested in a remote corner of southwestern France. First put before a court-martial, he refused to recognize its competence and insisted on his right as a peer to be tried by the upper chamber. As he had expect-

Trial and death

ed, he was sentenced to death in one of the most divisive trials in French history. On the morning of December 7, 1815, a firing squad in the Luxembourg Gardens ended what his soldiers had always regarded as a charmed life.

Ney was a soldier's soldier, wholly without political ambition or judgment. He was at his greatest in the campaigns for France's natural frontiers at the beginning and end of his career, but out of his depth in Napoleon's intricate strategy for the domination of Europe. He showed little interest in external distinctions or social success. The dignity with which he met his death effaced the memory of his political vagaries and made him, in an epic age, the most heroic figure of his time.

BIBLIOGRAPHY. ANDREW H. ATTERIDGE, *The Bravest of the Brave* (1912); JOHN B. MORTON, *Marshal Ney* (1958); and HAROLD KURTZ, *The Trial of Marshal Ney, His Last Years and Death* (1957), are the principal biographies, containing full-length bibliographies. Ney's second son published the *Mémoires du maréchal Ney* (1833; Eng. trans., 2 vol., 1833), but this work takes the story only up to 1805. The DUKE OF WELLINGTON'S attitude toward the captured and imprisoned Ney is amply documented in his *Supplementary Dispatches* . . . (1858–72), especially vol. 11; a vivid account of Ney's trial before the Upper Chamber in 1815 occurs in the *Personal Recollections of the Late Duc de Broglie, 1785–1820*, 2 vol. (1887). During his trial a number of English partisans were intent on liberating Ney in an action that misfired. Their exploits are described in GIOVANNI COSTIGAN, *Sir Robert Wilson* (1932); IAN BRUCE, *Lavalette Bruce* (1953); and HAROLD KURTZ (*op. cit.*).

(Ha.K.)

Niagara River and Falls

Flowing in a northerly direction from Lake Erie to Lake Ontario, a distance of about 35 miles (56 kilometres), the Niagara River constitutes part of the boundary between the United States and Canada, separating New York state from the province of Ontario. It is the drainage outlet for the four upper Great Lakes (Superior, Michigan, Huron, and Erie), the aggregate basin area of which is some 260,000 square miles (673,000 square kilometres). The relatively high flow and steep descent (about 326 feet [99 metres]) of the river combine to make it one of the best sources of hydroelectric power in North America.

About halfway along the river's course lies Niagara Falls, one of the major scenic spectacles of the continent. Although at 167 feet (51 metres) it is far from the highest waterfall in the world, its breadth and quantity of flow give the falls a grandeur and beauty attested to by over 4,000,000 visitors annually. United States and Canadian cities of the same name stand on either bank of the river. For many decades the falls were an attraction for honeymooners and for such stunts as walking over them on a tightrope or going over them in a barrel, but increasingly the appeal of the site has become its beauty and uniqueness as a physical phenomenon.

Course of the river. From its head at Lake Erie, the river flows through a single channel for about five miles. It is then divided into two channels by Strawberry and Grand islands, the eastern, or U.S., channel running for about 15 miles, the western, or Canadian, for about 12. At the foot of Grand Island the two merge again about three miles above the falls. From Lake Erie to the upper rapids the river descends about 10 feet, whereas in the short rapids it falls 50 feet before pouring over the falls.

Divisions of the falls

The falls are in two principal parts, separated by Goat Island. The larger division, adjoining the left or Canadian bank, is Horseshoe Falls; its height is 162 feet and the length of its curving crest line is about 2,600 feet. The American Falls, adjoining the right bank, is 167 feet high and 1,000 feet across.

Below the falls and extending for seven miles is the Niagara Gorge. The stretch of two and one-fourth miles from Horseshoe Falls is known as the Maid of the Mist Pool. It has a descent of only five feet and is navigable by excursion boats. Beyond this, the gorge descends another 93 feet, flowing northwestward first through the narrow Whirlpool Rapids for one mile to the Whirlpool. There the gorge makes a 90° bend to the northeast for two miles and turns north for another one and one-half miles to the foot of the Niagara Escarpment at Lewiston, New York. In its final seven miles, the river flows across a lake plain to Lake Ontario, falling less than one foot.

Geological history. The shaping of the gorge and the maintenance of the falls as a cataract depend upon peculiar geologic conditions. The rock strata from the Silurian Period (between 430,000,000 and 395,000,000 years ago) in the Niagara Gorge are nearly horizontal, dipping southward only about 20 feet to the mile. The uppermost layer of hard Niagaran dolomite is underlain by soft layers that are easily worn away. This disposition of the rock strata provides the conditions for keeping the water constantly falling vertically from an overhanging ledge during a long period of recession (movement upstream).

The river came into existence late in the Pleistocene, or Glacial, Epoch, more than 10,000 years ago, when the margin of a great continental ice sheet melted back and exposed the escarpment of Niagaran dolomite rock, allowing the discharge from the Lake Erie Basin to pour over it. Recession of the falls created the Niagara Gorge, the age of which, when calculated by dividing its length by the average rate of recession of the falls in recent time, is about 7,000 years. Other considerations led some geologists to estimate an age as great as 25,000 years. Determinations of the age of the last glacial ice advance in the area suggest, however, that the Niagara River is about 12,000 years old. The Whirlpool section differs from the rest of the gorge because there the river intersects an old channel that was formed before the last glacial ice advance and was later filled with glacial drift. Continued recession of the falls toward Lake Erie will ultimately cause the drainage of that lake, but such an event is not expected to occur within the next 25,000 years.

Age of the river and gorge

Human uses. *Navigation.* The river is navigable from Lake Erie to the upper rapids. Waterborne traffic along the Niagara passes through the upper single channel and the U.S. channel and enters the New York State Barge Canal at Tonawanda, New York. That canal, with minimum depth of 12 feet, connects with the Hudson River and has branches that connect with Lake Champlain and Lake Ontario. In winter Lake Erie ice floats into the Niagara River and tends to form a jam that restricts the flow. To prevent this, a floating timber boom is installed across the entrance, generally from early December to April. This obstruction may be bypassed by ships taking a route through the Black Rock Canal, from Buffalo Harbor to a point a few miles down the Niagara River, which extends the navigation period locally through a greater part of the winter. The principal shipping between Lakes Erie and Ontario, however, passes through the Welland Canal, an important link in the Great Lakes–St. Lawrence Seaway. Lying a few miles west of the Niagara River inside Ontario, it has a minimum depth of 27 feet. The ships that pass through it include vessels engaged in trade between the upper Great Lakes and Europe.

Generation of hydroelectricity. Canada and the U.S. agreed, in a treaty signed in 1950, to reserve sufficient amounts of water for flow over Niagara Falls to preserve their scenic value. The agreement provided for a minimum daytime flow during the tourist season of 100,000 cubic feet (3,000 cubic metres) per second and a minimum flow of 50,000 at all other times. All water in excess of these amounts, estimated to average about 130,000 cubic feet per second, was made available for diversion for power generation, to be divided equally between the U.S. and Canada. The total hydroelectric capacity of the river was fixed at about 4,000,000 kilowatts. This power is developed by public-authority power-plant installations on both sides of the gorge: the Robert Moses plant near Lewiston, New York, and the Sir Adam Beck plant at Queenston, Ontario, completed in the period 1960–63. The plants receive water diverted to them from the river above the falls and carried to them by tunnels and canals. The flow of the Niagara River averages 195,000 cubic feet per second but fluctuates from a minimum of 119,000 in winter to a maximum of 245,000 in summer. The amounts of water available for diversion range from about 110,000 cubic feet per second in winter to 120,000

on summer days and to 170,000 on summer nights. At night, when industrial power demand is low, turbine pumps deliver a part of the water to large storage basins about both plants; during daytime peak demand, the stored water is flowed back through the turbines to generate more electricity. Much of the energy is used in nearby electrochemical industries. The remainder of the energy is transmitted to various cities for miscellaneous uses. The maximum distance to which this power is transmitted is somewhat in excess of 350 miles.

The falls and other features. The water flowing over the falls is free of sediment, and its clearness contributes to the beauty of the cataract. In recognition of the importance of the waterfall as a great natural spectacle, the province of Ontario and the state of New York retained or acquired title to the adjacent lands and converted them into public parks.

In recent years the very large diversion of water above the falls for power purposes has lessened the rate of erosion. Elaborate control works upstream from the falls have maintained an even distribution of flow across both the U.S. and Canadian cataracts, and the curtains of the waterfalls have been maintained. Sightseers have found it difficult to believe that a large part of the great river above the falls is diverted and disappears into four great tunnels for use in the power plants downstream. Serious erosion of the American Falls has been occurring, however, and in 1969 the river was diverted from that area and some cementing of the bedrock was done, while at the same time an extensive boring and sampling program was carried out. River flow was returned to the American Falls in November 1969, and the effects of the work and results of the study continue to be evaluated.

Scenic sites

Excellent views of the falls are obtained from Queen Victoria Park on the Canadian side; from Prospect Point of the U.S. side at the edge of the American Falls; and from Rainbow Bridge, which spans the gorge about 1,000 feet downstream from Prospect Point. Visitors may cross from the U.S. shore to Goat Island by footbridge and may take an elevator to the foot of the falls and visit the Cave of the Winds behind the curtain of falling water. The Horseshoe Falls receded, or migrated upstream, at an average rate of nearly five feet per year in historic time, until control works and diversion of water decreased the erosion.

Another feature of the river is Ft. Niagara, a 504-acre (204-hectare) state park on the east bank at the mouth of the river. Located on the site of a blockhouse (Fort-Conti) built by the French in 1678–79, it includes old Ft. Niagara, which was built by the French in 1725–27 and is still standing. The fort was captured by the British in 1759 and was relinquished to the United States in 1796. Restoration of this and other structures in the park has provided an authentic historical flavour representing both the French and British periods of occupancy.

BIBLIOGRAPHY. UNITED STATES DEPARTMENT OF COMMERCE, LAKE SURVEY CENTER, *Great Lakes Pilot* (annual), full descriptions of physical conditions and the particulars of constantly changing conditions for the entire Great Lakes waterway system, including the Niagara River and vicinity; LLOYD GRAHAM, *Niagara Country* (1949), a popular-style history of the area from the days of Indian habitation to the period of cooperative development by Canada and the United States; WALTER M. EDWARDS, "Niagara Falls, Servant of Good Neighbors," *Natn. Geogr. Mag.*, 123:574–587 (1963), coverage of the area in well-captioned colour photographs.

(J.L.Ho.)

Nicaragua

The Republic of Nicaragua (República de Nicaragua) is located in the middle of Central America and is bounded by Honduras on the north, the Caribbean Sea on the east, Costa Rica on the south, and the Pacific Ocean on the west. Despite territorial losses to its neighbours, it is the largest of the Central American republics, with an area of 49,759 square miles (128,875 square kilometres). Its population of more than 1,900,000 represents a racial mixture that reflects the country's history. The national capital is at Managua.

Nicaragua can be characterized by its agricultural economy, its history of autocratic government, and its imbalance of regional development. Most of the population is engaged in subsistence farming, and the national economy is perhaps overdependent upon exports of cotton and coffee. Although several political parties are active, the government has been in the powerful hands of the family of former president Anastasio Somoza García since the 1930s. Almost all settlement and economic activities are concentrated in the west, and the rich tropical forests of the east remain unexploited. (For a related physical feature see NICARAGUA, LAKE. For a detailed discussion of history see CENTRAL AMERICAN STATES, HISTORY OF.)

The landscape. *Relief.* The western half of the country is occupied by a triangular region of mountains. One side of the triangle extends from northeast to southwest along the Honduras border, the second side runs from Honduras in the northwest to the Costa Rican border in the southeast, and the third side extends from north to south through the centre of Nicaragua. The short, steep, and rugged mountain ranges are separated by basins and contain fertile valleys. The intricately dissected region includes the Cordillera ("chain of mountains") Entre Ríos, on the Honduras border, the Cordilleras Isabella and Darién, in the north central area, and the serranías ("mountain range") Huapí, de Amerigue, and de Yolaina in the southeast. The mountains are highest in the north, where Pico Mogotón, in the Cordillera Entre Ríos, and Cerro Saslaya Peak, in the Cordillera Isabella, rise to 6,913 feet and 6,542 feet (2,107 metres and 1,994 metres), respectively.

The volcanic zone

To the west of the central mountains is a string of about 40 volcanoes that stretches from northwest to southeast along the Pacific coast. They are surrounded by low plains from the Golfo de Fonseca in the north to the Bahía (Bay) de Salinas in the south and are separated from the mountains by the great basin that contains Lakes Nicaragua, Managua, and Masaya. The volcanoes, some of which are active, are divided into two groups; in the north they are known as the Cordillera de los Marrabios and in the south as the Mesetas de los Pueblos. The highest volcanoes include San Cristóbal, which rises to 5,725 feet (1,745 metres); Concepción, which attains 5,282 feet (1,610 metres); and Momotombo, at 4,199 feet (1,280 metres).

The eastern half of Nicaragua is occupied by low, level plains. Among the widest Caribbean lowlands in Central America, they exceed 50 miles (80 kilometres) in width. The coast is broken by deltas and sandbars that enclose lagoons, and coralline reefs, islands, and banks that lie off shore.

Drainage and soils. The central mountains form the country's main watershed. The rivers that flow to the west empty into the Pacific Ocean or Lakes Managua and Nicaragua. They are short and carry a small volume of water; the most important are the Río Negro and Estero Real, which empty into the Golfo de Fonseca, and the Río el Tamarindo, which flows into the Pacific.

The eastern rivers are of greater length. The Río Coco (Río Segovia) flows for 295 miles along the Nicaragua–Honduras border with a total length of 425 miles and empties into the Laguna de Gracias a Diós on the extreme northern coast. The Río Grande de Matagalpa flows for more than 200 miles from the Cordillera Darién eastward across the lowlands to empty into the Caribbean north of Laguna de Perlas (Lagoon of Pearls) on the central coast. In the extreme south, the Río San Juan flows for 120 miles from Lake Nicaragua into the Caribbean in the northern corner of Costa Rica. Other rivers of the Caribbean watershed include the 120-mile-long Río Prinzapolca, the 60-mile-long Río Escondido, the 60-mile-long Río Indio, and the 37-mile-long Río Maíz.

Nicaragua's lakes

The west is a region of lakes. Lake Nicaragua, with an area of 3,191 square miles (8,264 square kilometres), is the largest lake in Central America. Located in the southern isthmus, the lake and its distributary, the Río San Juan, have long been discussed as a possible canal route between the Caribbean and the Pacific.

There are six freshwater lakes near the capital city of

Managua. They include the Lago de Managua, which covers an area of 405 square miles (1,049 square kilometres); Laguna de Asososca, which is 6,000 feet deep and acts as the city's reservoir of drinking water; and Laguna de Jiloá, which is slightly alkaline and is a favourite bathing resort. Laguna de Masaya is prized for its swimming and fishing facilities; the sulfurous waters of Laguna Nejapa have medicinal properties ascribed to them; and Laguna Tiscapa, which lies within the capital city, is the site of the Presidential Palace.

Other lakes in the Pacific watershed include Laguna de Apoyo, near Laguna de Masaya; Laguna Apoyeque, picturesquely located between two peaks on the Punta Chiltepe that juts into Lago de Managua; and the artificial Lago de Apanás on the Río Tuma, which generates much of the electricity consumed in the Pacific zone.

Soils on the east (Caribbean) coast are alluvial and are considered fertile. On the Pacific coast the soil is volcanic, and about 85 percent of its area is fertile.

Climate. The climate is slightly cooler and much wetter in the east than in the west. The Pacific side is characterized by a rainy season from May to November and a dry season from December to April. The annual average temperature is 81° F (27° C), and precipitation averages 75 inches (1,910 millimetres) yearly. On the Caribbean side of the country, the rainy season lasts for about nine months of the year, and there is no well-defined dry season. The annual average temperature is 79° F (26° C), and annual precipitation may total about 150 inches (3,810 millimetres). In the northern mountains temperatures are cooler and average about 64° F (18° C). Prevailing winds are from the northeast and are cool on the high plateau, warm and humid in the lowland.

Vegetation and animal life. Over 50 percent of the country is covered with a broad expanse of tropical forest. The numerous types of trees include the almond, walnut, broad-leaved male cedar, royal cedar, balsam, coast and okote pines, mahogany, and liquid amber. The quebracho, or axbreaker, yields hardwood; the guaiacum, or lignum vitae, yields one of the iron woods; the guapinal yields resin; and the medlar produces fruit similar to the crab apple. The western lowlands are comprised of savannas, the streams of which are lined with forests.

The hot and rainy regions are inhabited by various types of reptiles, including crocodiles, lizards, snakes, and turtles. Deer are common in the forested regions; wildlife also includes such cats as the puma and the jaguar, monkeys, and the peccary (a nocturnal mammal resembling the pig). There are many species of water and land birds, and the rivers, lakes, and lagoons are inhabited by freshwater and saltwater fishes and mollusks. Rodents and insects are abundant.

Traditional regions. The western volcanic mountains and surrounding lowlands and lakes contain more than 60 percent of the country's population, most of its cities, and most of its industry. The area also yields most of Nicaragua's agricultural produce and minerals. The valleys of the western central mountains contain about 30 percent of the population and yield about 25 percent of the national agricultural production. The highlands are also important for mineral extraction. Although they comprise about one-half of the country, the western lowlands are virtually uninhabited except along the coast and rivers. The region is important, however, as a vast resource of timber.

The four traditional regions

Patterns of settlement. In the early 1970s, about 52 percent of the population was rural. By far the largest city was Managua, on the southeastern shore of Lago de Managua. On December 23, 1972, the city was devastated by a series of earthquakes and subsequent fires that left approximately 7,000 dead, 15,000 injured, and half the city's enumerated population of 398,500 homeless. Other important urban centres included León (population, 55,600), Granada (35,000), Masaya (30,800), Chinandega (30,400), and Chichigalpa (15,000), all in the west. Matalgalpa, with a population of 21,000, and Estelí, with 20,200 inhabitants, were the largest cities of the central mountains. Bluefields, with 15,000 inhabitants, was the largest town on the Caribbean coast. All other urban centres had populations of fewer than 14,000.

People and population. *Population groups.* Most of Nicaragua's population of 1,911,543 is a product of racial admixture. Almost 70 percent of the people are mestizos (persons descended from mixed European and Amerindian ancestry), zambos (persons of mixed black and Amerindian descent), and mulattoes (persons of mixed European and black descent). About 20 percent are white and 10 percent are black. Of the original Amerindian population, only the Sumo, Miskito, and Ramaquie tribes remain; they occupy the basins of the northeastern rivers and other Atlantic areas.

Spanish is the official national language, and there is a widespread familiarity with English. The Amerindian languages have almost disappeared from use, but their influence remains in place names and in many nouns in Nicaraguan Spanish.

Roughly 85 percent of the population is Roman Catholic. The small Protestant and Jewish communities are concentrated in the larger cities. There is no official religion, and freedom of religious worship is guaranteed by law.

Demography. In the early 1970s the average population density was more than 41 persons per square mile. Because of the high birth rate of about 43 births per thousand of the population and the death rate of 8 per thousand, the population is young. In the early 1970s, 49 percent were under 15 years of age, and almost 47 percent were between the ages of 15 and 60. The natural rate of increase was about 35 per thousand of the population.

Nicaragua, Area and Population

	area*		population	
	sq mi	sq km	1963 census	1971 census†
Zones				
Atlantic				
Departments				
Río San Juan	2,876	7,448	16,000	21,000
Zelaya	22,816	59,094	89,000	149,000
North Central				
Departments				
Boaco	1,924	4,982	72,000	69,000
Chontales	1,910	4,947	76,000	69,000
Estelí	849	2,199	69,000	79,000
Jinotega	3,697	9,576	77,000	93,000
Madriz	679	1,758	50,000	54,000
Matagalpa	2,623	6,794	171,000	174,000
Nueva Segovia	1,290	3,341	46,000	66,000
Pacific				
Departments				
Carazo	398	1,032	66,000	71,000
Chinandega	1,800	4,662	129,000	155,000
Granada	372	964	66,000	72,000
León	2,021	5,234	150,000	168,000
Managua†	1,403	3,635	319,000	504,000
Masaya	210	543	77,000	94,000
Rivas	830	2,149	64,000	74,000
National district				
Distrito Nacional‡	...	...	...	...
Total Nicaragua	45,698 49,759	118,358 128,875	1,536,000§	1,912,000

*Areas for subdivisions are land areas. Of the two country total area figures, the first is the land area, the second the total area. †Preliminary figures. ‡All area and population figures for the Department of Managua include the National District. §Figures do not add to total given because of rounding.
Source: Official government figures.

The national economy. Nicaragua is rich in natural resources, most of which have yet to be exploited on a large scale. The economy is basically agricultural, and industry is in an incipient stage of development.

Natural resources. Mineral resources include known deposits of gold, silver, copper, iron ore, bauxite, lead, marble, and gypsum. The forests are a vast resource of hardwoods and softwoods, and the inland and coastal waters contain abundant food fishes. Nicaragua's potential hydroelectric resources are considered immense.

Sources of national income. Agriculture, forestry, and fishing engage about 46 percent of the labour force and

produce about 30 percent of the total national income. The major cash crops are cotton and coffee, most of which are produced for export. Cotton alone represents almost 30 percent of total exports. Other crops include maize (corn), rice, beans, sugarcane, sesame, henequen (sisal; a plant that yields strong fibres), manioc (yucca and cassava), and wheat. A variety of fruits and vegetables are also produced for the local market. Soybeans and peanuts have been introduced, and in the early 1970s the possibilities of cultivating the sunflower were being studied. There is room for great expansion of agriculture, for only 30 percent of the arable land is exploited.

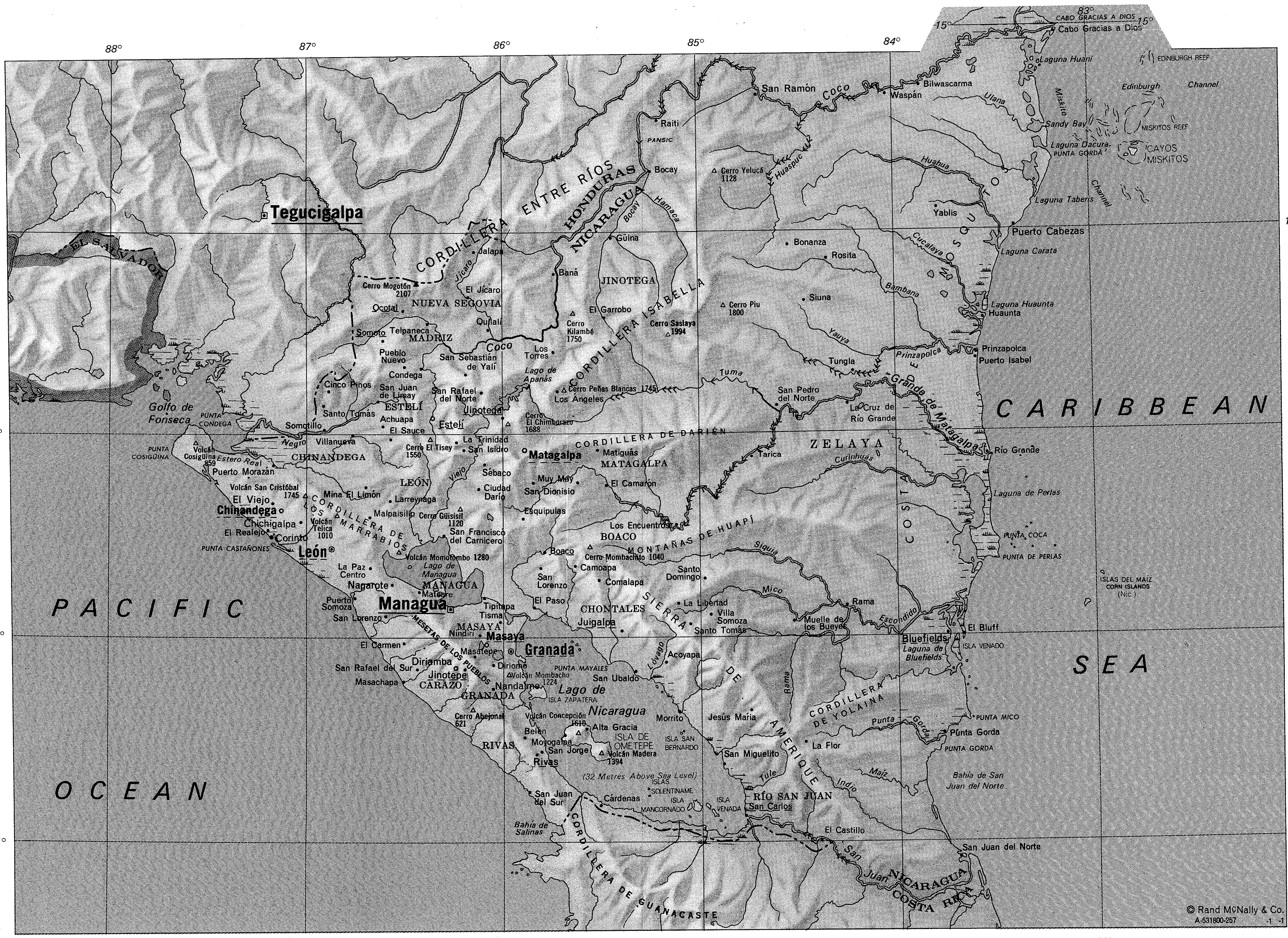
Cattle are significant as a source of dairy produce in the west and beef in the east. The cattle industry was increasing rapidly during the early 1970s, and exports of meat were reaching a competitive level with those of coffee. Other livestock includes goats, hogs, horses, and sheep.

Forestry is little developed because of the lack of adequate transportation facilities. Shrimping is the most important marine activity. About 95 percent of the shrimp, caught in both the Pacific and Caribbean, are exported; lobster is also exported in moderate quantities. Nicaragua's fish resources, however, are also little exploited because of lack of investment, and marine fishing remains largely a subsistence activity.

Of all the country's minerals, only its gold and copper are mined intensively. Both are mined for export; gold is extracted by Canadian and United States firms. Silver is also obtained on a smaller scale. Other reserves are as yet unexploited because of lack of financing.

The industrial sector

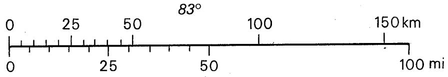
Nicaragua's infant industry is based on the production of consumer products, much of which require the importation of raw materials. In the early 1970s the government actively supported the diversification of production and the use of domestic raw materials. Products include refined petroleum, matches, footwear, soap and vegetable oils, cement, alcoholic beverages, and textiles. Upon the formation of the Central American Common Market in



NICARAGUA

Size of symbol indicates relative size of town • ○ ⊙ ⊡

Elevations in metres



1960, it was hoped there would be an increased demand for Nicaragua's processed foods, lumber, and fibres.

In the early 1970s the country had a total installed electrical capacity of more than 170,000 kilowatts at about 70 diesel and 11 hydroelectric stations. Construction continued on the Río Tuma scheme, which began to supply hydroelectric power to the western region in 1965. Research is being conducted on the applications of atomic power to engineering, medicine, and technology.

The Banco Central de Nicaragua, established in 1961, has the sole right of issue of the national currency, the cordoba. The financial system is dominated by the government-owned Banco Nacional de Nicaragua, which acts as a development bank and grants loans for agriculture and cattle raising. The Instituto de Fomento Nacional (Infonac; Institute for National Development) extends partial credit for industry, agriculture, and cattle raising. Long-term loans for housing construction are granted by the Banco de la Vivienda (Housing Bank), which covers less than 10 percent of the nation's needs. There is also an urgent need for financial institutions that can extend long-term credit for economic development.

Branches of U.S., British, and Canadian commercial banks operate in Nicaragua. There are also two private commercial banks and a savings and loan company.

Nicaragua's chief trading partners are the U.S., Japan, and West Germany. Imports are comprised largely of machinery, transport equipment, medicines, nonferrous metals, and petroleum. Exports include cotton, coffee, meat, oilseeds, copper, sugar, shrimp, and lobsters.

Management of the economy. Most government revenue is obtained through indirect taxes, such as import duties and surcharges, export taxes on coffee, and sales taxes on liquor, cigarettes, and other consumer goods. Direct taxes are levied only on property (land and houses).

Trade unions

Trade unions are small, and none of the larger organizations has more than 5,000 members. Organizations include the Confederación Nacional de Trabajadores de Nicaragua (National Confederation of Workers of Nicaragua), the Confederación General del Trabajo (General Confederation of Labour), the Federación de Transportadores Unidos Nicaragüense (United Transport Workers' Federation of Nicaragua), the Federación Sindical de Maestros de Nicaragua (Nicaraguan Teachers' Trade Union Federation), and the Movimiento Sindical Autónomo de Nicaragua (Autonomous Trade Union Movement). The independent union movement is growing, but most unions are still government controlled. Collective bargaining is used in settling some disputes.

Employers' associations include the Cámara Nacional de Comercio de Managua (National Chamber of Commerce of Managua). The Corporación Nicaragüense de Inversiones channels national and foreign financial resources into industrial development, and the Instituto Nacional de Comercio Exterior e Interior regulates trade balances and prices. The Comisión Nacional del Algodón is the official government agency for cotton development, and the Instituto Nicaragüense del Café is an autonomous government agency that controls the quality and export of coffee.

In the early 1970s the government was concerned with the unfavourable balance of trade and the nation's dependence upon cotton and coffee production. Imports were reduced, and the government continued modernizing and diversifying agriculture. Industries that used local raw materials continued to receive priority.

Transportation. Most of the country's transportation system is confined to the western zone. There is a network of highways estimated to be 8,740 miles (14,070 kilometres), of which more than half are impassable during the rainy season, 780 miles are paved, and about 750 miles are gravel covered. The system includes the 255-mile (410-kilometre) Nicaraguan section of the Pan-American Highway, which runs through the west from Honduras to Costa Rica. There is also an important road that runs from the Pan-American Highway, 24 miles from Managua, to Port Esperanza at Ciudad Rama, on the Río Escondido.

There are 235 miles (378 kilometres) of railways. The main line runs from Granada northwest to Corinto, on the Pacific Ocean. Branch lines lead north from Corinto to Puerto Morazán on the Golfo de Fonseca and north from León to the coffee area of Carazo.

The chief ocean ports of Puerto Morazán, Corinto, and San Juan del Sur serve the Pacific coastal area. The Caribbean ports include Puerto Cabezas, Port Isabel, and Bluefields, which is connected to the river port of Port Esperanza by regular small craft service. The short rivers in the west are navigable for canoes and small motorboats; the lower course of the Estero Real, however, can accommodate small commercial vessels. In the east, the Río Coco is navigable in its lower course for medium-size vessels. The Río Grande de Matagalpa is limited to small craft, as is the Río San Juan.

Air transport services

The international De Las Mercedes Airport, seven miles from Managua, offers daily jet service to North America and Latin America. The second largest airport at Puerto Cabezas can also accommodate jets; there are five other airports with scheduled domestic flights. Air services are offered by Lanica, a private airline, and by American and Central American airlines. Craft Airlines offers daily service between Managua and San José, Costa Rica.

Administration and social conditions. *Constitutional framework.* Nicaragua had nine constitutions between 1838 and 1972. Under the 1950 constitution, which remained in effect until 1972, the national government was headed by a strong president, who was elected to a one-time, five-year term. Given extensive executive powers, he could detain persons, suspend constitutional guarantees, and enact laws when the legislature was not in session. The president appointed his Cabinet of ministers, who could not be members of the legislature.

Nicaragua had the only bicameral legislature in Central America. It consisted of a 42-member Chamber of Deputies and a Senate of 16 elected members, all serving for six years, and all former presidents, who were appointed for life. The legislature acted as a rubber stamp of executive authority; it could amend the constitution with a majority vote.

Under an agreement reached in 1971 between Nicaragua's two major political parties, the president was succeeded on May 1, 1972, by a triumvirate of party leaders. The legislature was replaced with a 100-member Constituent Assembly that had been elected in 1971 and charged with the task of producing a new constitution by 1974. The triumvirate was to retain executive authority until a president was elected under the new constitution.

Local government. In 1971 the country was divided into 16 departments and the national district of Managua. Each was administered by a political head appointed by the president. The departments were subdivided into 123 *municipios*, each headed by a mayor, or alcalde. Except in Managua, mayors are elected for two years by a council; the council itself is elected for a four-year term. In Managua mayoral duties are performed by a presidential appointee.

The political process. The franchise extends to all persons over 21 years of age and all over 18 years who have earned an elementary school-leavers certificate, who are literate, or who are married. Voting is carried out by secret ballot.

Under the 1950 constitution, elections were controlled by the Supreme Electoral Tribunal, which ranked as the fourth branch of the government; it counted votes, verified the election, and determined the winners. Deputies were elected from a national list and senators from four electoral districts. Seats in the legislature were awarded by a system of proportional representation in which at least one-third were reserved for minority political parties. In 1971 seats in the Constituent Assembly were proportionately divided between the two major parties, as was the membership of the tribunal one year later.

Nicaragua's leading political organization is the Nationalist Liberal Party. The Nicaraguan Conservative Party is the second strongest party, although it is not con-

sidered by some to represent true political opposition to the government. The Traditionalist Conservatives and the Independent Liberals are clearly anti-government; they generally do not participate in elections, however, because of a lack of confidence in the electoral tribunal.

Justice and the armed forces. The judicial system is headed by the Supreme Court in Managua, which may initiate legislation. There are five courts of second instance in León, Granada, Masaya, Matagalpa, and Bluefields; more than 150 courts of first instance; and a labour court. Under the 1950 constitution, judges were appointed by the legislature. The supreme court magistrates are seven, of which four must belong to the majority party; all are appointed by Congress.

The National Guard

The National Guard serves as both the army and police force. Originally trained by the U.S. Marines, it is composed of about 5,400 active and 4,000 reserve personnel. Members enlist for a three-year period, but service can be made compulsory at any time. The Guard is politically important and has been the power behind the government since the 1930s. There is also a small coast guard and a well-equipped air force of about 1,500 men.

Education, health, and welfare. Primary education is free and compulsory for those between the ages of six and 13. The nation's facilities and teaching staff are inadequate, however, and can accommodate only about 50 percent of the school-age children. Of those who attend school, most leave before completing the secondary level, and over 40 percent of the population is illiterate.

The universities

Institutions of higher learning include the Nicaraguan branch of the Central American University at Managua, founded in 1961; and the National Autonomous University of Nicaragua at León, founded as the University of León in 1814. The Central American Institute of Business Administration offers post-graduate courses under the guidance of Harvard University.

Health and welfare services are inadequate. In the late 1960s there was one doctor for every 1,674 persons and one hospital bed for every 433. Welfare services are offered by the labour unions and the limited social security program is offered for inhabitants of Managua. The shortage of adequate housing in both urban and rural areas remains acute, despite the government's housing program that began in 1959.

Social conditions. In response to the nation's overall poverty, the government initiated minimum wages in 1963; agricultural workers receive food allowances and housing loans. These programs have had little effect, and the average annual income—including that of the multimillionaire families that dominate the economy—was US $380 in 1969.

Malaria, which was previously responsible for 30 percent of the nation's annual deaths, was eradicated by the early 1970s. Major causes of death are old age, infections of the digestive system, homicide, and war.

Cultural life. Nicaragua's literary tradition was consolidated in the late 19th and early 20th centuries with the poetry and prose of Rubén Darío, known as the "prince of Spanish-American literature." Later writers of importance include Santiago Argüello, Lino Argüello, Salomón de la Selva, and Salvador Buitrago Díaz.

José de la Cruz Mena is to Nicaraguan music what Darío is to its literature. Although he has not received international recognition, his waltzes are masterpieces of melody and harmony. The composers Edwin Krüger and José Ramirez excel in folkloric music. There are also many creative artists and sculptors.

The National Library at Managua maintains branches in most of the major cities. The National Museum is located in the capital city, and the Tenderi Museum of Indian artifacts is at Masaya. The Grand Theatre "Ruben Dario" in Managua rivals the best in Latin America.

The most important Nicaraguan daily newspapers are *La Prensa* and *Novedades* in Managua, *El Centroamericano* in León, and *El Diario* in Granada. There are also two popular weekly papers: *Semana* (owned by *La Prensa*) and *Extra.*

The government-owned National Radio operates 72 broadcasting stations throughout the country, and there is one radio receiver for every 18 persons. Three television stations broadcast from Managua.

Prospects for the future. Like many Latin American nations, Nicaragua faces problems of economic and social imbalance. The rising birth rate, combined with an uneven distribution of settlement on the land, and the concentration of economic control in the hands of relatively few, while the majority of the population lives at the subsistence level, all constitute direct or indirect obstacles to rapid development. A further factor is of particular importance—the country is dependent upon its agricultural exports, such as cotton, and national income is thereby subject to fluctuations in the world market. Tourism, however, has provided needed revenue, and prospects for its further expansion are bright. Hopes for the construction of a trans-Nicaraguan canal to supplement the Panama Canal have yet to materialize, although should they ever do so, the country's economic and strategic position would be transformed. A devastating earthquake hit Managua in December 1972; it compounded Nicaragua's economic difficulties and created serious problems of reconstruction and resettlement. It was planned to rebuild Managua's business district about six miles from its original site.

BIBLIOGRAPHY. Three good works on the geography of the country are PABLO LEVY, *Nicaragua* (1864); LOS HERMANOS CRISTIANOS, *Geografía de Nicaragua* (1930); and JAIME INCER BARQUERO, *Nueva Geografía de Nicaragua* (1970). In English, the accounts of the PAN-AMERICAN UNION, *Nicaragua* (1964) and *Visit Nicaragua* (1957), are also useful. For a description of the flora and fauna of the area, see THOMAS BELT, *The Naturalist in Nicaragua* (1874 and 1911).

(R.A.)

Nicaragua, Lake

Lake Nicaragua (Spanish, Lago de Nicaragua), the largest of several freshwater lakes in southwestern Nicaragua and the dominant physical feature of the country, is also the largest lake in Central America. Its aboriginal name was Cocibolca, meaning "sweet sea"; the Spanish called it Mar Dulce (freshwater sea); its present name is said to have been derived from that of Nicarao, an Indian chief whose people lived on the lake shores. Oval in shape, the lake has an area of 3,190 square miles, is 110 miles in length, and has an average width of 36 miles. About 60 feet deep in the centre, its waters reach a depth of 200 feet to the southeast of its largest island, Ometepe (Spanish, Isla de Ometepe). Its surface is 95 feet above sea level.

It is believed that Lake Nicaragua, together with Lake Managua (Lago de Managua) to the northwest, originally formed part of an ocean bay which, as a result of volcanic eruption, became an inland basin containing the two lakes, which are linked by the Tipitapa River (Río Tipitapa). The ocean fish thus trapped adapted themselves as the water gradually turned from saltwater to freshwater. Lake Nicaragua is the only freshwater lake containing oceanic animal life, including sharks, swordfish, and tarpon.

More than 40 rivers drain into the lake, the largest being the Tipitapa River. The San Juan River (Río San Juan) drains out of the lake, following a 112-mile (180-kilometre) course that runs from the southeastern shore of the lake through a densely forested region to empty into the Caribbean Sea. For part of its course, the San Juan forms the boundary between Nicaragua and Costa Rica. To the southwest, the lake is separated from the Pacific Ocean by a narrow land corridor—the Rivas Isthmus—which is 12 miles (19 kilometres) wide.

Contrary to popular belief, the lake is tideless, although there is a daily fluctuation in the water level caused by east winds blowing up the San Juan Valley. The water level also falls during the dry season, December to April, and rises during the rainy season, May to October. There are several currents in the lake; the principal one runs from southeast to northeast on the surface, while beneath it a deeper current flows in the opposite direction. Surface water temperature usually remains at 75° F (24° C), and

bottom temperature at 60° F (16° C). Due to the chemical composition of the volcanic rocks forming parts of the lake's bed and shores, the lake waters contain high proportions of dissolved magnesium and potassium salts.

There are over 400 islands in the lake, of which 300 are within five miles (eight kilometres) of the city of Granada on the northwest lake shore. Most of the islands are covered with a rich growth of vegetation, which includes tropical fruit trees. Some of the islands are inhabited. Ometepe, which, as mentioned, is the largest island, is 16 miles (26 kilometres) long and 8 miles (13 kilometres) wide. It is formed of what originally were two separate volcanoes—Concepción, which is 5,282 feet (1,610 metres) high and which last erupted in 1944, and Madera, which is 4,015 feet (1,224 metres) high. Lava from bygone eruptions forms a bridge between them, called the Tistian Isthmus. (A third volcano associated with the lake is Mombacho, 4,413 feet [1,345 metres] high, which stands on the western shore.) Ometepe Island is the pre-eminent site in Nicaragua for pre-Columbian examples of statuary, ceramics, and other archaeological remains, some of which, it is believed, represent vestiges of ancient South American, as well as North American, civilizations. Coffee, cocoa, corn, bananas, and fruit are grown on the island as well as a little cotton and tobacco.

The lake is a traditional means of communication between the cities of the west coast of Nicaragua and those in the south and east of the country. Steamships have operated on the lake since 1882. Lake steamers based at Granada visit small lakeside towns, such as San Jorge and La Virgen on the west shore; Cárdenas and San Carlos on the southern shore; San Ubaldo, Puerto Díaz, San Miguelito, and El Morrito on the eastern shore. Granada has a population of about 35,000; the next largest lakeside town is San Jorge, with a population of about 2,000.

In previous centuries, piratical raids from the Caribbean were sometimes made on the lakeside towns, until the building of fortifications in the 17th century on the San Juan River blocked the pirates' ingress. From the time of the ending of Spanish rule in the 1820s, the possibility of constructing a transoceanic canal from the Atlantic to the Pacific, which would run up the San Juan River, cross the lake, and be completed by a channel dug through the Rivas Isthmus, has been mooted. Surveys to this end were made in the 1830s. After the discovery of gold in California in 1848, Cornelius Vanderbilt, the New York millionaire, developed the Vanderbilt Road—a route over which gold prospectors from New York were transported up the river and over the lake, completing the final few miles to the Pacific by stage coach in order to take ship for San Francisco in California. The arrangement revived interest—which lasted for many years—in the possibilities of a transoceanic canal. After the completion of the Panama Canal in 1914, interest in the project once more subsided. Since 1916, however, by the provisions of the Bryan–Chamorro Treaty concluded between Nicaragua and the United States, the United States has had the exclusive right to build such a canal.

Lake Nicaragua remains, at present, largely undeveloped. Should a transoceanic waterway be constructed, however, its situation would be transformed. Meanwhile, because of its scenic beauty, and its archaeological interest, as well as its facilities for fishing, sailing, and swimming, it constitutes a focus for tourism. Commercial fishing also offers good prospects.

BIBLIOGRAPHY. There are no books exclusively on Lake Nicaragua. MILES P. DU VAL, *Cadiz to Cathay: The Story of the Long Diplomatic Struggle for the Panama Canal* (1968), has some material; the PAN-AMERICAN UNION, *Nicaragua* (1964) and *Visit Nicaragua* (1957), are also useful. For a description of the flora and fauna of the area, see THOMAS BELT, *The Naturalist in Nicaragua* (1874 and 1911).

(R.A.)

Nicephorus II Phocas

Byzantine emperor from 963 to 969, Nicephorus II Phocas was a member of a great aristocratic landowning family that had distinguished itself under the Macedonian emperors in wars against the Arabs. He was the son and grandson of generals and before his accession he had commanded the Byzantine armies of the East (954–955) and had directed a victorious expedition against the Arabs in Crete in 960–961. In Byzantine history he is considered the military emperor *par excellence*, the artisan of Byzantine grandeur in the world of the 10th century, and the restorer of that imperial power in the East that had been fought for in battles against the infidel from the beginning of the 10th century.

Nicephorus II Phocas, coin, 10th century. In the British Museum.
Peter Clayton

Early life

Nicephorus Phocas was born in 912, the son of Bardas Phocas, an important Byzantine general in Anatolia, on the borders of the empire. He quickly embraced a military career of arms and as a young patrician distinguished himself at his father's side in a war against the Ḥamdānid Arabs in the East. In 954–955 the emperor Constantine VII Porphyrogenitus named him commander in chief of the armies of the East, to replace the aging Bardas. Nicephorus proceeded to restructure the army to reinforce discipline and improve recruiting. At this point he probably wrote the treatises on military tactics that are attributed to him, although proof is lacking.

The emperor Romanus II named him commander of a wartime expedition to liberate Crete (which had been controlled by the Arabs ever since 826), at great cost to Aegean populations and international commerce. This enterprise mobilized the entire Byzantine fleet and close to 24,000 men. Nicephorus gained the island with the capture of Chandax, now Iráklion, on March 7, 961. In a general massacre, the inhumanity of which revealed his fierce nature, he broke all Arab resistance. Aided by the monks, among whom was Athanasius, his spiritual director and founder of the Greek Orthodox monastery on Mt. Athos, Nicephorus achieved the reconsolidation of Christianity. He then returned to Constantinople with ʿAbd al-ʿAzīz, the last *amīr* of Crete, as his captive. This exploit, sung by the poet Theodosius the Deacon, realized the Byzantine dream (after dozens had failed to liberate Crete) of imperial mastery of the eastern Mediterranean. Later, as emperor, Nicephorus could state proudly that he controlled the seas. By that time, however, he had recovered Cilicia and the island of Cyprus and had captured other Muslim naval bases.

At the beginning of 962, Nicephorus attacked the Arabs of Cilicia and Syria, capturing more than 60 fortresses. After the death of Romanus II on March 15, 963, the situation in the capital changed. The Emperor's will had left a eunuch, Joseph Bringas, in charge of the affairs of state and the 20-year-old empress, Theophano, as acting regent for the legitimate emperors, Basil and Constantine, aged six and three, respectively. These circumstances do not seem to have tempted Nicephorus.

Rise to power

In spite of his great popularity, there was no indication that Nicephorus—whose physical appearance was report-

edly not very agreeable and who seemed destined under the influence of Athanasius the Athonite to embrace the monastic life—would end up seducing and being seduced by the young and beautiful empress. If such a plan existed at the time (and there is reason to believe it did) it was probably the brainchild of the ambitious Theophano, who was unhappy with Bringas' government. The people of Constantinople, aroused by Basil the chamberlain, revolted against Bringas; and the imperial army, through the intermediation of John Tzimisces, Nicephorus' faithful lieutenant, "obliged" the soldier to accept the crown at Caesarea on July 3, 963, and to march against Constantinople. On August 16, 963, Nicephorus was crowned in the Hagia Sophia by the patriarch Polyeuctus, and on September 20 he celebrated his marriage to Theophano.

Smitten with the young woman and influenced by his brother Leo Phocas, whose self-interested machinations (he was accused of speculating on the price of wheat) stirred up the discontent of the people of Constantinople, Nicephorus gradually became taciturn and suspicious even of his best advisers, who, one after another, were removed from office. As emperor, Nicephorus continued his exploits against the Arabs until finally, abandoned by all, he retired to the fortified palace of Boukoleion, which he had built for his personal safety. During the night of December 10, 969, he was killed there by former friends, guided by Tzimisces and advised by Theophano.

The contradictions in Nicephorus' life and character also marked his domestic politics. His government evoked unanimous discontent: the hostility of the people to the new fiscal charges and coinage debasement required by military needs; the exasperation of ecclesiastical authorities over decisions against enrichment of the monasteries; the remonstrances of his spiritual director, Athanasius, against his private life; and the apprehensions of Theophano that her children would be ousted through the machinations of Leo Phocas. These all created a climate of intrigue, which resulted in Nicephorus' assassination and brought John Tzimisces to the throne.

Military achievements

The failure of Nicephorus' domestic policies did not cast a shadow on his military achievement, which made his reign one of the Byzantine Empire's most glorious. In the words of C. Schlumberger, his most exhaustive biographer, he inaugurated the Byzantine era in the East. In fact, though known primarily for his exploits against the infidel, Nicephorus also carried the imperial frontier beyond the Euphrates to Syria. Nor did he neglect the other imperial frontiers in the conception of Byzantine grandeur. To counteract the Bulgar menace he spurred Russian intervention in the Danubian area, a policy that was not without danger for Byzantium, especially after his death. Also, to stop expansionist plans of the Germanic sovereign Otto I, who was re-creating the Carolingian heritage, Nicephorus opposed Otto's title of emperor, while trying with more or less success to consolidate the Byzantine presence in Italy. Nicephorus II's policies, seen in their entirety, indicate that his purpose was to assure Byzantium of its place as international arbiter, which he accomplished through the use of arms.

Phocas was indeed a Nicephorus (Bringer of Victory) for the empire. The Byzantines surnamed him Kallinikos, artisan of good victories; the Arabs called him Nikfour, the Saracen hammer. His death caused joy in the Muslim world and shook Christianity. His legend was quickly nourished with stories of his exploits and tragic death. Byzantine and even Bulgar poets were inspired by his exploits, and posterity has kept his memory alive: he is celebrated in the epic poetry of the frontier; the church beatified him (an *acolouthie* was composed in his honour); and the monks of Mt. Athos still venerate as their benefactor and founder Nicephorus, emperor and martyr. His life was summed up in the phrase inscribed on his sarcophagus: "You conquered all but a woman."

BIBLIOGRAPHY. C. SCHLUMBERGER, *Un Empereur byzantin au X^e siècle: Nicéphore Phocas* (1890), is a comprehensive biography. The same author's *L'Épopée byzantine à la fin du X^e siècle*, 3 vol. (1896–1905); and A.A. VASILIEV, *Byzance et les Arabes*, vol. 2, *La dynastie macédonienne* (1968), are useful for Nicephorus Phocas' career in the East. Also of interest is H. GREGOIRE, "The Amorians and the Macedonians, 842–1025," in *The Cambridge Medieval History*, new ed., vol. 4, pt. 1, pp. 147–156 (1966).

(H.Ah.)

Nichiren

Nichiren, a Buddhist monk of the Kamakura period (1192–1333), is one of the most significant figures in Japanese Buddhism. Together with the other Buddhist founders of that period (see HONEN and SHINRAN), he greatly contributed to the complete adaptation of Buddhism to the Japanese religious mentality. Because of his uncompromising commitment, militant personality, and concern with national regeneration through true religion, he has been called "the Japanese prophet."

Nichiren was born in 1222, the son of a fisherman, in the village of Kominato, on the Pacific coast of the present Bōsō-hantō in eastern Japan. When he was 11 years old, he entered the Buddhist monastery of Kiyosumi, near Kominato, and after four years of noviceship received the Buddhist orders. Buddhism in Japan had become more and more doctrinally confused, and the identity of the various sects was based more on institutional aspects than on doctrinal tenets. Though the monastery of Kiyosumi officially belonged to the Tendai sect, centred on the *Lotus Sūtra* text and realization of the universal Buddha-nature, the doctrine practiced there was a mixture of different Buddhist schools with a strong emphasis on Shingon, an esoteric school that emphasized an elaborate symbolic ritual.

Quest for true Buddhist teaching

The young monk was too intense and too sincere in his quest for the true doctrine of the Buddha to be satisfied with such prevailing confusion of doctrine. Soon his central spiritual problem was to find, through the maze of scriptures and doctrines, the authentic teaching the historical Buddha, Gautama (Śākyamuni), had preached for the salvation of mankind. So he undertook a thorough study of all the major Buddhist schools existing in Japan. In 1233 he went to Kamakura, where he studied Amidism—a pietistic school that stressed salvation through the invocation of Amida, the Buddha of infinite compassion—under the guidance of a renowned master. After having persuaded himself that Amidism was not the true Buddhist doctrine, he passed to the study of Zen, which had become popular in Kamakura and Kyōto. He then went to Hiei-zan, the cradle of Japanese Tendai Buddhism, where he found the original purity of the Tendai doctrine corrupted by the introduction and acceptance of other doctrines, especially Amidism and esoteric Buddhism (see BUDDHISM, *Shingon*). To eliminate any possible doubts, Nichiren decided to spend some time at Kōyasan, the centre of esoteric Buddhism, and also in Nara, Japan's ancient capital, where he studied the Ritsu sect, which emphasized strict monastic discipline and ordination.

By 1253 Nichiren had reached his final conclusion: the true Buddhism was to be found in the *Lotus Sūtra*, and all other Buddhist doctrines were only temporary and provisional steps used by the historical Buddha as a pedagogical method to lead men to the full and final doctrine contained in the *Lotus Sūtra*. Moreover, the Buddha himself had decreed that this doctrine was to be preached to men during the age of the Latter Law (Mappo)—the last, degenerate period after his death, the present age—and that a teacher would then appear to preach this true and final doctrine.

In the spring of 1253, Nichiren returned to Kiyosumi where he proclaimed his faith before his old master and his fellow monks, adding that all other forms of Buddhism were to be banished, for they were false and were misleading the people. Neither the monks of Kiyosumi nor the feudal lord of the region accepted his doctrine, and their angry reaction was such that he had to escape to save his life.

Expelled from his monastery, Nichiren lived in a little hut in Kamakura and spent his days preaching his doctrine at the busiest crossroads of the city. His constant attacks against all other sects of Buddhism attracted an ever-increasing hostility and finally open persecution

from Buddhist institutions and from the authorities. The country was at the time afflicted by epidemics, earthquakes, and internal strife. Reflecting on this sad situation, Nichiren is said to have read once again all the Buddhist scriptures and in 1260 published a short tract, *Risshō ankoku-ron* ("The Establishment of Righteousness and the Pacification of the Country"), in which he stated that the deplorable state of the country was due to the fact that the authorities and the people refused to follow true Buddhism and were supporting false sects. The only salvation was for the authorities and the people of Japan to accept Nichiren's doctrine as the national faith and banish all the other sects. If this was not done, the state of the country would become even worse, and Japan would be invaded by a foreign power. The military government in Kamakura reacted to this prophetic admonition by exiling the monk to a deserted place in the Izu-hantō (Izu Peninsula), in the present Shizuoka Prefecture, in June 1261. He was pardoned in 1263, but upon his return to Kamakura he renewed his attacks.

Nichiren's message to Japan

In 1268 an embassy from the Mongols—who had conquered China—arrived in Japan demanding that the Japanese become a tributary nation to the new Mongol dynasty. Nichiren saw in this the fulfillment of his prophecy of 1260. Once again he sent copies of his *Risshō ankoku-ron* to the authorities and the heads of the major Buddhist institutions, insisting again that if his doctrine was not accepted and the other sects were not banished, Japan would be visited with all sorts of calamities.

Again the authorities and the older Buddhist sects were enraged by the extraordinary audacity of this troublesome monk, and in 1271 Nichiren was arrested and condemned to death. The penalty was commuted at the last moment, and instead he was exiled to the island of Sado, in the Sea of Japan, where in 1272 he wrote his systematic work *Kaimokushō* ("The Opening of the Eyes"). According to Nichiren's account and the belief of his adherents, he was saved from execution by a miraculous intervention that struck the sword from the executioner's hand. While the fiery monk was in exile, a second and a third Mongol embassy arrived, threatening an invasion if Japan persisted in her refusal to become a vassal nation. Nichiren's prophecy and the pressure of his influential friends in Kamakura moved the government, and an edict of pardon was issued in the spring of 1274. In May, Nichiren arrived in Kamakura, where he met with high government officials and reiterated his stern requests. Though this time the authorities treated him with deference and respect, they still refused to comply with his demands.

Full of indignation, Nichiren left Kamakura in June and with a small number of disciples retired to a solitary place on Minobu-san, in the present Yamanashi Prefecture. There he spent his last years instructing his followers and writing. Among the main works of this period are the "Selection of the Time," a synthetic exposition of his philosophy of history, and "In Recompense of Indebtedness," in which a good life is seen as one of practical gratitude toward one's parents, all creatures, one's sovereign, and the Buddha.

The hardships and persecutions endured for so many years began to take their toll, and Nichiren's state of health grew worse and worse. His final illness was probably a cancer of the intestinal tract. In the fall of 1282 he left his hermitage at Minobu and took residence in the mansion of one of his disciples in the district of Ikegami (in what is now Tokyo), where he died on November 4, 1282.

Contribution to Japanese Buddhism

Nichiren is perhaps the most controversial figure in the history of Japanese Buddhism. Incapable of accepting any compromise, he described himself as a most intractable character. Yet letters he wrote to his disciples and friends reveal how loving, understanding, and even delicate he could be. He dearly loved Japan and wanted her to fulfill her mission of being the chosen country of Buddhism, from which Buddha's salvation was to spread to the entire world. His Buddhism was typically Japanese in the sense that it could not be confined to mere speculation or even to individual salvation but had to be concerned with the salvation of society and temporal institutions; hence, the importance he gave to the right understanding of history and human affairs. The continuing vitality of the religious system he founded in the 13th century is attested by the fact that many of the modern Buddhist sects now flourishing in Japan are, in various degrees, based on Nichiren's doctrines.

BIBLIOGRAPHY. MASAHARU ANESAKI, *Nichiren, the Buddhist Prophet* (1916, reprinted 1966), a highly readable work that gives a good exposition of Nichiren's life and doctrine—the only available biography in English; GEORGE B. SANSOM, "Nichiren," in SIR CHARLES ELIOT, *Japanese Buddhism* (1935, reprinted 1969), a brief but adequate exposition of Nichiren's doctrine.

(P.P.del C.)

Nicholas V, Pope

As pope from 1447 to 1455, Nicholas V (Tommaso Parentucelli), perhaps the most influential of the Renaissance popes, is notable for his sponsorship of the new Humanist learning and thus for the attempted conciliation of religion with secular culture. He was the founder of the world-famous Vatican Library.

Alinari

Nicholas V, effigy on his tomb in St. Peter's Basilica, Vatican City, 15th century.

Parentucelli was born November 15, 1397. His father died when he was nine. Later he studied at Bologna, but for lack of funds had to interrupt his studies there. Then, to earn money, he acted as tutor for two years in two wealthy, cultured Florentine families, and this contact with the early Renaissance coloured all his life. After returning to the university and completing his studies, at the age of 22 he entered the household of Niccolò Albergati, the cardinal-archbishop of Bologna, whom he served devotedly for 20 years, accompanying him on his many diplomatic missions throughout Europe. Pope Eugenius IV recognized Parentucelli's merit and experience and, on Albergati's death, made him bishop of Bologna (1444), but he was prevented from entering the city by its rebellious inhabitants, who sought independence from papal rule. At the Council of Ferrara–Florence (1438–45), he led the discussions with the Armenians, Copts, and Jacobites that attempted to end their doctrinal differences with the Latin Church, and he later journeyed on missions for the Pope. After his diplomatic success in pacifying the German Electors at the Diet of Frankfurt in 1446, he was created cardinal and only three months later, on March 6, 1447, was elected pope. Thenceforward, his chief aims were, in his words, "without using arms other than those which Christ has given me for my defence, that is to say, His Cross," to work for ecclesiastical and political peace, to reform the church, and to make Rome architecturally and artistically the worthy centre of Christianity.

Achievements in diplomacy and church reform

When Nicholas became pope, the remnant of the rebellious Council of Basel (1431–37)—which advocated control of the church through councils rather than through the pope—and its antipope Felix V still challenged Rome. Nicholas, by demanding little and yielding much, brought the schism to an end. He restored peace in

the papal states, not by mercenary armies, which he disbanded, but by strategically situated castles entrusted to carefully chosen governors. In Rome he was conciliatory with the nobility and granted concessions to the restive citizens. He allowed the town of Palestrina, destroyed by his predecessor, to be rebuilt; pacified Bologna by granting it virtual independence; won the allegiance of Poland by concessions; and, by promising coronation as Holy Roman emperor to Frederick III, gained the support of Austria. In order to give peace to Italy by preserving the status quo and to organize a crusade against the Turks, he initiated the Peace of Lodi among Venice, Milan, Florence, and Naples and solemnly ratified it on February 25, 1455.

Nicholas helped the church by furthering legislation against the old abuses of simony (buying and selling of church offices) and clerical concubinage and by encouraging bishops to govern their dioceses wisely. In Germany Cardinal Nicholas of Cusa (1401–64), a notable scholar and thinker, and the Franciscan theologian and preacher, Saint John of Capistrano (1386–1456), laboured hard at reform with much success. A wider measure to stimulate piety and restore the papal reputation was the proclamation of the 1450 Jubilee Year. Vast numbers of pilgrims visited Rome, and the project did much good, though it was marred by an outbreak of the plague and a tragedy when 172 people died in a panic-crush on the St. Angelo Bridge.

Pope Nicholas is best remembered for his influence on the Renaissance in Rome. "Of all Renaissance popes," says Eugène Müntz, a famous curator and art historian, "Nicholas is the one who ventilated the greatest number of architectural ideas: his successors only executed one or another element of his programme." He had plans for building a new St. Peter's Church but was able only to rebuild what was crumbling, to reconstruct the Vatican Palace, and to surround the whole with a wall. The restoration and embellishment of many Roman architectural treasures, such as the senatorial palace on the Capitoline Hill, are credited to him, but he pillaged ancient monuments to quarry materials for his new constructions. At his initiative, Rome became a centre for goldsmiths and silversmiths; he employed French, Belgian, and German tapestry makers; he commissioned artists of note, among them the great Florentine painter Fra Angelico (1387–1455) to beautify his constructions.

Architectural and Humanistic achievement

He had the Humanist's passion for books. On his diplomatic missions he sought them out, and as pope he spent vast sums on buying them. His court became a centre for Humanists, some of them more pagan in outlook than Christian; they were employed in copying and translating ancient texts, among them the works of Homer, Herodotus, Thucydides, and many Greek Church Fathers.

His last years were saddened by a plot against his life. Twice he dealt mercifully with the ringleader; the third time, in 1453, he had him and his accomplices executed. Also in 1453, Constantinople, the seat of Eastern Christianity, was captured by the Turkish sultan. Nicholas had ordered a fleet to aid the beleaguered city, but it arrived too late. This was a military reverse of great religious and cultural significance.

Nicholas was a man of gentle character who achieved more by wise concession than others did by force. Of him Vespasiano da Bisticca (1421–98), the biographer of 15th-century luminaries, wrote: "He did not know what avarice was: indeed, if he retained anything of his own, it was simply because no one had asked him for it." His diplomatic efforts on behalf of peace in Italy and elsewhere and his patronage of the arts, especially of literature, restored to the church much of the ancient prestige it had lately lost. His failure to promote sufficiently religious reform, however, was destined to result in the Reformation in the 16th century.

BIBLIOGRAPHY. VESPASIANO DA BISTICCI, *The Vespasiano Memoirs: Lives of Illustrious Men of the XVth Century* (Eng. trans. 1926), written by a Florentine closely involved in the Humanistic circles of the city; G. MANETTI, "Vita Nicolai V summi pontificis ex manuscripto codice Florentino," in *Muratori RIS* 3.2:907–960 (1734), written by a Florentine Humanist and politician who became apostolic secretary in the court of Nicholas V; L. PASTOR, *History of the Popes from the Close of the Middle Ages*, vol. 2, pp. 3–314 (1899), a classic work; E. MÜNTZ, *Les Arts à la cour des papes pendant le XV^e et XVI^e siècle*, vol. 1, pp. 68–189 (1878); E. MÜNTZ and P. FABRE (eds.), *La Bibliothèque du Vatican au XV^e siècle* (1887), authoritative books based on close study of Vatican registers and accounts.

(J.Gi.)

Nicholas I of Russia

Nicholas I, emperor of Russia from 1825 to 1855, was by circumstance and choice a military man, whose uncompromising bearing, personality, and deeds made him the personification of the classic autocrat. A reactionary, he has been called the emperor who froze Russia for 30 years. He may have summarized his philosophy of life in the following passage:

> Here [in the army] there is order, there is a strict unconditional legality, no impertinent claims to know all the answers, no contradiction, all things flow logically one from the other; no one commands before he has himself learned to obey; no one steps in front of anyone else without lawful reason; everything is subordinated to one definite goal, everything has its purpose. That is why I feel so well among these people, and why I shall always hold in honour the calling of a soldier. I consider the entire human life to be merely service, because everybody serves.

By courtesy of Mrs. Merriweather Post, Hillwood, Washington, D.C.

Nicholas I, watercolour by Christina Robertson, 1840. In the collection of Mrs. Merriweather Post, Hillwood, Washington, D.C.

Early life. Nicholas (Nikolay Pavlovich) was born on July 6 (June 25, old style), 1796. His parents were Grand Duke Paul and Grand Duchess Maria. Some three and a half months later, following the death of Catherine II the Great, Nicholas' father became Emperor Paul I of Russia. Nicholas had three brothers, two of whom, the future emperor Alexander I and Constantine, were 19 and 17 years older than he. It was the third, Michael, his junior by two years, and a sister, Anne, who became his childhood companions and intimate lifelong friends.

Family relations

Paul was extremely neurotic, overbearing, and despotic. Yet it is believed that he showed kindness and consideration to his younger children and that, in fact, he loved and cherished them tenderly. He was killed in a palace revolution of 1801, which made Alexander emperor when Nicholas was not quite five years old. Maria, on the contrary, remained formal and cold in her relationship to the children, very much in keeping with her general character. She belonged, apparently, among those human beings who combine numerous conventional virtues

with a certain rigidity and lack of warmth. In the words of a competent observer: "The only failing of this extraordinary woman was her being excessively, one may say, exacting of her children and of the people dependent on her."

Education. The future emperor's first guardian and instructress was a Scottish nurse, Miss Jane Lyon, who was appointed by Catherine II to care for the infant and who stayed with Nicholas constantly during the first seven years of his life. From Miss Lyon the young grand duke learned even such things as the Russian alphabet, his first Russian prayers, and his hatred of the Poles (at least he liked later to trace the origin of his bitter antipathy toward that people to the stories told by his nurse about her painful experience in Warsaw in the turbulent year of 1794). In 1802–03 men replaced women in Nicholas' entourage, and his regular education began. As directed by Gen. Matthew Lamsdorff, it emphasized severe discipline and formalism. The growing grand duke studied French and German as well as Russian, world history, and general geography in French, together with the history and geography of Russia. Religion, drawing, arithmetic, geometry, algebra, and physics were added to the curriculum. Nicholas received instruction also in dancing, music, singing, and horseback riding and was introduced at an early age to the theatre, costume balls, and other court entertainment. In 1809 a more advanced curriculum went into effect, with courses ranging from political economy, logic, moral philosophy, and natural law to strategy. English, Latin, and Greek were added to the language program. Though, on the whole, a belief that Nicholas had not been trained for his role of Russian sovereign is wrong, he did profit little from the instruction, which he found rigid and tedious. He loved only military science, becoming a fine army engineer and expert in several other areas of military knowledge. Moreover, he always remained in his heart a dedicated junior officer.

Love for military science

Circumstances also favoured militarism. Nicholas' education, as well as that of his younger brother, was interrupted and largely terminated by the great struggles against Napoleon in 1812–15. The grand dukes were allowed to join the army in 1814; and, although they saw no actual fighting, they lived through the heady emotions of those momentous years and also enjoyed the opportunities to stay in Paris and other places in western and central Europe. On November 4, 1815, at a state dinner in Berlin, Alexander I and King Frederick William III rose to announce the engagement of Nicholas and Princess Charlotte of Prussia (Alexandra, after she became Orthodox). The solemn wedding followed some 20 months later, on July 13, 1817. The match represented a dynastic and political arrangement sought by both reigning houses, which had stood together in the decisive years against Napoleon and after that at the Congress of Vienna—the peace settlement following the Napoleonic Wars—and it proved singularly successful. Not only was Nicholas in love with his wife, but he became very closely attached to his father-in-law as well as to his royal brothers, one of whom was later to be his fellow ruler as King Frederick William IV. Beyond that, Nicholas was powerfully attracted by the Prussian court and even more so by the Prussian Army. He felt remarkably happy and at home in his adopted family and country, which for many years he tried to visit as often as he could.

To complete his training, Grand Duke Nicholas was sent on two educational voyages—an extensive tour of Russia that lasted from May to September in 1816 and a journey to England, where the future emperor spent four months late that same year and early in 1817. The Russian trip covered much ground at great speed and was quite superficial, but it has interest for the historian because of the notes that Nicholas, following the instructions of his mother, took on everything seen and heard. The grand duke's observations deal, typically, with appearances rather than with causes and reflect a number of his prejudices, including his bitter dislike of Poles and Jews. Such quick inspection tours later became almost an obsession of the Emperor. In England, Nicholas stayed mostly in London, although he travelled to a score of other places. While he did attend the opening of the houses of Parliament and in general obtained some knowledge of English politics, his only recorded comments on that score were unfavourable. The future emperor found it much more congenial to examine military and naval centres. His favourite English companion was the Duke of Wellington. Less than a year after his return to Russia and a few months after his marriage, Nicholas was appointed inspector general of the army corps of engineers. In subsequent years he held several other military positions but of secondary significance.

Ascension to the throne. Alexander I's unexpected death in southern Russia on December 1, 1825, led to a dynastic crisis. Because Alexander I had no direct male successor, Constantine was next in line for the throne. But the heir presumptive had married a Polish woman not of royal blood in 1820 and renounced his rights to the crown. Nicholas was thus to become the next ruler of Russia, the entire matter being stated, in 1822, in a manifesto confirmed with Alexander I's signature. But the manifesto remained unpublished, and Nicholas questioned the legal handling of the whole issue and the reaction in the country, which expected Constantine to succeed Alexander. In any case, Constantine and the Polish kingdom of which he was commander in chief swore allegiance to Nicholas; but Nicholas, the Russian capital, and the Russian Army swore allegiance to Constantine. It was only after Constantine's uncompromising reaffirmation of his position and the resulting lapse of time that Nicholas decided to publish Alexander's manifesto and become emperor of Russia. On December 26, 1825 (December 14, O.S.), when the guard regiments in St. Petersburg were to swear allegiance for the second time in rapid succession, this time to Nicholas, liberal conspirators staged what came to be known as the Decembrist Rebellion. Utilizing their influence in the army, in which many of them were officers, they started a mutiny in several units, which they entreated to defend the rightful interests of Constantine against his usurping brother. Altogether some 3,000 misled rebels marched in military formation to the Senate Square—now the Decembrist Square—in the heart of the capital. Although the rebellion had failed by nightfall, it meant that Nicholas I ascended the throne over the bodies of some of his subjects and in actual combat with the dreaded revolution.

Dynastic crises

Personality. Nicholas I has come down in history as the classic autocrat, in appearance and manner as much as in behaviour and policy. To quote Andrew Dickson White, a United States diplomat:

Appearance and manner

> With his height of more than six feet, his head always held high, a slightly aquiline nose, a firm and well-formed mouth under a light moustache, a square chin, an imposing, domineering, set face, noble rather than tender, monumental rather than human, he had something of Apollo and of Jupiter . . . Nicholas was unquestionably the most handsome man in Europe.

Or to refer to Adolphe, marquis de Custine, whose lasting literary fame rests on his denunciation of the Russia of Nicholas I: "Virgil's Neptune . . . one could not be more emperor than he." In short, Nicholas I came to represent autocracy personified: infinitely majestic, determined and powerful, hard as stone, and relentless as fate. Yet, on closer acquaintance, the other side of the Emperor emerged. The detachment and the superior calm of an autocrat, which Nicholas I tried so often and so hard to display, were essentially a false front. The sovereign's insistence on firmness and stern action was based on fear, not on confidence; his determination concealed a state approaching panic, and his courage fed on something akin to despair. Nicholas' violent hatred could concentrate apparently with equal ease on an individual, such as the French king Louis-Philippe; a group, such as the Decembrists; a people, such as the Poles; or a concept, such as revolution. His impulse was always to strike and keep striking until the object of his wrath was destroyed.

Aggressiveness, however, was not the Emperor's only method of coping with the problems of life. He also used regimentation, orderliness, neatness, and precision, an

enormous effort to have everything at all times in its proper place. Nicholas I was by nature a drill master and an inspector general; the army remained his love, almost an obsession, from childhood to the end of his life. But, in every other sphere of activity and existence too, the Emperor insisted on minute and precise regulation, with nothing to be left to chance. Position, circumstances, and his own character placed an almost intolerable burden on his shoulders. Still, he managed to carry it for three decades, sustained by his overwhelming sense of duty and devotion to hard work, by his sincere religious convictions, and by his family. His outlook, however, became ever more pessimistic and fatalistic, until in the disaster of the Crimean War the autocrat declared simply: "I shall carry my cross until all my strength is gone. Thy will be done."

Ideology. Nicholas' views fitted his personality to perfection. In contrast to Alexander I, he had been brought up at the time of wars against Napoleon and of reaction, which he accepted wholeheartedly as his own cause. Eventually the Russian wing of European reaction, represented by Nicholas I and his government, found its ideological expression in the doctrine of so-called Official Nationality. Formally proclaimed in 1833 by Count Sergey Uvarov, the Emperor's minister of education, Official Nationality rested on three principles: Orthodoxy, autocracy, and nationality. Autocracy meant the affirmation and maintenance of the absolute power of the sovereign, which was considered the indispensable foundation of the Russian state; in foreign relations it was transformed into legitimism and a defense of the Vienna settlement. Orthodoxy referred to the official church and its important role in Russia and also to the ultimate source of ethics and ideals that gave meaning to human life and society. Nationality (*narodnost*) described the particular nature of the Russian people, considered as a mighty and dedicated supporter of its dynasty and government. Whereas Alexander I had never quite abandoned dreams of change, Nicholas I was determined to defend the existing order in his motherland, especially autocracy.

Doctrine of Official Nationality

Reign. Nicholas I's rule reflected in a striking manner both his character and his principles. The new regime became pre-eminently one of militarism and bureaucracy. The Emperor surrounded himself with military men, to the extent that late in his reign there were almost no civilians among his immediate assistants. Also, he relied heavily on special emissaries, most of them generals of his suite, who were sent all over Russia on particular assignments to execute immediately the will of the sovereign. Operating outside the regular administrative system, they represented an extension, so to speak, of the monarch's own person. In fact, the entire machinery of government came to be permeated by the military spirit of direct orders, absolute obedience, and precision, at least as far as official reports and appearances were concerned. Corruption and confusion, however, lay immediately behind this facade of discipline and smooth functioning.

In his conduct of state affairs, Nicholas I often bypassed regular channels and generally resented formal deliberation, consultation, or other procedural delay. The importance of the Committee of Ministers, the State Council, and the Senate decreased in the course of his reign. Instead of making full use of them, the Emperor depended more and more on special bureaucratic devices meant to carry out his intentions promptly while remaining under his immediate and complete control. As one favourite method, Nicholas I made extensive use of ad hoc committees that stood outside the usual state machinery. The committees were typically composed of a handful of the most trusted assistants of the Emperor; because these were few in number, the same men in different combinations formed these committees throughout Nicholas' reign. As a rule, the committees carried on their work in secret, adding further complication and confusion to the already cumbersome administration of the empire. The failure of one committee to perform its task merely led to the formation of another. For example, some nine committees tried to deal with the issue of serfdom during Nicholas' reign.

Administration of the empire

The propensities of the autocrat found expression also in the development and the new role of His Majesty's Own Chancery. Organized originally as a bureau to deal with matters that demanded the sovereign's personal participation and to supervise the execution of the Emperor's orders, it acquired five new departments: in 1826 the Second and the Third, to deal with the codification of law and the newly created corps of gendarmes, respectively; in 1828 the Fourth, to manage the charitable and educational institutions under the jurisdiction of the empress dowager Maria; in 1836 the Fifth, to reform the condition of the state peasants (soon replaced by the new Ministry of State Domains); in 1843 the Sixth, to draw an administrative plan for Transcaucasia. The departments of the Chancery served Nicholas I as a major means of conducting a personal policy that bypassed the regular state channels. Its Third Department, the political police, acted as the autocrat's main weapon against subversion and revolution and as his principal agency for controlling the behaviour of his subjects and for distributing punishments and rewards among them. Its assigned fields of activity ranged from "all orders and all reports in every case belonging to the higher police" to "reports about all occurrences without exception!" The two successive heads of the Third Department—Count Aleksandr Benckendorff and Prince Aleksey Orlov—probably spent more time with Nicholas than did any of his other assistants; they accompanied him, for instance, on his repeated trips of inspection throughout Russia. During his entire reign the Emperor strove to follow the principle of autocracy—to be a true father of his people concerned with their daily lives, hopes, and fears.

Yet Nicholas I could do little for them beyond the minutiae. Determined to preserve autocracy, afraid to abolish serfdom, and suspicious of all independent initiative and popular participation, the Emperor and his government could not introduce in their country the much-needed basic reforms. In practice as well as in theory they looked backward. Important developments took place only in a few areas in which change would not threaten the fundamental structure of the Russian Empire. Thus Count Mikhail Speransky codified law, and Count Pavel Kiselev changed and improved the lot of the state peasants; but even limited reforms became impossible after 1848. Frightened by European revolutions, Nicholas I became completely reactionary. During the last years of the reign the Emperor's once successful foreign policy collapsed, leading to isolation and to the tragedy of the Crimean War. A dauntless champion of legitimism and a virtual hegemon of eastern and central Europe following the revolutions of 1848–49, Nicholas—in part because of his own miscalculations, rigidity, and bluntness—found himself alone fighting the Crescent (the Ottoman Empire), supported by such countries of the Cross as France, Great Britain, and Sardinia.

Isolation and tragedy

Although it is unlikely that Nicholas committed suicide, as several historians have claimed, death did come as liberation to the weary and harassed Russian emperor. An ordinary cold picked up in late February 1855 turned into pneumonia, which the once mighty, but now apparently exhausted, organism refused to fight. To the end the autocrat retained lucidity and dignity.

Nicholas died in St. Petersburg on March 2 (February 18, O.S.), 1855. His last words to his heir and his family were: "Now I shall ascend to pray for Russia and for you. After Russia, I loved you above everything else in the world. Serve Russia." Nicholas I was survived by his wife, Empress Alexandra, and their six children: Emperor Alexander II, grand dukes Constantine, Nicholas, and Michael, and grand duchesses Maria and Olga. Another daughter, Grand Duchess Alexandra, had died in 1844.

BIBLIOGRAPHY. The leading full account of Nicholas I and his reign, especially valuable on foreign relations and with numerous documentary appendixes, is THEODOR SCHIEMANN, *Geschichte Russlands unter Kaiser Nikolaus I*, 4 vol. (1904–19). An emperor- and court-centred history, incomplete but very rich in primary sources, is НИКОЛАЙ ШИЛДЕР, *Император Николай Первый, его жизнь и царствование*, 2 vol.

(1903). A readable popular study is CONSTANTIN DE GRUNWALD, *La Vie de Nicolas Ier* (1946; Eng. trans., *Tsar Nicholas I*, 1954). See also NICHOLAS V. RIASANOVSKY, *Nicholas I and Official Nationality in Russia, 1825–1855* (1959).

(N.V.R.)

Nicholas II of Russia

The forced abdication of Tsar Nicholas II in 1917 marked the end of centuries of imperial government in Russia; and the execution by the Bolsheviks of him and his family during the ensuing civil war effectively extinguished the chances of a restoration of the monarchy. By nature timid and vacillating, Nicholas had early embraced the God-given role of autocrat of all Russia. His reactionary attitude, coupled with his generally inept handling of domestic and foreign affairs, contributed much to the popular discontent that led to his downfall and to four years of revolution and civil war.

By courtesy of Mrs. Merriweather Post, Hillwood, Washington, D.C.

Nicholas II, watercolour by an unknown artist. In the collection of Mrs. Merriweather Post, Hillwood, Washington, D.C.

He was born at Tsarskoye Selo, one of the tsar's summer palaces near St. Petersburg, on May 18 (May 6, old style), 1868, the eldest son of the tsarevich Aleksandr Aleksandrovich (emperor as Alexander III from 1881) and his consort Maria Fyodorovna (Dagmar of Denmark). Succeeding his father on November 1, 1894, he was crowned in Moscow on May 26, 1895.

Neither by upbringing nor by temperament was Nicholas fitted for the complex tasks that awaited him as autocratic ruler of a vast empire. He had received a military education from his tutor, and his tastes and interests were those of the average young Russian guards officer of his day. He had few intellectual pretensions but delighted in physical exercise and the minutiae of army life: uniforms, insignia, parades. Yet on formal occasions he felt ill at ease. Though he possessed great personal charm, he was by nature timid; he shunned close contact with his subjects, preferring the privacy of his family circle. His domestic life was serene. To his wife, Alexandra, whom he had married on November 26, 1894, Nicholas was passionately devoted. She had the strength of character that he lacked, and he fell completely under her sway. Under her influence he sought the advice of spiritualists and faith healers, most notably Rasputin, who eventually acquired great power over the imperial couple. Nicholas also had other irresponsible favourites, often men of dubious probity who provided him with a distorted picture of Russian life, but one that he found more comforting than that contained in official reports. He distrusted his ministers, mainly because he felt them to be intellectually superior to himself and feared that they sought to usurp his sovereign prerogatives. His view of his role as autocrat was childishly simple: he derived his authority from God, to whom alone he was responsible, and it was his sacred duty to preserve his absolute power intact. He lacked, however, the strength of will necessary in one who had such an exalted conception of his task. In pursuing the path of duty, Nicholas had to wage a continual struggle against himself, suppressing his natural indecisiveness and assuming a mask of self-confident resolution. His dedication to the dogma of autocracy was an inadequate substitute for a constructive policy, which alone could have prolonged the imperial regime.

Influence of Alexandra

Soon after his accession Nicholas proclaimed his uncompromising views in an address to liberal deputies from the *Zemstva*, the self-governing local assemblies, in which he dismissed as "senseless dreams" their aspirations to share in the work of government. He met the rising ground swell of popular unrest with intensified police repression. In foreign policy, his naïveté and lighthearted attitude toward international obligations sometimes embarrassed his professional diplomats; for example, he concluded an alliance with the German emperor William II during their meeting at Björkö in July 1905, although Russia was already allied with France, Germany's traditional enemy.

He was the first Russian sovereign to show personal interest in Asia, visiting in 1891, while still tsarevich, India, China, and Japan; and later he nominally supervised the construction of the Trans-Siberian Railway. His attempt to maintain and strengthen Russian influence in Korea, where Japan also had a foothold, was partly responsible for the Russo-Japanese War (1904–05). Russia's defeat not only frustrated Nicholas' grandiose dreams of making Russia a great Eurasian power, with China, Tibet, and Persia under its control, but also presented him with serious problems at home, where discontent grew into the revolutionary movement of 1905.

Nicholas considered all who opposed him, regardless of their views, as malicious conspirators. Disregarding the advice of his prime minister, Sergey Yulyevich Witte, he refused to make concessions to the constitutionalists and workers until events forced him to yield more than might have been necessary had he been more flexible. On March 3, 1905, he reluctantly agreed to create a national representative assembly, or Duma, with consultative powers; and by the manifesto of October 30 he promised a constitutional regime under which no law was to take effect without the Duma's consent, as well as a democratic franchise and civil liberties. Nicholas, however, cared little for keeping promises extracted from him under duress. He strove to regain his former powers and ensured that in the new "fundamental laws" (May 1906) he was still designated an autocrat. He furthermore patronized an extremist right-wing organization, the Union of the Russian People, which sanctioned terrorist methods and disseminated anti-Semitic propaganda. Witte, whom he blamed for the October manifesto, was soon dismissed, and the first two Dumas were prematurely dissolved as "insubordinate." Pyotr Arkadyevich Stolypin, who replaced Witte and carried out the coup of June 16, 1907, dissolving the second Duma, was loyal to the dynasty and a capable statesman. But the Emperor distrusted him and allowed his position to be undermined by intrigue. Stolypin was one of those who dared to speak out about Rasputin's influence and thereby incurred the displeasure of the Empress. In such cases Nicholas generally hesitated but ultimately yielded to Alexandra's pressure. To prevent exposure of the scandalous hold Rasputin had on the imperial family, Nicholas interfered arbitrarily in matters properly within the competence of the Holy Synod, backing reactionary elements against those concerned about the Orthodox Church's prestige.

Foreign affairs

After its ambitions in the Far East were checked by Japan, Russia turned its attention to the Balkans. Nicholas sympathized with the national aspirations of the Slavs and was anxious to win control of the Turkish straits but tempered his expansionist inclinations with a sincere desire to preserve peace among the great powers. After the assassination of the Austrian archduke Francis Ferdinand at Sarajevo, he tried hard to avert the impending war by diplomatic action and resisted, until July 30, 1914, the pressure of the military for general, rather than par-

tial, mobilization. The outbreak of World War I temporarily strengthened the monarchy, but Nicholas did little to maintain his people's confidence. The Duma was slighted, and voluntary patriotic organizations were hampered in their efforts; the gulf between the ruling group and public opinion grew steadily wider. Alexandra turned Nicholas' mind against the popular commander in chief, his father's cousin the grand duke Nicholas; and on September 5, 1915, he dismissed him, assuming supreme command himself. Since the Emperor had no experience of war, almost all his ministers protested against this step as likely to impair the army's morale. They were overruled, however, and soon dismissed.

Nicholas II did not, in fact, interfere unduly in operational decisions, but his departure for headquarters had serious political consequences. In his absence, supreme power in effect passed, with his approval and encouragement, to the Empress. A grotesque situation resulted: in the midst of a desperate struggle for national survival, competent ministers and officials were dismissed and replaced by worthless nominees of Rasputin. The court was widely suspected of treachery, and antidynastic feeling grew apace. Conservatives plotted Nicholas' deposition in the hope of saving the monarchy. Even the murder of Rasputin failed to dispel Nicholas' illusions: he blindly disregarded this ominous warning, as he did those by other highly placed personages, including members of his own family. His isolation was virtually complete.

Abdication and death

When riots broke out in Petrograd on March 8, 1917, Nicholas instructed the city commandant to take firm measures and sent troops to help restore order. It was too late. The government resigned, and the Duma, supported by the army, called on the Emperor to abdicate. At Pskov, on March 15, with fatalistic composure, Nicholas renounced the throne—not, as he had originally intended, in favour of his son Alexis but in favour of his brother Michael, who, however, refused the crown.

Nicholas was detained at Tsarskoye Selo by Prince Lvov's provisional government. It was planned to send him and his family to England; but instead, mainly because of the opposition of the Petrograd Soviet, the Revolutionary Workers' and Soldiers' Council, they were removed to Tobolsk, in western Siberia. This step sealed their doom. In April 1918 they were taken to Yekaterinburg (now Sverdlovsk) in the Urals. When the anti-Bolshevik "White" Russian forces approached the area, the local authorities were ordered to prevent a rescue; and in the night of July 16/17 the prisoners were all slaughtered in the cellar of the house where they had been confined. The bodies were burned and cast into an abandoned mine shaft, but the facts were established by investigation after Yekaterinburg had been taken by the "White" forces.

BIBLIOGRAPHY. R.K. MASSIE, *Nicholas and Alexandra* (1967), is a popular study of life at the imperial court. A reliable general survey of Nicholas' reign is given by RICHARD CHARQUES in *The Twilight of Imperial Russia* (1958); on the last years, compare SIR BERNARD PARES, *The Fall of the Russian Monarchy* (1939 and 1961); and GEORG KATKOV, *Russia 1917: The February Revolution* (1967), the latter is more favourable to Nicholas. СЕРГЕЙ ПЕТРОВИЧ МЕЛЬГУНОВ, *Судьба Императора Николая II после отречения; историко-критические очерки* (1951), gives a thorough account of the Emperor's fate after his abdication. For a selection from his correspondence with Alexandra, see *The Letters of the Tsar to the Tsaritsa, 1914-1917*, ed. by C.E. VULLIAMY (1929; reprinted, with Alexandra's letters, in 2 vol., 1970).

(J.L.H.K.)

Nickel Products and Production

Nickel was used industrially as an alloying metal almost 2,000 years before it was isolated and recognized as a new element. Although it is best known for its use in coinage, nickel has become much more important for its many industrial uses, particularly as a constituent in alloy steels.

History. As early as 200 BC, the Chinese made substantial amounts of a white alloy from a copper–nickel ore found in Yunnan province and zinc. The alloy, known as *pai-t'ung*, was exported to the Middle East and even to Europe.

Later, miners in Saxony encountered what appeared to be a copper ore but found that processing it yielded only a useless slaglike material. They considered it bewitched and ascribed it to "Old Nick." Thus, it became known as kupfernickel (Old Nick's copper). It was from this ore, studied by A.F. Cronstedt, that nickel was isolated and recognized as a new element in 1751. In 1776 it was established that *pai-t'ung*, now called nickel-silver, was composed of copper, nickel, and zinc.

Demand for nickel-silver was stimulated in England around 1844 by the development of silver electroplating, for which it was found to be the most desirable base. The use of pure nickel as a corrosion-resistant electroplated coating developed a little later; both these uses continue, with nickel plating much the more important.

Ores and mining. Small amounts of nickel were produced in Germany in the mid-19th century. More substantial amounts came from Norway, and a little from a mine at Gap, Pennsylvania. A new source, New Caledonia in the South Pacific, came into production about 1877 and dominated until the development of the copper-nickel ores of the Copper Cliff-Sudbury, Ontario, region in Canada, which since 1905 has been the world's largest source of nickel. Several producers are active in Canada.

Canadian ores are sulfides containing nickel, copper, and iron. The most important nickel mineral is pentlandite $(Ni,Fe)_9S_8$, followed by pyrrhotite, usually ranging from FeS to Fe_7S_8, in which some of the iron may be replaced by nickel. Chalcopyrite, $CuFeS_2$, is the dominant copper mineral in these ores, with small amounts of another copper mineral, cubanite, $CuFe_2S_3$. Some gold, silver, and the six platinum-group metals also are present, and their recovery is important. Cobalt, selenium, tellurium, and sulfur also may be recovered from the ores.

Table 1: World Production of Nickel (1969)

	metric tons
Canada	192,700
U.S.S.R.	105,000
New Caledonia	90,474
Cuba	35,200
U.S.	14,167
Europe	11,294
Australia	10,796
Rhodesia	8,000
Indonesia	7,025
South Africa	5,500
Brazil	1,089

Source: U.S. Department of the Interior, Bureau of Mines, *Minerals Yearbook, 1969*.

Other important classes of ore are the laterites, which are the result of long weathering of peridotite initially containing a small percentage of nickel. Weathering in subtropical climates removes a major portion of the host rock, but the contained nickel dissolves and percolates downward and may reach a concentration sufficiently high to make mining economical. The nickel magnesium silicate, garnierite, is the richest in nickel. The New Caledonian deposits are of this type, and numerous other laterite deposits are scattered around the world, presenting a wide range of mining, transport, and recovery problems. The nickel content of laterites varies widely: at Le Nickel in New Caledonia, for example, the ore delivered to the smelter in 1900 contained 9 percent nickel; currently it contains around 3 percent, but production from this source is increasing substantially.

Laterite deposits

The search for new ore bodies is an essential part of a mining enterprise. Though prospectors using visual surface indications found the earliest ore bodies, many sulfides, particularly in Canada, are covered with clay and give no surface clues. Geophysical methods are thus required, and magnetic and electrical methods have been employed. Nickel ore bodies may contain magnetic components that are reasonably good conductors of electricity. Much of the surface, however, is wooded, rough, or boggy, so efforts have been made to devise airborne prospecting devices. An air magnetometer developed dur-

ing World War II as an antisubmarine device was thus employed after the war by International Nickel Company. A second method, developed by the same company, is the airborne conductor sensor or electromagnetic device that detects changes in the conductivity of the earth by inductive methods. It opened a new era in prospecting, not only for the nickel–copper sulfides but also for other sulfide ores and is now used throughout the world.

Refining. The Sudbury sulfide deposits were originally identified as copper deposits, and the Canadian Copper Company was organized in 1885 to develop the mines. But the new ore, which contained 2 percent of nickel, along with copper and iron, baffled copper refiners. The owner of a copper smelter at Bayonne, New Jersey, who contracted to buy 100,000 tons of the new ore, found that it could not be handled by normal copper smelting. After extensive experimentation, it was discovered that by adding sodium sulfide and allowing the molten mixture to cool slowly, two layers, which could be broken apart, formed. The top layer contained most of the sodium sulfide and copper sulfide; the bottom was largely nickel sulfide. If the bottom was again melted with more sodium sulfide and slowly cooled, the separation was improved. After several repetitions, a nickel sulfide of good purity resulted. This could be roasted to remove the sulfur, then reduced to yield nickel of adequate purity for use. The process, known as "Orford tops and bottoms," was used until 1948, when it was supplanted by a scheme first proposed in 1932, when it was determined that the low-temperature mutual solubility of copper sulfide (Cu_2S) and nickel sub-sulfide (Ni_3S_2), which constituted the mixture of sulfides (the matte), was very small and could be selectively floated as mechanical mixtures of the two sulfides. Though first efforts in selectively floating solidified matte containing these two components were not successful, good separation was achieved after much finer grinding of the solidified matte was employed.

In the present practice, the ore is crushed and ground with water to liberate the sulfides from the less dense gangue (the waste part of the ore). Density differences as well as magnetic properties are utilized in this step. The sulfide minerals can then be separated from the other minerals by selective flotation, in which chosen minerals attach themselves to air bubbles and form a froth. This process has been refined to achieve good separation of the copper-containing chalcopyrite from the nickel-containing fractions. The nickel-containing fraction of pyrrhotite-pentlandide from the flotation process, plus material recovered by magnetic concentration, is roasted to remove a portion of the sulfur, and the material is then smelted to remove much of the iron and silica. The resulting matte consisting of nickel, copper, and iron sulfides is further oxidized in a Bessemer converter to yield essential copper and nickel sulfides. These are separated by slowly cooling the matte, grinding it, and selectively floating it to yield essentially nickel sulfide and copper sulfide. In subsequent operations, the two sulfides are reduced to metal and converted into anodes for electrolytic refining to yield high-purity nickel and copper. Special methods have been devised to recover nickel from pyrrhotite and to produce a marketable iron ore.

Selective flotation

At another Canadian mine, impure nickel sulfide is cast into anodes that are heat-treated to render them less brittle and are then decomposed electrolytically in a sulfate solution, a process that yields nickel cathodes of good purity. The sulfur, precious metals, and copper are also recovered.

A unique method for recovering high-purity nickel from matte is employed at a nickel plant at Clydach, Wales, where the impure sulfide is roasted to remove all of the sulfur and then processed to form finely divided metal, which is treated with carbon monoxide gas at about 55° C that reacts to produce volatile nickel carbonyl, $Ni(CO)_4$. This in turn is decomposed at about 230° C to yield pure nickel in the form of small pellets.

Another version of the carbonyl nickel process is used at Clydach to produce very fine nickel powder for many powder-metallurgy applications. In this process, the reduced nickel reacts with carbon monoxide at about 120° C at a pressure of 300 pounds per square inch. The resulting $Ni(CO)_4$ is condensed, fractionated to remove iron carbonyl, and decomposed at high temperature to yield high-purity nickel powder of sizes and shapes required for various end uses, including the nickel–cadmium storage battery.

Refining laterite ores is difficult because of the impossibility of concentrating the nickel fraction by gravity or flotation. In processing laterites by pyrometallurgy, all the diluent material must be removed by means of a flux, a process requiring large amounts of fuel or electrical energy. In some cases, sulfur is added in the form of gypsum to produce a matte that can be further processed in the manner employed for sulfide ores. The copper content usually is low, however, so processing is somewhat simpler. When sulfur is not added, the nickel plus considerable iron may be reduced with carbon or silicon to yield a ferronickel alloy.

An alternative to pyrometallurgy is chemical extraction of the nickel. With one type of laterite, the extraction of nickel, plus some iron and cobalt, can be accomplished by treatment with a sulfuric solution under high temperature and pressure. Alternatively, extraction can be effected by treatment with an aqueous solution containing ammonia or ammonia plus carbon dioxide.

Chemical extraction

It is difficult to generalize about the worldwide methods used to refine nickel since the methods depend upon the ore, the cost of electric power, and the market for the various end products that are possible. In the U.S.S.R., most of the production comes from the north, Norilsk in western Siberia and the Pechenga area of the Kola Peninsula. The ores involved are the copper-nickel-iron-sulfide type similar to those of Canada. The methods employed for processing them are similar to those used in Canada, including selective flotation of the matte. The remainder of Russian production comes from the southern Urals. The various oxide ores of this area are smelted with gypsum to provide enough sulfur to produce a matte, the same method long used at New Caledonia.

Production. Tables 1 and 2 give the world production of nickel and the distribution of uses in the U.S.

Characteristics and uses of nickel and its alloys. Pure nickel possesses a useful combination of properties including corrosion resistance, good strength, and high ductility, even at extremely low temperatures. It also possesses useful electronic properties and special magnetic properties. Nickel is a particularly good catalyst for reactions involving hydrogen, and substantial quantities of it are used in hydrogenating natural oils. Nickel catalyzes the addition of hydrogen to unsaturated compounds in vegetable, animal, and fish oils, converting them from liquids to solids. Natural oils treated in this way are used in such products as shortening, oleomargarine, and soap.

Table 2: Distribution of Nickel Use in U.S. (1969)

use	percent of total production
Stainless steel	30.0
High nickel alloys	21.6
Alloy steels	10.7
Ferrous castings	11.0
Iron–nickel alloys	2.0
Nonferrous alloys	6.5
Plating and chemicals	17.4
Coinage and powder	0.8

The white colour of nickel is attractive, and most of its alloys with copper are substantially white. Its ability to form strong, ductile alloys with many metals, including iron, chromium, cobalt, copper, and gold, is utilized in industry.

Nickel is resistant to corrosion by fluorine, alkalies, and a variety of organic materials. It remains bright on indoor exposure but tarnishes outdoors, although its corrosion rate is very low. Its low corrosion rate, coupled with its resistance to corrosion by sodium chloride and other chlorides used on roads during the winter, makes it essential as an undercoat on chromium-plated automotive trim (see ELECTROPLATING). About one-half of the nickel used

for electroplating in the United States is used in this way. Nickel containing some phosphorus can be deposited by chemical reduction; this "electroless nickel" is finding new uses.

Heavy nickel plating is employed as a lining for tank cars and as a coating for the inner walls of large pipes and similar equipment in the chemical industry.

Nickel is essential as the base for oxide-coated cathodes used in all television tubes and all but the largest radio power tubes. Alloyed with about 2 percent tungsten plus a trace of magnesium, nickel is used as the cathode base in amplifiers for submarine cables that are expected to function for 20 years without attention.

Nickel also is an essential component of white-gold alloys widely used for jewelry. These alloys also contain nickel, copper, and zinc, all of high purity.

Monel

The addition of copper to nickel provides a series of useful alloys. Monel metal, 67 percent nickel and the balance essentially copper, is stronger than nickel and has broad corrosion-resisting applications. Extremely resistant to rapidly flowing seawater, it has many marine uses. The addition of a small percentage of aluminum and titanium renders it precipitation hardenable; this high-strength version is widely used for propeller shafts. Increasing copper to 55 percent produces an electrical resistance alloy known as Constantan, which is used as a thermocouple in conjunction with pure copper.

The 30 percent and 10 percent nickel–copper alloys, usually containing 0.5 percent and 1.5 percent iron, are widely used in the form of tubes for heat interchangers and condensers. Their resistance to seawater corrosion makes them important in desalination plants. Copper-base alloys containing a small percentage of nickel become precipitation hardenable if 5–8 percent of tin or a smaller amount of silicon or phosphorus is added. These have special uses.

The ancient Chinese alloy *pai-t'ung*, now known as nickel-silver, contains 10–30 percent nickel with the balance copper plus zinc. This alloy continues as a favoured base for silver-plated ware. It is also used as a spring material for relays and has numerous other applications.

An alloy of 25 percent nickel and 75 percent copper is essentially white in colour and was adopted for coinage by Belgium in 1860 and by the U.S. five years later. More recently it has been employed in the U.S. as the outer layer of copper-centred 10- and 25-cent coins. Pure nickel was adopted by the Swiss for coinage in 1881; this use has spread to many other countries.

The fact that nickel changes in length as it is magnetized makes it useful as an ultrasonic transducer in various underwater defense devices. Alloying nickel with about 21 percent of iron has a spectacular effect in producing alloys with extraordinarily high magnetic permeability in weak fields. This type of alloy, known as Permalloy, discovered at Bell Telephone Laboratories in 1916, has had a great value in long-distance telephone transmission, including undersea cables. Other alloys of about 45–50 percent nickel, and the balance iron, have been developed for magnetic uses at higher field strengths.

Invar, an alloy containing 36 percent nickel, with the balance iron, is notable for its extremely small thermal expansion. Discovered in 1898, it has, along with later developed nickel alloys, many applications ranging from thermostats to balance wheels for watches, metal-to-glass seals essential to electric lights, and radio tubes.

A remarkable group of nickel-containing permanent-magnet alloys was developed, beginning in Japan in the early 1930s. An early example contained 25 percent nickel, 12 percent aluminum, and the balance iron. More powerful versions, such as Alnico V (containing 8 percent aluminum, 14 percent nickel, 24 percent cobalt, 3 percent copper, balance iron), developed in The Netherlands, were heat-treated in a magnetic field. These materials had a profound effect on the design of many electrical devices, including magnetic separators, dc motors, and automobile generators.

The first major market for nickel was in the production of nickel and nickel–chromium steels for armour plate, an application based on the work of James Riley of Glasgow in 1889 and tests by the U.S. Navy in 1891 on armour plate from a French steel producer. Military demands supported the industry for many years, but with the development of steam-turbine power plants, the automobile, agricultural machines, and aircraft, a whole new group of high-strength steels containing from 0.5 to about 5 percent nickel along with chromium, molybdenum, etc., were developed. More recently, with a demand for steels for ultra-low-temperature use with liquefied gases, steel of 9 percent nickel and alloys of higher nickel content have come into demand. These steels rely on carbon for hardening by heat treatment. The nickel toughens the steel and slows the hardening process so that larger sections can be heat-treated. A carbon-free iron alloy known as maraging steel has been developed. It contains 18 percent nickel, plus cobalt, titanium, and molybdenum. This alloy can be heat-treated to provide a tensile strength of some 300,000 pounds per square inch, coupled with an elongation of 5 percent to 10 percent.

High-strength steels

Nickel is resistant to oxidation at high temperatures and to electrical erosion. For these reasons, alloys that are high in nickel, such as the 4 percent manganese alloy, are used for spark plug electrodes in automobiles and for other types of ignitors. The addition of 15–20 percent of chromium to nickel vastly improves oxidation resistance so that such alloys are used for electric resistance heaters. Alloys containing 15–20 percent chromium, plus various amounts of iron, and an alloy containing 35 percent chromium, 20 percent nickel, with the balance iron, find extensive industrial use where high strength and corrosion resistance over a wide range of temperatures are essential. The addition of small amounts of aluminum and titanium permits the alloy that is high in nickel and chromium to be further strengthened by precipitation treatment. Alloys of this general type made the jet-aircraft engine possible. Gas turbines require the same alloys and are growing in importance for industrial power uses.

A large group of alloy steels ranging from 18 percent chromium, 8 percent nickel, and 25 percent chromium, 20 percent nickel, to 20 percent chromium, 35–40 percent nickel are employed where corrosion resistance is a major requirement. The stainless steels, of which the 18 percent chromium–8 percent nickel variety is the best known, are widely used where stain and corrosion resistance must be coupled with high strength. The largest single use of nickel is in the production of stainless steel.

Important compounds. Nickel sulfate hexahydrate, $NiSO_4 \cdot 6H_2O$, is employed in the electrolytic refining of nickel as well as in most nickel electroplating baths. Nickel chloride hexahydrate, $NiCl_2 \cdot 6H_2O$, is often used in conjunction with the sulfate in plating baths; while the nickel sulfamate, $Ni(SO_3NH_2) \cdot 4H_2O$, and the nickel fluoborate, $Ni(BF_4)_2$, are employed in some of the newer types of electroplating baths.

Nickel dimethylglyoxime is an insoluble salt useful in analytical chemistry in precipitating nickel. Nickel carbonyl, $Ni(CO)_4$, a liquid at room temperature, is employed in a carbonyl nickel refining process. Like all other carbonyls, it is poisonous. Nickel sub-sulfide, Ni_3S_2, is the nickel component of matte involved in pyrometallurgy. Nickel oxide, NiO, is involved in refining processes and also may be an end product.

Pentlandite, $(Ni,Fe)_9S_8$, is the most important sulfide mineral; but pyrrhotite (FeS to Fe_7S_8) with some nickel replacing iron also is important in nickel and iron-recovery processes. Millerite, NiS, is a rare mineral. In the laterites, garnierite, $(NiMg)_6Si_4O_{10}(OH)_8$, is the richest in nickel; but nickeliferous limonite, $(Fe,Ni)O(OH) \cdot nH_2O$, constitutes a major portion of the laterites.

BIBLIOGRAPHY. The development of the world nickel industry is described by J.F. THOMPSON and N. BEASLEY in *For the Years to Come* (1960). A summary of present world sources and a description of the processes employed for recovery may be found in J.R. BOLDT, JR., and P. QUENEAU, *The Winning of Nickel* (1967). See also E.J. PRYOR, *Mineral Processing* (1965), for an excellent discussion. The properties of nickel and many of its alloys are discussed in the *Metals Handbook*, 8th ed., vol. 1 (1961); and in *Monograph 106* of the NATIONAL BUREAU OF STANDARDS (1968). The corrosion resistance of nickel and its alloys is summarized by H.H. UHLIG,

Corrosion Handbook (1948). The development of nickel steels from the earliest time to the present is covered by R.B.G. YEO and O.O. MILLER in *The Sorby Centennial Symposium on the History of Metallurgy* (1965). Nickel plating was an early use of nickel that has grown substantially over the years. Much practical information on this aspect may be found in a publication of the INTERNATIONAL NICKEL CO., *Nickel Plating Processes and Properties* (1967). A summary of the use of nickel in various industrial fields is given in *Applications of Nickel*, MAB 248, Metallurgical Advisory Board, National Academy of Sciences (1968). World production statistics plus other information on developments may be found in the *Annual Reports of the U.S. Bureau of Mines.*

(E.M.W.)

Niebuhr, Reinhold

Reinhold Niebuhr was one of the most influential American theologians of the 20th century. His special contribution was his stress on the issues on the borderline between theology and social ethics. His criticism of the prevailing theological liberalism of the 1920s significantly changed the intellectual climate within the Protestant churches in the United States, and his enormous influence on political thought, both inside and outside the church, caused Hans J. Morgenthau, an eminent political scientist, to say that Niebuhr was "the greatest living political philosopher of America." He was probably the most popular preacher in university chapels from the early 1920s to the early 1950s. Many sophisticated Christians today trace their conviction that Christianity makes sense to the influence of his preaching. He was not a specialized scholar in the technical sense in any field, including theology, but his broad learning and his original and incisive thought made him the subject of many theses and other scholarly writings, and he exercised a seminal influence on scholarship and thought in a variety of fields.

By courtesy of the Rare Book Department, Union Theological Seminary Library, New York

Niebuhr, 1963.

Niebuhr was born in Wright City, Missouri, on June 21, 1892. He was the son of Gustav and Lydia Niebuhr, who had emigrated to the United States at an early age from Germany. Gustav Niebuhr was a minister of the Evangelical Synod of North America, a denomination with a Lutheran and Reformed German background that merged into the Evangelical and Reformed Church in 1934 and is now a part of the United Church of Christ. At an early age Reinhold Niebuhr decided to emulate his father and become a minister. He graduated from his denomination's Elmhurst College, Illinois (1910), and Eden Theological Seminary, St. Louis, Missouri (1913), and completed his theological education at Yale, receiving a bachelor of divinity degree (1914) and a master of arts (1915). He was ordained to the ministry of the Evangelical Synod in 1915.

Niebuhr served as pastor of Bethel Evangelical Church in Detroit from 1915 to 1928. This experience—and especially his exposure to American industrialism, specifically, the automobile industry, before labour was protected by unions and by social legislation—caused him to become a radical critic of capitalism and an advocate of Socialism. His *Leaves from the Notebook of a Tamed Cynic* (1929), an account of his years in Detroit, reveals his spirit and bent of mind, which continued to characterize him throughout his life. He never lost the pastoral concern for individuals that is so well expressed in this book. Though he was formidable as a polemicist in his writings and in his public life, his personal relations were warm and friendly. This was especially true of his relations with his students. Even people who were targets of his polemics on issues of theology or of ultimate philosophical outlook were often his coworkers in many a social or political issue.

Years as a pastor

He left the pastoral ministry in 1928 to teach at Union Theological Seminary in New York City, where he was a great intellectual and personal force until his retirement in 1960. As a theologian he is best known for his "Christian Realism," which emphasized the persistent roots of evil in human life. In his *Moral Man and Immoral Society* (1932) he stressed the egoism, the pride and hypocrisy of nations and classes. Later he saw these as ultimately the fruit of the insecurity and anxious overdefensiveness of man in his finiteness; here he located "original sin." He emphasized the tendency for sin—in the form of destructive pride—to appear on every level of human achievement, especially where claims to perfection were made, either in religious or political terms. His powerful polemics against liberal beliefs in assured progress and radical utopian hopes have caused a neglect of his more hopeful teaching concerning the image of God in all men that is never completely destroyed by sin and concerning "common grace" that is not dependent on recognized Christian redemption in personal or collective life. Also, he was himself a hopeful political activist and emphasized the good that could be achieved if pretentions were overcome. His outlook is well expressed in his statement that "the saints are tempted to continue to see that grace may abound, while sinners toil and sweat to make human relations a little more tolerable and slightly more just." He always had faith in what he called "indeterminate possibilities" for humanity in history as long as men did not deceive themselves into thinking that absolute solutions of historical problems were in their control. Though he did much to encourage the revival of the theology of the Reformation, with its emphasis on sin and grace—so-called Neo-orthodoxy—his salient theological work, *The Nature and Destiny of Man*, was planned by him as a synthesis both of the insights of the Reformation and of the Renaissance, with its hopefulness about cultural achievements.

His distance from the strongly Christocentric forms of Protestant Neo-orthodoxy can be seen in his unusual attitude toward the Jewish community. He was perhaps the first Christian theologian with ecumenical influence who developed a view of the relations between Christianity and Judaism that made it inappropriate for Christians to seek to convert Jews to their faith.

Political activities

His early political activities were influenced by his Socialist convictions (he was a founder of the Fellowship of Socialist Christians) and he ran for office several times on the Socialist ticket. In the 1930s he broke with the Socialist Party over its pacifist or noninterventionist attitude in foreign policy, and in the 1940s he became a left-wing, anti-Communist Democrat. He was a founder and for a time chairman of the Americans for Democratic Action and he was vice chairman of the Liberal Party in the state of New York. In the 1930s he was much influenced by Marxist theory, but he rejected Marxist absolutism and both the tactics of Communists in the United States and Stalinism in the Soviet Union.

He did much to persuade Christians influenced by pacifism to support the war against Hitler. He himself had been a pacifist as a result of his revulsion against World War I, but during the 1930s he became the strongest theological opponent of any form of pacifism that claimed to have universally applicable nonviolent solu-

tions of political problems. Identifying himself with the resistance to Hitler within Germany, he opposed a vindictive peace. After World War II, he had considerable influence with the policy planners in the State Department. He was a strong supporter of the Cold War resistance to Russian political expansion in Europe. His political activity ended during the early stage of the Cold War, but his later thought showed his capacity to transcend the outlook of that period. His book *The Irony of American History* (1952), while justifying American anti-Communist policies, gave much attention to criticism of American messianism and the American tendency to engage in self-righteous crusades. He always attacked American claims to special virtue. Early he favoured the recognition by the United States of mainland China, and he was an early opponent of the Vietnam War. He regarded as an error attempts to impose U.S. solutions on the new countries that emerged after World War II.

Niebuhr was an editor of *The World Tomorrow*, a religious pacifist and Socialist journal; *Christianity and Crisis*, a biweekly with wide-ranging social and religious concerns; and a quarterly, now discontinued, first named *Radical Religion* and later *Christianity and Society*.

He married Ursula M. Keppel-Compton in 1931. His wife was herself a teacher of religion at Barnard College in New York City, and they worked closely together.

After 1952 his public activities were seriously limited as the result of a stroke, but he was able to continue much of his teaching and writing. He died on June 1, 1971, at Stockbridge, Massachusetts.

BIBLIOGRAPHY. CHARLES W. KEGLEY and ROBERT W. BRETALL (eds.), *Reinhold Niebuhr: His Religious, Social, and Political Thought* (1956), includes an intellectual autobiography, interpretive essays with Niebuhr's response, and a comprehensive bibliography. More recent is RONALD H. and JOANN M. STONE (comps.), "Writings of Reinhold Niebuhr, 1953–1971," *Union Seminary Quarterly Review*, 27:9–29 (1971). JUNE BINGHAM, *Courage to Change: An Introduction to the Life and Thought of Reinhold Niebuhr* (1961), is an Impressionistic biography based on careful research. GORDON HARLAND, *The Thought of Reinhold Niebuhr* (1960), is one of the best overall expositions with theological emphasis.

Writings of Niebuhr (in addition to works mentioned above) include: *Faith and History: A Comparison of Christian and Modern Views of History* (1949), theological orientation; *The Self and the Dramas of History* (1955), probably his profoundest philosophical work; and *The Structure of Nations and Empires* (1959), his chief systematic discussion of international relations. Four volumes of essays, some of which are essential for understanding Niebuhr's thought and his influence on events, are *Christianity and Power Politics* (1940); *Christian Realism and Political Problems* (1953); *Pious and Secular America* (1958); and *Faith and Politics: A Commentary on Religious, Social, and Political Thought in a Technological Age*, ed. by RONALD H. STONE (1968). *Love and Justice*, ed. by D.B. ROBERTSON (1957), is a collection of shorter writings showing Niebuhr's response to events; *Children of Light and Children of Darkness: A Vindication of Democracy and a Critique of Its Traditional Defence* (1944), a brief but comprehensive discussion of social ethics.

(J.C.Be.)

Niedersachsen

Niedersachsen, or Lower Saxony, is one of the more important of the *Länder*, or states, of the Federal Republic of Germany. The second largest in size, with an area of 18,304 square miles (47,408 square kilometres), it occupies an important band of territory across the northern part of the nation, stretching from The Netherlands border and the coast of the North Sea to the frontier, with the German Democratic Republic in the east. The neck of land occupied by Schleswig-Holstein, Hamburg, and, farther north, Denmark borders it on the north, while its neighbours to the south are the *Länder* of Nordrhein-Westfalen, containing the industrial zone of the Ruhr, and Hessen. By the early 1970s, its population exceeded 7,000,000, making the *Land*, with about 12 percent of the total population of West Germany, fourth in rank. There was, by contrast, a low population density of some 386 persons per square mile, due mainly to the low level of industrial development. At the end of the 1960s, the gross national product per capita was below the average for West Germany. This economic retardation was to be the key to Niedersachsen's situation in the 1970s.

The landscape. With the exception of a small highland area to the south, the landscapes of the state are dominated by the great north German plain. In the northwest, the Ostfriesische Inseln—12 islands in the North Sea—and about 325 square miles of coastal land are actually below sea level, protected from inundation by dikes similar to those in the nearby Netherlands. The highland area, which lies south of the important Mittelland Kanal (Midland Canal) running across the state, contains the Weser, Deister, and the rugged Harz Mountains. More than half of Niedersachsen is drained by the Weser and its tributaries, the Fulda and the Werra, although the major settlement of Bremerhaven (at the mouth of the Weser) and Bremen itself (40 miles up the river) form a separate political entity, the Freie Hansestadt Bremen, smallest of the West German *Länder*. The territory also contains two large lakes: Steinhuder Meer (12 square miles) and Dümmer (6 square miles). At the mouths of the rivers flowing into the North Sea fertile marshes are found, mostly supporting a pasture economy. In the northeastern region there is a less fertile area of land partly covered with forests. This contains Lüneburger Heide (heath), noted for its old-fashioned red farmhouses and the ancient megalithic structures known as "graves of giants." It is now a celebrated nature preserve.

Regional divisions

The troughlike valleys of the forested southern uplands provide good-quality land, as do their foothills farther north. The latter form part of the treeless belt of rich, often windblown, soils known as the Börde, which runs in a narrow east–west zone across the state. In addition to supporting an arable farming population this area, situated on the boundary between plain and upland, became a historical nucleus for the growth of a string of small towns. Niedersachsen's climate offers mild winters, moderately warm summers, and a steady year-round rainfall ranging from 24 to 35 inches (600 to 900 millimetres).

Niedersachsen was established on November 1, 1946, by the British military government, which merged the former Prussian province of Hanover with the states of Braunschweig, Oldenburg, and Schaumburg-Lippe. A quarter of a century later, nearly one-fifth of its population was living in large cities (defined as those with more than 100,000 inhabitants), more than a quarter in places with 10,000 to 100,000 inhabitants, and more than half in communes of fewer than 10,000 inhabitants.

Ethnic origins

The people. The population of Niedersachsen regards itself as lower German, linked by a common ancient Saxon origin and use of the lower German language known as Plattdeutsch. The latter, a dialect closely related to Dutch, Frisian, and English, is quite distinct from the official High German. Some regional literature is still produced in this form, and it remains the language of the home in much of the state. This feeling of cultural unity helps to bind together such diverse areas as the parts of ancient Hanover east of the Weser, the younger regions of Braunschweig, Emsland, Osnabrück, and South Oldenburg (which were formerly under Westphalian influence), and the Frisian portions of northern Oldenburg and Ostfriesland. The northern European character of social life is further illustrated by the adherence of about 77 percent of the population to Protestantism and 19 percent to Roman Catholicism.

In 1939 the population of Niedersachsen as presently defined amounted to 4,500,000. By 1946 the influx of refugees from other areas of war-torn Europe had caused an increase to 6,200,000, and this in spite of war losses. By 1950 the population had reached 6,744,000. During the 1950s over 340,000 refugees were transferred to other states of the Federal Republic of Germany that were able to offer better living conditions. By the middle of 1968 the population passed the 7,000,000 mark. This growth was mainly caused by natural increase and, to a certain degree, by immigration. By the early 1970s, the average birthrate was about 18 per 1,000, the average death rate 12 per 1,000, the rate of increase about 6 per 1,000. Niedersachsen entered the 1970s with seven cities of over 100,000

population: Hannover, 518,000; Braunschweig, 225,000; Osnabrück, 141,000; Oldenburg, 131,000; Salzgitter, 118,000; Göttingen, 114,000; and Wilhelmshaven, 103,000.

The economy. Agriculture, the traditional mainstay of the local economy, is still important. By the 1970s about 61.5 percent of the area of the state was used for agriculture (grain land, 24.1 percent; other arable land, 8.2 percent; permanent grassland, 27.1 percent; other agriculturally utilized land, 2.1 percent). Some 20 percent of the land was covered by forests, 2.5 percent consisted of uncultivated moors, almost 4 percent was wasteland, 2 percent water surface, and about 10 percent used for roads and buildings. Industrial turnover in 1968 exceeded 46,000,000,000 Deutsche Marks ($11,500,000,000).

Industrial Employment in Niedersachsen
(1968)

industry	percent of total workers
Vehicles	15.9
Machines	11.0
Electrical goods	9.5
Investment goods (other than those mentioned above)	9.3
Raw materials, industrial intermediates*	20.5
Consumer goods†	20.7
Food, beverages, tobacco	9.5
Mining‡	3.6
Total	100.0

*Chemicals, rubber and asbestos manufactures, steel. †Textiles, clothing, furniture, etc. ‡Iron ore (4,200,000 tons), potash (1,100,000 tons), petroleum (6,400,000 tons), natural gas, and peat.

Hannover Industrial Fair

Once a year Hannover, the state capital, becomes the meeting point of the industrial world, when the German Industries Fair, the largest of its kind all over the world, is held there. In the 1970s it was attracting more than 600,000 visitors a year from over 100 countries and 5,500 exhibitors, among whom were 1,100 from 33 foreign countries, including eastern Europe and the German Democratic Republic.

Transportation. Niedersachsen is well provided with transport facilities. By 1970, 1,750,000 road vehicles in the state could travel on a network made up of autobahns, federal highways, state highways, and numerous district and commune roads. In addition, the railway network transported over 3,000,000 tons of goods annually. The state's importance in the regional economy of West Germany was enhanced by such inland waterways as the Mittelland Kanal, the Dortmund–Ems Kanal, and a host of others, facilitated by the flat landscapes. In addition, the major rivers, notably the Weser and the Elbe, were navigable for considerable distances.

More than 40,000,000 tons of goods pass through the ports of Wilhelmshaven, Emden, Nordenham, and Brake each year, an indication of Niedersachsen's importance in regional and world trade.

The state's main airport of Hannover-Langenhagen handles some 2,000,000 passengers annually, together with 15,750 tons of freight and mail. The scenic beauties of Lüneburger Heide and the southern uplands, and also the seaside resorts of the northwest, attract a considerable tourist traffic, amounting by the 1970s to almost 20,000,000 overnight stays, all but 5 percent of which were by German citizens. The nearly 6,000 hotels included 7.6 percent in large towns, 32.5 percent in watering places and climatic health resorts, and 33.2 percent in seaside resorts. The latter proved attractive to inhabitants of the great industrial regions bordering the state to the south.

Administration and social conditions. The governmental structure is comprised of a prime minister, state chancellory, and eight ministries. Regional governmental divisions include the six *Regierungsbezirke* (Hannover, Hildesheim, Lüneburg, Stade, Osnabrück, and Aurich) and two *Verwaltungsbezirke* (Oldenburg and Braunschweig). According to an administrative and regional reform initiated in 1965, it was anticipated that the 4,000 or so small communes would shrink to about a quarter of that number of larger, more functional units. Niedersachsen entered the 1970s with its two main political parties evenly balanced: results of Landtag elections held in June 1970 showed the Social Democrats (left-wing) had 46.2 percent of the vote, the Christian Democrats (centre right) 45.7 percent, and the Free Democrats (partner of the Social Democrats in the federal government) and the National Democrats each with less than the 5 percent necessary to enter the state parliament. Justice is administered by means of a constitutional court, three courts of appeal, 11 regional courts, and 132 local courts. The armed forces in this strategically important state, together with British, Dutch, and Belgian forces, participate in national defense through the framework of NATO. By the 1970s, there were some 700,000 pupils at primary schools, 100,000 pupils at middle schools (*Mittelschulen*), and 120,000 pupils at secondary schools, with a further 25,000 pupils being educated in special schools.

Educational facilities

University education was offered by the University of Göttingen, with an enrollment of 10,000 students; the seven institutes of the Max-Planck-Gesellschaft (Society for the Advancement of Science); the technical universities of Hannover and Braunschweig (5,000 enrollment each); and a number of smaller institutes. The forested landscapes of the state have given rise to no fewer than 12 research institutes for forestry attached to the University of Göttingen. The second oldest German Academy of Science (founded in 1751) is located in Göttingen, where it flourishes with 22 separate committees.

Cultural life and institutions. In common with other West German *Länder* and indeed with many other parts of Europe, Niedersachsen has a thriving and well-subsidized cultural life. There are state theatres at Hannover, Oldenburg, and Braunschweig. Hannover, the cultural as well as the state capital, can boast of three other theatres, among them the Landesbühne, which gives performances in more than 40 towns in the region. Other notable theatres are, in Wilhelmshaven, the Landesbühne Niedersachsen-Nord; in Göttingen, the Deutsches Theater; in Hildesheim, the Stadttheater; and in Celle, the Schlosstheater, whose plays are performed in a fine baroque building dating from 1674. In addition, nearly 500 cinemas (with 16,000,000 patrons annually) cater to a more popular audience. The *Hannoversche Allgemeine Zeitung* and *Hannoversche Presse* are the leading state newspapers, each with a circulation of more than 150,000. Hannover is also the centre of the agricultural and forestry press. A famous cultural periodical, the *Westermanns Monatshefte*, is edited from Braunschweig. Radio and television are broadcast by Norddeutscher Rundfunk (NDR), based in Hamburg but with studios at Hannover and Oldenburg. There are many facilities for sports, both indoor and outdoor, and there are nearly 1,000,000 members of sports clubs—one in seven of the entire population.

Problems and prospects. In the years to come the industrial basis of Niedersachsen will be considerably expanded, bringing a fresh flow of capital investment and, it is hoped, a more diversified and prosperous regional economy. Plans were underway in the early 1970s for an aluminum plant at Stade and another chemical complex at Wilhelmshaven. Nuclear power plants were planned or under construction at nearby Würgassen (upper Weser), at Stade (lower Elbe), and at Esenshamm (lower Weser), and the once backward state appeared in the forefront of a national economic development.

BIBLIOGRAPHY. For current information on this subject, see *Statistisches Jahrbuch für Niedersachsen*, a statistical handbook published annually.

(W.G./G.Dn.)

Nietzsche, Friedrich

Friedrich Nietzsche, a 19th-century German philosopher, was one of the most influential thinkers of modern times. He was an impassioned critic of the culture and ethos of his time, and particularly of Christianity, conformism, nationalism, and resentment. Almost all 20th-century German philosophers, as well as the greatest German poets, novelists, and psychologists, are profoundly in-

debted to him. In other countries, too, and most especially in France, he ranks among the most influential philosophers since Kant and Hegel.

Louis Held—Deutsche Fotothek, Dresden.

Nietzsche, 1888.

Early life and studies. Nietzsche was born October 15, 1844, at Röcken, in the Prussian province of Saxony, the son and grandson of Lutheran ministers. His father christened him Friedrich Wilhelm, after the reigning king of Prussia. (Nietzsche later dropped the Wilhelm.) The King became mad a few years later, as did Nietzsche's father; and in January 1889 Nietzsche himself became insane. The father's illness was diagnosed as "softening of the brain," and after his death in 1849 his skull was opened and the diagnosis confirmed. But there is some question about the meaning of softening of the brain; most experts agree that the philosopher's eventual insanity was not inherited.

In January 1850 Nietzsche's mother moved her family to Naumburg (on the Saale). The household there included Nietzsche's younger sister Elizabeth, his maternal grandmother, and two maiden aunts. This may help to account for some of Nietzsche's snide remarks about women. In 1858 Nietzsche was admitted to one of the most distinguished boarding schools in Germany, Pforta, near Naumburg. He was frequently at the head of his class, received a superb classical education, and graduated in the year 1864.

Student of classical philology

He went to the University of Bonn to study theology and classical philology but gave up theology the following year and followed Friedrich Ritschl, a professor of classics, to Leipzig. Ritschl, founder of a famous school of classical scholarship, who had never before included the work of a graduate student in his scholarly journal, published some of Nietzsche's papers there. One paper won a prize, but Nietzsche wrote his friend Erwin Rohde, later an eminent classicist too, that he himself found the paper "repulsive." By now he was more interested in philology than in philology.

He may have contracted syphilis while he was a student, which might account for his later madness. It seems certain that during his adult life he was, sexually, an ascetic. The most that has been claimed is that as a student he may have visited a brothel twice. In October 1867 he commenced his military service, but in March he hurt himself seriously when his chest hit the pommel of the saddle as he jumped on his horse. He rode on as if nothing had happened but then was hospitalized and suffered for a long time.

Back at Leipzig, he wrote Rohde on November 20, 1868, about his disgust with philology—"the whole molish business, the full cheek pouches and blind eyes, the delight at having caught a worm, and indifference toward the true and urgent problems of life." He still published some reviews of scholarly books, but at one point thought of writing his dissertation on Immanual Kant's philosophy; and early in 1869 he considered switching to chemistry, "throwing philology where it belongs, with the household rubbish of our ancient ancestors."

At that point the University of Basel (Switzerland) was looking for a professor of classical philology. Nietzsche had not yet written his doctor's thesis, much less the additional dissertation that a German Ph.D. must have accepted before lecturing at a university (after which he generally writes a book before obtaining a professorship). But Ritschl wrote Basel that in almost 40 years of teaching he had never seen a student like Nietzsche. When Nietzsche was offered the chair, Ritschl assured Basel that, although most of Nietzsche's work had been in Greek literature and philosophy, "with his high gifts he will work in other fields with great success. He will simply be able to do anything he wants to do." Without any dissertation or examination, he got his doctorate, and in April 1869 he went to Basel and became a Swiss subject.

When the Franco-Prussian War broke out in 1870, Nietzsche obtained a leave to serve as a medical orderly with the Prussian army. After a month, he returned to Germany with dysentery and diphtheria. Although his health was shattered, he returned to Basel in October to resume a very heavy teaching load. Whenever possible, he visited Richard Wagner, the great operatic composer, in Tribschen, near Lucerne, Switzerland. Wagner was then in voluntary exile from Germany, where his work was not yet widely accepted; he shared Nietzsche's admiration for the philosophy of Schopenhauer, and he appreciated Nietzsche as a professorial apostle of his music. For Nietzsche this was his first encounter with a major artist; and he could do, if only briefly, with a father figure (Wagner was the same age as his father). The basic conflict between the two was to emerge later.

First works. His first book, *Die Geburt der Tragödie* (1872; *The Birth of Tragedy*, 1968), eschewed all the customary trappings of the classical scholar. He dispensed with footnotes and Greek quotations and made no show whatever of his learning. Nietzsche acknowledged the Greeks' genius of measure, restraint, and harmony, which he called "the Apollinian"; but he argued that one must not ignore "the Dionysian"—the irrational passions that had to be harnessed to make possible Greek literature and art. His thesis was that tragedy was born out of the fusion of the two, and was killed by rationalism. He proposed his own ideas about the birth and death of Greek tragedy in a mere 85 pages and ended with a considerable addition on the "rebirth of tragedy" from the spirit of Wagner's music. A young philologist, Ulrich von Wilamowitz-Möllendorff, who was later to become a distinguished classicist, seized this opportunity to demonstrate his erudition by publishing a pamphlet in which he tried to demolish the book and the reputation of its author. It occasioned a heated controversy with Nietzsche's friend Rohde. By 1912 F.M. Cornford, a leading British classicist, was to call *The Birth of Tragedy* "a work of profound imaginative insight, which left the scholarship of a generation toiling in the rear." Until World War II, it was Nietzsche's ideas about the birth of tragedy that were most influential; since then his ideas about the death of tragedy have been widely echoed. Despite trenchant criticisms of the work's thesis by present-day scholars, however, Nietzsche's first book retains considerable importance. It brought to an end the era in which Greek culture could be summed up as "sweetness and light."

Thesis of *The Birth of Tragedy*

In his later work Nietzsche extols Dionysus, by then not in contrast with Apollo but as standing for the synthesis of the Apollinian and the Dionysian, the sublimation of passion, and the creative affirmation of this world and this life. In the later writings, Dionysus is opposed to Christ, to otherworldliness, and to the extirpation (not sublimation) of the passions.

In *The Birth of Tragedy* Nietzsche emancipated himself from Arthur Schopenhauer's "Buddhistic negation of the will," as he called it. He had admired Schopenhauer, an influential 19th-century German philosopher, for his frank recognition of the suffering in this world, but in Greek tragedy Nietzsche found that it was possible to face up to

the horrors of existence and to affirm life, nevertheless, as beautiful in spite of everything. He argued that tragedy had died of the spirit of rationalism and optimism that was embodied in Socrates, but near the end of the book he envisaged the possibility of "an artistic Socrates": a philosopher with a passion for poetry and music as well as an intellectual conscience. This was clearly his conception of his own mission, and it remains a splendid image of his spirit. Nietzsche followed up his first book with four *Unzeitgemässe Betrachtungen* (1873–76; *Thoughts Out of Season*, 1909) published one at a time. They are fine essays, but Nietzsche's fame rests squarely on his later works. In 1878 he published the first of five collections of aphorisms. He called it *Menschliches, Allzumenschliches* (1878; *Human All-too-Human*, 2 vol., 1909–11) and dedicated it to the memory of Voltaire on the 100th anniversary of his death. Nietzsche's emerging break with Wagner was sealed shortly thereafter when Wagner received this deeply antiromantic book of aphorisms. In August Wagner attacked Nietzsche in the *Bayreuther Blätter*. The breach had been inevitable. As long as Wagner was a lonely genius in exile, his hatred of the French and the Jews did not matter that much to Nietzsche, who admired these peoples; but by 1878 Wagner had moved to Bayreuth, made his peace with the new German empire, and become a major influence. That Nietzsche also considered Wagner's *Parsifal*, of which he received an inscribed copy, an insincere obeisance to Christianity was a relatively minor matter.

The break with Wagner

The great creative decade (1879–89). In 1879 Nietzsche resigned from the university, pleading his poor health, and was granted a modest pension. His doctors advised him to do as little reading and writing as possible to save his poor eyes. But during the next ten years he devoted himself entirely to his writing, living very frugally and driving himself relentlessly. He published a book every year, and each of them represented a triumph over his half-blind eyes, terrible migraine headaches, and manifold physical agonies. His major works were written in utter solitude in Sils Maria (Switzerland), in Nice and Mentone on the Riviera, and in Rome and Turin in Italy. Until 1888 they received virtually no attention either in the press or from scholars. Without any response, Nietzsche kept writing.

Of his former colleagues at Basel, only Franz Overbeck, a church historian but an unbeliever, remained utterly loyal to him until the end. An unsuccessful young composer, Heinrich Köselitz—Nietzsche called him Peter Gast—worshipped Nietzsche and helped him occasionally with proofreading and other chores. Nietzsche's relationship with two other friends was much more troubled. He first met Paul Rée in 1873. Two years later Rée received his doctorate in philosophy and published *Psychologische Beobachtungen* ("Psychological Observations"), a work that Nietzsche liked. They exchanged letters, and a genuine friendship developed. Rée had a light touch and worked on some of the same problems that concerned Nietzsche. He inscribed a copy of his *Der Ursprung der moralischen Empfindungen* (1877; "The Origin of Moral Feelings") to Nietzsche: "To the father of this essay, most gratefully from its mother."

Friendship with Lou Andreas-Salomé

In 1882 he wrote Nietzsche about a young woman, Lou Salomé (1861–1937). Nietzsche met her, was greatly impressed, and the friendship became three-cornered. Gradually and subtly, Rée's feelings for Lou undermined his relationship to Nietzsche. But the jealousy of Nietzsche's sister, Elizabeth, was uncompromising. She led Nietzsche to believe that his friends had betrayed him, speaking evil about him behind his back; and by the end of that year Nietzsche felt lonelier and more forsaken than ever. After Rée's death (1901), Lou, by then a well-known writer, claimed that both men had proposed marriage to her and that Nietzsche asked Rée to transmit his proposal. Others embellished the story by adding that, unknown to Nietzsche, she was Rée's mistress at that time. But Rudolph Binion has argued in his biography *Frau Lou* (1968) that she remained a virgin until ten years later and that Nietzsche never proposed marriage to her, although she was apparently waiting for him to do so. After her marriage Lou became known as Lou Andreas-Salomé.

Once, Nietzsche actually had asked a woman (Mathilde Trampedach) to marry him. He did not know her well at all and evidently felt relieved when she turned him down. But the break with Lou and with Rée hurt him deeply. It was in despair and utter solitude that he began writing *Also sprach Zarathustra* (Eng. trans., *Thus Spoke Zarathustra,* 1954), a rhapsodic and aphoristic attempt to present his thought as a whole. The first three parts appeared separately (1883–84) and met with no response at all. Of the fourth part Nietzsche had only 40 copies printed privately and then distributed seven among friends. The first public edition appeared in 1892. After that, *Zarathustra* became Nietzsche's most popular book, widely read but rarely understood. It is now generally considered one of the masterpieces of world literature, daring in form and full of ideas—especially, but by no means only, about morals and psychology.

In *Jenseits von Gut und Böse* (1886; *Beyond Good and Evil*, 1968) and *Zur Genealogie der Moral* (1887; On *the Genealogy of Morals*, 1968) Nietzsche tried to explain his ideas in more explicit prose. In form and manner these books come closer to British and American philosophy than Nietzsche's other works, which are aphoristic or lyrical. But no English-speaking philosopher has ever subjected the faith and morals of Western man to such merciless questioning.

In 1888 Nietzsche published *Der Fall Wagner* (1888; *The Case of Wagner*, 1968), an often very funny book of roughly 50 pages that says some very serious things about culture and decadence. That year Georg Brandes, a Danish Jew who was a literary critic and scholar, began to lecture on Nietzsche at the University of Copenhagen; this was the first significant public notice of Nietzsche's works and thought. Nietzsche completed four more books in 1888. In *Götzen-Dämmerung* (1889; *Twilight of the Idols*, 1954) he tried to summarize his philosophy in a mere 100 pages. In *Der Antichrist* (1895; *The Antichrist*, 1954) he presented a single-minded attack on Christianity. In *Ecce Homo* (1908; Eng. trans. 1968) he reflected on his own significance and, one by one, on his books. In *Nietzsche contra Wagner* (1895; Eng. trans. 1954), finally, he brought together passages on Wagner from his earlier works, making a few slight stylistic changes. This little book is perhaps his most beautifully written work.

Period of collapse. Early in January 1889, Nietzsche collapsed in the street in Turin. Carried back to his room, he sent out some mad but meaningful letters and post cards. Overbeck came to take him back to Basel. He spent the last 11½ years of his life first in an asylum, then in his mother's care in Naumburg, and finally in Weimar, where his sister took him after his mother's death (1897). He died in Weimar on August 25, 1900.

During this last period he wrote nothing and was incapable of conversation. Informed opinion favours the diagnosis of an atypical general paralysis, which would indicate tertiary syphilis. If so, the disease was dormant in him during his creative years. But his works cannot be written off as the products of a madman, as some opponents of his thought claim. The insanity that comes of syphilis hardly ever lasts 11 years or more. The breakdown in January 1889 is very clear. The books—all written before the collapse—must be judged and can afford to be judged on their merits.

Nietzsche's sister's distortion of his thought

His sister had married a prominent German anti-Semite, Bernhard Förster, in 1885, and they were in Paraguay when Nietzsche collapsed. Before her marriage she had been very attached to her brother without understanding his philosophy; and she found little to admire in it after his break with Wagner. In 1889 she was absorbed in her husband's attempt to found a Teutonic colony in South America. But the settlers felt swindled by the Försters; her husband committed suicide; and her attempts to make a national hero of him came to nothing. Meanwhile, Nietzsche's fame was spreading. Then his sister transferred all her energies to his cause and changed her name to Förster-Nietzsche. The irony of her new name was

lost on her: Nietzsche had loathed the anti-Semitic nationalism of her husband; but, as the chief apostle of her brother's philosophy, she always remained Förster first and misrepresented Nietzsche. She even stooped to forgery and published as addressed to her some drafts of letters to others. She also published a selection of some of his most interesting notes as if it were his systematic main work: *Der Wille zur Macht* (1901; *The Will to Power*, 1967). These notes merit serious study, but it is important to distinguish them from the books that Nietzsche himself polished for publication. As he remarked repeatedly about himself, few writers require to be read with so much regard for context.

Evaluations. At one time Nietzsche was widely associated by interpreters of his thought with Darwin and evolution, then with the Nazis, and more recently with the Protestant theologians who liked to cite his dictum that "God is dead." Still others see him as an Existentialist. All such perspectives are partial and misleading. It is no less legitimate to stress his relationship to Sigmund Freud and psychoanalysis or to such writers as Thomas Mann, Hermann Hesse, Rainer Maria Rilke, and André Gide.

As a writer of German prose, he remains unexcelled. His poetry has also had considerable influence. Freud paid lavish tribute to his psychology and often remarked that Nietzsche had a more thorough self-knowledge than any other man had had or ever would have. But his greatest importance was as a philosopher. In 1950 Gottfried Benn, a leading German poet, said:

> Virtually everything my generation discussed, tried to think through—one might say, suffered; one might also say, spun out—had long been expressed and exhausted by Nietzsche, who had found definitive formulations; the rest was exegesis.

MAJOR WORKS

PHILOSOPHY: *Die Geburt der Tragödie* (1872; *The Birth of Tragedy* in *Basic Writings of Nietzsche*, 1968); *Unzeitgemässe Betrachtungen*, 4 vol. (1873–76; *Thoughts Out of Season*, 2 vol., 1909) comprising: *David Strauss der Bekenner und der Schriftsteller*, *Vom Nutzen und Nachtheil der Historie für das Leben*, *Schopenhauer als Erzieher*, and *Richard Wagner in Bayreuth*; *Menschliches, Allzumenschliches* (1878; *Human All-too-Human*, 1909–11); *Vermischte Meinungen und Sprüche* (1879); *Der Wanderer und sein Schatten* (1880: the last two included in later editions of *Menschliches, Allzumenschliches* as part 2); *Morgenröte* (1881; *The Dawn of Day*, 1911); *Die fröhliche Wissenschaft* (1882); new ed. augmented by book 5 and *Lieder des Prinzen Vogelfrei* (1887; *The Joyful Wisdom*, 1910; *The Gay Science*, 1974); *Also sprach Zarathustra* pt. 1–3 (1883–84) and pt. 4 (1891; *Thus Spoke Zarathustra* in *The Portable Nietzsche*, ed. by Walter Kaufmann, 1954); *Jenseits von Gut und Böse* (1886; *Beyond Good and Evil* in *Basic Writings of Nietzsche*); *Zur Genealogie der Moral* (1887; *On the Genealogy of Morals* in *Basic Writings of Nietzsche*); *Der Fall Wagner* (1888; *The Case of Wagner* in *Basic Writings of Nietzsche*); *Götzen-Dämmerung* (1889; *Twilight of the Idols* in *The Portable Nietzsche*); *Der Antichrist* (1895; *The Antichrist* in *The Portable Nietzsche*); *Nietzsche contra Wagner* (1895; Eng. trans. in *The Portable Nietzsche*); *Der Wille zur Macht*, selections from Nietzsche's notebooks (1901; rev. and enlarged ed., 1910–11; *The Will to Power*, 1967); *Ecce Homo* (1908; Eng. trans. in *Basic Writings of Nietzsche*, trans. and ed. by Walter Kaufmann, 1968).

VERSE AND MUSIC: Nietzsche's poems—first collected in *Gedichte und Sprüche* (1898)—are included in the first two collected editions, listed below. Not so his musical compositions, which are of minor interest: *Hymnus an das Leben* (for chorus and orchestra, text by Lou Andreas-Salomé, 1887) and 16 songs with piano accompaniment, collected in the incomplete edition of his *Musikalische Werke*, vol. 1 (1924).

COLLECTED EDITIONS OF THE WRITINGS: Of the various collected editions the most satisfactory is the *Musarion-Ausgabe*, 23 vol. including 2½ vol. of indexes (1920–29). *Kröners Taschenausgabe*, 12 vol. (including index volume) that can be bought separately, is the handiest edition, but contains fewer notes and fragments. *Werke in drei Bänden*, 3 vol. (1954–56) contains fewer poems, far fewer notes and fragments, but a selection of letters. Vol. 4 contains an index. *Kritische Gesamtausgabe* of works (projected: about 30 volumes) and letters (projected: about 15 volumes), ed. by Giorgio Colli and Mazzino Montinari, began to appear in 1967. There are also various collections of Nietzsche's letters, notably including *Gesammelte Briefe*, 5 vol. (1900–09), supplemented by *Nietzsches Briefwechsel mit Franz Overbeck* (1916). The English collected edition by Oscar Levy, 18 vol. including index (1909–13), comprises unreliable translations by various hands. Translations of the major works are provided by Walter Kaufmann (ed.), *The Portable Nietzsche* (1954) and *The Basic Writings of Nietzsche* (1968).

BIBLIOGRAPHY. The *International Nietzsche Bibliography*, ed. by HERBERT W. REICHERT and KARL SCHLECTA, 2nd enlarged ed. (1968), lists over 4,500 studies in 27 languages. Some studies by otherwise reputable scholars are unreliable and have been refuted on crucial points; *e.g.*, CRANE BRINTON, *Nietzsche* (1941); and ARTHUR C. DANTO, *Nietzsche As Philosopher* (1965). GEORGE A. MORGAN, JR., *What Nietzsche Means* (1941), is scholarly but ignores Nietzsche's development. R.J. HOLLINGDALE, *Nietzsche: The Man and His Philosophy* (1965), is an intellectual biography (partially dated by more recent scholarship), but the author is not a philosopher. The two leading German Existential philosophers have written a great deal about Nietzsche: KARL JASPERS, *Nietzsche* (1936; Eng. trans., 1965) and *Nietzsche und das Christentum* (1938; Eng. trans., *Nietzsche and Christianity*, 1961); MARTIN HEIDEGGER, *Nietzsche*, 2 vol. (1961), in German. The present article is based on WALTER KAUFMANN, *Nietzsche: Philosopher, Psychologist, Antichrist*, 3rd rev. and enlarged ed. (1968), which contains a 26-page bibliography. Kaufmann's *From Shakespeare to Existentialism*, rev. ed. (1960), contains five chapters on Nietzsche, and his translations of eleven of Nietzsche's works feature commentaries.

(W.Ka.)

Niger

The Republic of the Niger, a landlocked West African country, takes its name from the river Niger, which flows through the southwest part of its territory. The name Niger derives in turn from the phrase *gher n-gheren*, meaning "river among rivers," in the Tamashek language. The republic has an area of 458,075 square miles (1,186,408 square kilometres) and a population of more than 4,200,000. It is bounded on the northwest by Algeria, on the northeast by Libya, on the east by Chad, on the south by Nigeria and Dahomey, and on the west by Upper Volta and Mali. The capital is Niamey, on the Niger River.

Physically the tenth largest country in Africa, Niger is only sparsely populated. The Sahara covers the northern part of its territory. Peanuts (groundnuts) and cattle are its principal products. The majority of its peoples are Muslim. Though the country obtained its independence on August 3, 1960, after 50 years of French rule, economically, it still relies heavily on France, and French is its official language. Hamani Diori has been president of the republic since independence.

The lack of transport has always constituted an obstacle to Niger's economic development. The recent discovery of large deposits of uranium at Arlit (Arhli) near Agadez, which are to be mined by a Franco-Niger company, has, however, opened new prospects of economic development. This circumstance offers new hope to a country that had previously been among the most remote and underdeveloped of African nations. (For coverage of related physical features, see CHAD, LAKE; NIGER RIVER; and SAHARA. For coverage of historical aspects, see WEST AFRICA, HISTORY OF.)

The landscape. *Relief.* Niger extends for about 750 miles from north to south, and about 930 miles from east to west. It tends to monotony in its features, is intersected by numerous depressions, and is dominated by arid highlands in the north. Rainfall increases as one proceeds southward so that the country divides naturally into three distinct zones—a desert zone in the north; an intermediate zone, where nomadic pastoralists raise cattle in the centre; and a cultivated zone in the south, where the greater part of the population, both nomadic and settled, is concentrated.

Three geographic zones

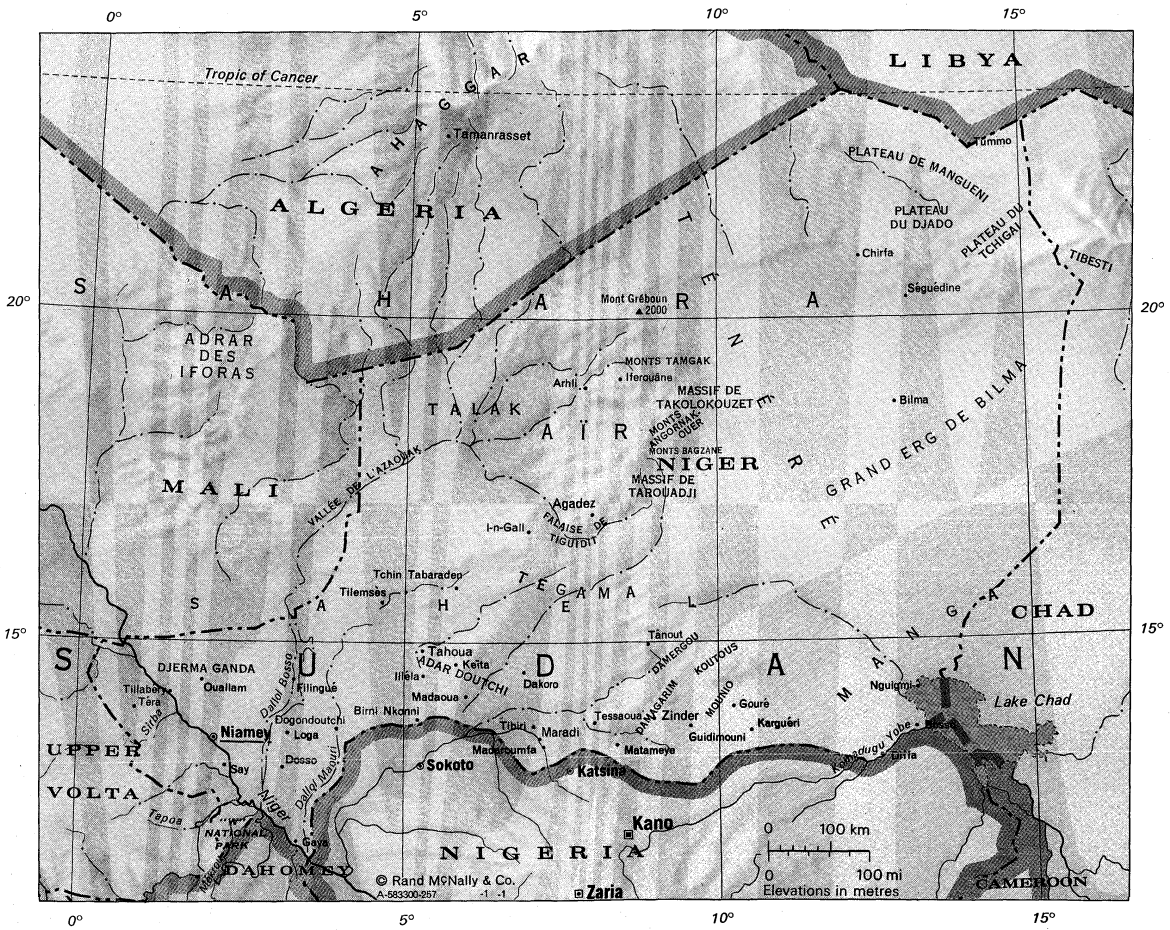
The highlands of the north are cut by valleys (*kori*) of the Massif de l'Aïr, which are an extension of the Ahaggar (Hoggar) Mountains of Algeria, and consist of a range running north to south in the centre of Niger, with individual mountain masses forming separate "islands": from north to south these are firstly Tazerzaït, where the highest point in the country is Mont Gréboun, which

reaches an altitude of 6,562 feet; Tamgak; Takolokouzet; Angornakouer; Bagzane; and Tarouadji. To the northeast are a series of high plateaus, which form a bridge between the Ahaggar Mountains of Algeria and the Tibesti mountains of Chad. From west to east, these are the plateaus of Djado, Mangueni, and Tchigaï.

The sandy regions of the Nigerian Sahara extend to either side of the Massif de l'Aïr. To the west, the Talak region includes the Tamesna area in the north (where fossil valleys are filled with moving sand dunes) and the Azaoua area in the south. East of the Massif de l'Aïr is the Ténéré region, covered partly by an expanse of sand called an erg, partly by stony plain called a reg.

The plateaus of the south, which form a belt about 900 miles long, may be divided into three regions. To the west is the Djerma Ganda region. Its large valleys are filled with sand, while *dallol* (fossilized valleys of rivers that formed tributaries of the Niger in ancient times) descend from the Massif de l'Aïr and the Adrar des Iforas massif (mountainous mass) of neighbouring Mali. The central region consists of the rocky Adar Doutchi and Majia areas; it is the region of the *gulbi* (dried-up valleys of former tributaries of the Sokoto River) and the Tegama —a tableland of sandstone, ending, towards the Massif de l'Aïr, at the Tiguidit scarp. To the east the underlying rock reappears in the Damagarim, Mounio, and Koutous regions, to the north of which is the region of Damergou, consisting of clays. In the Manga region, in the east, traces of ancient watercourses appear on the sandy plain.

Hydrography and soils. It is convenient to make a distinction between the ancient hydrographic system, which allowed agriculturalists, fishermen, and pastoralists to live in the Aïr region about 5,000 or 6,000 years ago, and the present simple system, which forms the basis



NIGER

MAP INDEX

of the marked difference between the northern and southern parts of the country. The present system includes to the west the Niger River Basin, and to the east the basin of Lake Chad; between the two occur vestiges of the older system, such as the *dallol* and the *gulbi*.

To the west, the Niger River crosses about 370 miles of Niger's territory. Because of the change in river flow, which occurs because of the dispersal of its waters in its interior "delta" region in Mali, it is only in January and February that it flows past Niamey in flood. At other times the river is fed by certain temporary watercourses that flow in from the right bank. These are the Gorouol, the Dargol, the Sirba, the Goroubi, the Djamangou, the Tapoa, and the Mékrou; the last two flow through the "W" National Park (so called because the Niger flows through the area in the form of a W). On the left bank, proceeding eastward, appear the *dallol*, the vestiges of the older watercourses. Generally running from north to south, they constitute zones of dampness, although a few still contain waters that flow towards the Niger. The best known are the Dallol Bosso, the Dallol Foga, and the Dallol Maouri. Other vestiges consist of the *kori*, which runs down from the Massif de l'Aïr and from former tributaries that had their sources in the massif of the Adrar des Iforas, and which flowed to a confluence at what is now the wadi (channel of a watercourse that is dry except during periods of rainfall) of Ti-m-merhsoï. No waters flow through the *kori* now, but water is still to be found beneath their sands. Other remnants of the old system are formed by the *gulbi*, through which water still flows annually, occasionally causing damage.

Ancient watercourses

To the east is situated the basin of Lake Chad, a large, shallow lake, which at its highest level has an area of approximately 9,650 square miles; of this, Niger possesses about 1,100 square miles. Its extent is considerably reduced during the dry season. The Komadougou Yobé River, which flows into Lake Chad from the west, forms part of the frontier between Niger and Nigeria. Its water level, which begins to rise in August, from January to May consists only of some stagnant pools.

In addition to the drainage system described, it may be noted that rainwater collects in several basins, so that some permanent lakes or pools also exist; these are found at Keïta and Adouna in the Adar Doutchi region, at Madaroumfa in the Maradi *gulbi*, and at Guidimouni to the east of Zinder. The water table underground can be tapped by means of artesian wells.

The soils fall into three natural regions. In the Saharan region in the north, which extends over an area of about 308,000 square miles, the soil remains infertile, except in a few oases where water is found. In the region known as the Sahel, which forms a transitional zone between the Sahara and the cultivated region to the south, the soils are thin and white, being covered with salty deposits that form an infertile crust. The third region (in the south) is cultivated. In this area the soils are associated with extensive dunes or uplands or with basins or depressions. Some of the soils in the latter, such as those in the Niger Basin and in the *gulbi* are rich. Black soils occur in the Kolo Basin. Throughout the region, however, and above all on the plateaus, layers of laterite (leached iron-bearing soil) occur.

Climate. Niger extends southward from the Tropic of Cancer, so that the greatest part of its territory lies in the dry tropical zone. In the southern part of the territory the climate is usually of the type known as Sahelian, which is characterized by a single rainy season. In January and February the continental equivalent of the northeast trade winds blow, dry and fresh, from the Sahara toward the Equator, meeting the harmattan, a dry wind which blows from east to west between the Tropic of Cancer and the Equator, hindering normal living conditions on the southern fringe of the desert. From April to May the southern trade winds blowing from the Atlantic reach the Equator and are diverted toward the Sahara where they meet with the harmattan—an encounter that results in tornadoes. A long dry season occurs between October and June, while the winter rains occur between June and October. The winter season begins with the June tornadoes, and lasts from between one to four months, according to the latitude; August is everywhere the rainy month.

Temperature and rainfall

Niger lies in one of the hottest regions of the world. Temperatures rise to the maximum from February to May, and drop during the winter rainy season, rising again somewhat before reaching the annual minimum in January or February. Annual minimum temperatures vary between 29° F (−2° C) at Bilma and 61° F (16° C) at Tillabéry. Tillabéry, which is on the Niger River, is also the hottest place, recording an annual average maximum of 106° F (41° C). The daily range is greater in the north than in the south, and is also more extreme during the dry season. The severity of the temperature increases from south to north: at Niamey the absolute extremes of temperature are 47° F (8° C) and 114° F (46° C); in the Massif de l'Aïr temperatures of −23° F (−31° C) have been recorded, while the absolute maximums are in the region of 122° F (50° C) in the shade.

Rainfall varies according to location as well as season. In general, it diminishes from south to north; the limit of assured annual rainfall is reached to the north of Agadez, where, furthermore, temperatures are excessive. The ten-inch isohyet (line on a map connecting points having equal rainfall) follows a line from Tahoua to Gouré, in effect marking the northern limit of nomadic pastoral life, for the rainfall permits a sparse vegetation to grow. To the south the 30-inch isohyet marks the southern limit of this zone, after which the southern agricultural zone begins. The extreme southwest receives the most abundant rainfall; the Gaya region is the first and the last to receive rainfall each year. Seasonal variations in rainfall are striking. In the course of the same winter rainy season a most irregular pattern of rainfall may occur, while from one year to another the amount of rainfall may double; in addition, the rainy season itself may arrive early or late, thus jeopardizing the crops.

Vegetation and animal life. The vegetation of the desert zone clusters around the oases; it includes the date palm and corn; its animal life, which must be able to endure hunger and thirst, includes the dromedary.

In the Sahel zone, where the doum palm and the cram cram (a prickly grass) appear, the vegetation has a short life cycle, and is principally used for grazing. Animal life includes the ostrich and the gazelle.

In the cultivated zone the vegetation includes acacia trees, doum palms, and rônier palms, as well as baobabs. Wild life, which has partially disappeared, includes antelope, elephants, and warthogs; giraffe are found in the Zarmaganda and Damergou regions, and hippopotamuses and crocodiles on the banks of the Niger. The extreme southwest is a savanna (grassy parkland) region where baobabs, kapok trees, and tamarind trees occur. Animal life is preserved in the "W" National Park, where antelope, lions, buffalo, hippopotamuses, and elephants may be seen.

Traditional regions. The southern part of Niger's territory is situated in the vast region of Africa known as the Sudan, in which, in former times, large political states arose, such as Ghana, Mali, and Songhai, as well as the Hausa states, the empire of Sokoto, and Bornu. The northern part of Niger remains the domain of the Tuareg. As a result of this geographic and historical heritage, Niger contains a multitude of traditional regions the names of which remain despite the establishment of contemporary administrative divisions.

Tuareg predominate in northern Niger's Aïr (Azbine) and Azaouak regions and have also constituted two further regions further to the south known as Imanan and Tagazart. In the northeast is the Tibesti region, which is the land of the nomadic Teda (or Toubou).

People of the south

In the south, from east to west, three groups of peoples are found. The first consists of the Kanuri, who occupy the Manga region, as well as parts of the regions known as the Damergou and the Damagarim, which are traditionally associated with Bornu in Nigeria. The second group is formed by the Hausa who occupy the traditional regions of Daura, Katsina, Gober Toudou, Adar Doutchi, Aréoua, and Kourfey. The third group consists of the

Zerma (Djerma) and Songhai peoples of southwestern Niger. The Zerma occupy the Zidji (or Zarmatarey), Boboye, Zarmaganda, and Fakara regions, while the Songhay are to be found in the Anzourou region to the north, and the Dendi region to the south, as well as on the right bank of the Niger. The Fulani, or Peul, people are found scattered in all these regions, and are found in compact groups only in the Boboye and Manga areas.

All these regions have a fluctuating political, economic, and geographic significance: the Hausa regions, for example, have been cut in two and divided between Niger and Nigeria. Most regions, moreover, have been and remain zones where contact takes place between different peoples—between the Hausa and the Tuareg in the Adar Doutchi region, for example; between the Tuareg and the Kanuri in the Damergou region; and between Hausa and Zerma in the Aréoua area.

The landscape under human settlement. Only about 8 percent of the population are town dwellers. The rural population is divided into nomads and sedentary peoples. There are about 8,000 villages, out of which 5,000 have fewer than 200 inhabitants; there are practically no villages in the desert zone. The Fulani herdsmen, who breed horned cattle and oxen, and the Tuareg, who raise goats, sheep, and dromedaries, tend to travel over the northern region during the winter. They meet together for the salt cure (*i.e.*, to permit the cattle to lick the salty soil of the area) in the In Gall region during August and September but move southward during the dry season. Both Fulani and Tuareg live in tribal groups and gain their subsistence from their livestock. The Fulani subsist above all on milk in various forms; while travelling they live in shelters that are often of a temporary nature. The Tuareg live on meat and dates and shelter themselves in tents.

The Fulani and the Tuareg

Sedentary peoples, such as the Hausa, the Songhai-Zerma, and the Kanuri, who inhabit the Niger and Chad Basins, live largely by agriculture. They raise millet, rice, corn, peanuts, and cotton. They also work as blacksmiths and shoemakers, while on the banks of Lake Chad and of the Niger the Buduma and Sokoro (Sorko) peoples are fishermen. Sedentary peoples live in dwellings that vary from those made of straw to those made of banco (hardened mud), although the Woko (Wogo) people live in tents of delicate matting.

There is a tendency among the nomads to settle down and the already sedentary peoples are expanding the lands under cultivation toward the north. Rural life, above all in its sedentary form, tends to slow its pace during the long dry season; it is at this time of year that migration to the towns or other countries occurs.

It was approximately in the 15th century that a few towns, such as Agadez or Zinder, were first established as halting places, or depots, on the trans-Saharan caravan routes. As commercial routes gradually developed on the coasts, however, these northern towns lost their former economic importance, while other centres, such as Birni Nkonni, and Tessaoua, declined in the course of the 19th century as a result of the colonial era.

At the present time there are four principal towns in Niger. Niamey, the political capital, has experienced rapid growth. A town of about 100,000 people, it has a cosmopolitan character; its transient population may rise to as many as 10,000. Its characteristic life varies between the European and African rural styles, including various intermediate steps, of which the style of life of the *évolués* (educated Africans) is the most distinctive. Zinder, for which the African name is Damagaram, is an older town than Niamey; a Hausa town, it was the capital of Niger until 1926 and has a number of skilled craftsmen, especially leatherworkers and dyers. The town is experiencing some industrial growth and has close links with Nigeria. Its population amounts to about 40,000 people. Maradi has a population of about 35,000 and has developed rapidly. The town is situated in the heart of the peanut-growing region near the Nigerian frontier. Many European companies have established branches there; the town is particularly renowned for its red goats, the skins of which are exported to Europe and the Americas. Tahoua has grown up on the edge of the desert. There, with its 31,000 inhabitants, it forms a large livestock market, where nomad pastoralists and sedentary farmers meet. But all of the towns remain little more than modest administrative and commercial centres. Because of the discovery of uranium ore it is expected that Agadez will experience a spectacular growth. Maradi, Zinder, and Tahoua all have buildings in the striking Sudanese architectural style.

Population centres

People and population. *Linguistic groups.* The largest linguistic group is formed by the Hausa, whose language, also spoken in Nigeria, is one of the most important in West Africa; 85 percent of the inhabitants of Niger understand Hausa, which possesses an abundant literature that has been printed in Latin characters in neighbouring Nigeria. Songhai is the second most important language; also called Songhoi, it is also spoken in Mali, in northern Upper Volta, and in northern Dahomey. In Niger itself it is divided into various dialects, such as Songhai proper, Zerma, and Dendi. The language of the Fulani is Fulfulde; in Niger it has two dialects, eastern and western, the demarcation line between them running through the Boboye district. Tamashek is the language of the Tuareg, who often call themselves the Kel Tamagheq, or Tamashek-speakers. The language is also spoken in Algeria and Mali and possesses its own writing, called *tifinagh*, which is in widespread use. Kanuri is spoken not only in Niger but also in Cameroon and Nigeria; the tongue is called *beriberi* by the Hausa. While these five languages are the principal ones spoken in Niger, there is also an important Teda, or Toubou, linguistic group in the Tibesti region. In addition, many of the peoples of Niger speak Arabic, and a still larger number read and write in that language; Agadez possesses one of the oldest Arabic schools in Africa. The use of the Arabic alphabet resulted in Fulfulde and Hausa becoming written languages; the resulting script is called *ajami*; a search for more old manuscripts in *ajami* is currently being conducted.

By using Hausa and Songhai, one may make oneself understood from one end of the country to the other. French, however, remains the official language, as well as the language of instruction, although it remains understood only by a small minority. English is taught as the principal foreign language in secondary schools.

Ethnic groups. Ethnic groups correspond to the five linguistic groups already mentioned. The Hausa are the most important, constituting more than half of the present population of Niger, even though the majority of the Hausa people live in Nigeria. The Hausas occupy the centre of southern Niger as far as Dogondoutchi. The Songhai-Zerma are found in the southwest; the Songhai proper live along the Niger, where they are assimilating the Kurtey and Woko peoples. The majority of the Songhai people as a whole, however, live in Mali. The Zerma live on the left bank of the Niger, remaining in close contact with the Maouri and Aréoua peoples. The Fulani, who are dispersed throughout the country, are 80 percent nomadic; they are also found dispersed throughout West Africa. The Tuareg, also nomadic, are divided into three subgroups—the Aulliminden of the Azaouak region in the west, the Asben (Kel Aïr) in the Aïr region, and the Itesan (Kel Geres) to the south and east of Aïr. The Tuareg people are also found in Algeria and in Mali. The Kanuri, who live to the east of Zinder, are divided into a number of subgroups—the Manga, the Dogara (Dagara), the Mober, the Buduma, and the Kanembu; they are also found living in Chad, Cameroon, and Nigeria. Apart from the Teda (Toubou), who constitute an important minority, the remainder of the population consists of Arabs, black Africans from other countries, and Europeans, of whom the greater part are French. In statistical terms it has been estimated that the Hausa form 54 percent of the population, the Songhai-Zerma 24 percent, the Fulani 11 percent, the Beriberi-Manga 9 percent, and the Tuareg and others 1 percent each.

Religious groups. Christianity (Catholicism and Protestantism) remains above all a religion of the towns, particularly of Niamey. There are several Christian missions in the Songhai and Aréoua areas. Christianity remains primarily a European religion, although it is also

practiced by some black Africans from other countries. The traditional animist religions of the black Africans continue to manifest themselves in strength. Though the Annaawaa group of Hausa have always refused to accept Islām, as have a group of Fulani, the Wodaabe—who distinguish themselves from other Fulani for this reason—Islām remains the religion of the majority of the peoples of Niger. It is practiced by at least 90 percent of them, is increasing its number of adherents, and constitutes a unifying force within the country. The various religious groups coexist harmoniously.

Demography. In 1972 the population numbered an estimated 4,243,000 inhabitants, representing a density of about nine persons per square mile. According to estimates made in 1970, the birthrate was 5.2 percent and the death rate 2.3 percent a year, giving a rate of population increase of 2.9 percent. The Kanuri people, however, have a much reduced rate of increase apparently due to the lower fecundity of Kanuri women; in general, the nomads have a lower rate of population increase than the sedentary peoples.

Rate of population increase

Immigration is insignificant; in 1960, for example, about 20,000 black Africans, Arabs, and Europeans entered the country, the Africans coming from Dahomey, Upper Volta, Mali, and Nigeria. Emigration or migration, on the other hand, is more important—whether to the towns of Niger itself or to other countries. In 1960, out of an estimated 130,000 persons who had migrated, 80,000 had travelled to the towns, and 50,000 abroad. An estimated 500,000 nationals of Niger are living abroad. The migrants usually leave the country after the harvest, passing the dry season in another country and returning before the opening of the rainy season. Many spend years abroad—the Songhai-Zerma people in Ghana and Togo, and the Hausa in the Cameroon, Nigeria, and Chad.

The distribution of population shows a great imbalance. Except for a few oases, the desert is virtually empty of human life. The majority of the population lives in the south, occupying about a quarter of the national territory.

Niger, Area and Population

	area		population	
	sq mi	sq km	1959-60 census	1972 estimate
Départements				
Agadez	244,869	634,209	55,000	82,000
Diffa	54,138	140,216	46,000	142,000
Dosso	11,970	31,002	332,000	560,000
Maradi	14,896	38,581	494,000	740,000
Niamey	34,862	90,293	727,000	919,000
Tahoua	41,188	106,677	608,000	871,000
Zinder	56,151	145,430	600,000	929,000
Total Niger	458,075*	1,186,408	2,864,000*	4,243,000

*Figures do not add to total because of rounding.
Source: Official government figures.

The national economy. Niger's only exports are peanuts and cattle. Previously one of the poorest of the French West African colonies, it has remained, after independence, one of the least endowed of the developing countries, with an average per capita income of about $70. Agriculture and livestock account for about 86 percent of the gross domestic product. In the production of both peanuts and livestock, Niger is third among West African countries.

Natural resources. Salt is traditionally exploited in the Kaouar and Aïr regions, as well as in the *dallol*, and in the Manga district. Annual production, which is insufficient, is estimated at 3,000 tons. Natron (hydrated sodium carbonate) is extracted locally. Cassiterite (the most important ore of tin) is mined at open workings in the Massif de l'Aïr, about 120 tons being produced in 1969. Gold is obtained by panning in the Sirba River; production amounted to about 11,000 pounds. About 30,000,000 tons of limestone and an important deposit of gypsum have been located at Malbaza and in the Ader Doutchi and Majia region. Cement production at Malbaza amounted to about 22,000 tons in 1967. The exploitation of uranium ore was due to begin in the Aïr region in 1971; known reserves in the Arlit region, amounting to about 20,000 tons, rank among the most important in the world. Annual production is expected to amount to 1,500 tons in 1973. Apart from tungsten, of which a little has been worked in the Aïr region, traces of copper, lignite (a brownish-black coal), molybdenum (a silver-white metal used as an alloy with iron in making high-speed cutting tools), zinc, oil, phosphates, and titanium (a metallic element used in making steel) have been found and are the subject of further prospection. In addition, a reserve of 250,000,000 tons of iron ore, with an iron content of 55 percent, has been located in the Say region, but is not considered commercially exploitable at this time.

Uranium ore

The exploitation of plant resources has long been practiced but on a small scale. The doum palm and the palmyra (rônier) palm provide wood for construction, while the 120,000 date palms of the Manga oasis produce 3,500 tons of dates a year. Small amounts of kapok (a silky down from the kapok tree, used for insulation, in making life jackets, and so forth) and of gum from the acacia gum tree are exported. Skins of ostriches, crocodiles, and snakes are used for making handicrafts that are exported to Europe. Fish from the Niger River and Lake Chad are exported southward to the coastal countries.

Energy. Hydrocarbon fuels, brought up from the port of Cotonou in Dahomey, first by rail and then by truck, constitute virtually the only source of industrial power; they are used either directly, or to drive diesel engines to generate electricity. In 1969 electricity production amounted to 30,000,000 kilowatt-hours; total installed capacity in 1970 amounted to 12,600 kilowatts. The Office of Solar Energy has been pursuing research, and has already produced solar batteries, which are used by the educational television program. Peanut shells have been experimentally used to supplement hydrocarbon fuels since 1968. Wood is the traditional domestic fuel. The possibility of establishing hydroelectric projects on the Mékrou and Niger rivers is being studied. It is anticipated that the production of uranium at Arlit may eventually result in a new source of energy for Niger.

Manufacturing. Some manufacturing industries have been established, mostly at Niamey. They produce chemicals, food products, textiles, transport equipment, and metal furniture. There are many small craft industries in the principal towns.

Management of the economy. The economic system is based upon planning but accords an important role to private enterprise. The three main policy objectives are the maintenance of national unity, the elevation of the living standards of the population, and the attainment of economic independence. The current four-year plan, establishing economic priorities, covers the period from 1970 to 1973. The private sector of the economy consists partly of a multitude of small enterprises, and partly of enterprises belonging to large French or international companies, such as the banks of Paris and Unilever Ltd. The government, through the agency of the Banque de Développement de la République du Niger, which is funded partly by aid from abroad, has promoted the establishment of about 30 companies, including real estate, road transport, air transport, and agricultural processing enterprises. In 1969, 16 companies were financed partly by the government and partly by private enterprise. There is a Chamber of Commerce, Agriculture, and Industry that promotes both private and public enterprise.

The private sector of the economy

Trade unions. In 1970 Niger had fewer than 15,000 wage earners, all of whom were members of the Union Nationale des Travailleurs du Niger, the only national trade union; its members are engaged in such activities as agriculture, education, and health services.

Economic perspectives. The economy has suffered from the fall in the prices of agricultural commodities, which, in addition to the drought that occurred in 1968, caused substantial losses. Industrialization and an increase in mining production are therefore envisaged as means of strengthening the economy. Problems associated with energy generation, communications, and of outlets to the sea will, however, have to be solved. Mean-

while, Niger is encouraging the strengthening of economic links between African countries. Apart from its membership in the Organization of African Unity, Niger is a member—together with the Ivory Coast, Dahomey, Upper Volta, and Togo—of the Conseil de l'Entente, a regional cooperative grouping, as well as of Organisation Commune Africaine, Malgache et Mauricienne, a larger grouping of French-speaking African states.

Transportation. While the economically active zone of the country consists of a belt running from east to west across the southern part of the country, the principal lines of communication run southwards towards the coast. The two ports used by Niger—Cotonou in Dahomey and Lagos in Nigeria—are each more than 600 miles away, and Niger possesses no railroad. Traditional systems of transport and communication are still largely relied upon. These include camel caravans in the northern Sahel region, canoes on Lake Chad and the Niger, and individual travel on horseback or on foot. Only a small tonnage of goods is transported. Out of about 11,000 vehicles, some 2,500 trucks maintain transport communications between Maradi and Zinder in Niger and Kano in Nigeria, and between Niamey, the capital, and Parakou in Dahomey; between them they transport virtually all imports and exports. Only a small amount of goods are transported by air.

Roads

The principal road axis enters the country from Gao in Mali, runs on the banks of the Niger as far as Niamey, and then continues eastwards to Zinder. It is currently being tarred, and is to be extended to Nguigmi on the shores of Lake Chad. From this central route roads branch off southwards. Toward the north, routes running via Tahoua and Tânout converge near Agadez, linking Niger to Algeria via Tamanrasset. The Air Niger Company is responsible for domestic air services linking the airports of Tahoua, Maradi, Zinder, Agadez, and Arlit. An International Airport at Niamey links Niger with the West African capitals of Abidjan, Libreville, and Bamako, and with Porto Novo in Dahomey, as well as with European capitals.

A bridge over the Niger near Niamey was completed in 1970. When the exploitation of uranium begins at Arlit, the development of a trans-Sahara route is to be considered, as well as the possibility of extending the railroads from Parakou in Dahomey and from Ouagadougou in Upper Volta to Niger's frontiers.

Administration and social conditions. The present administrative system combines the principles of western-style democracy with the traditional political system of the Niger region.

The structure of government. A constitution promulgated in November 1960 established a presidential type of regime, in which the president wields executive power, nominates cabinet ministers, and stands at the head of the army and of the administration. It is mandatory, however, that certain matters be considered by the cabinet. Legislative authority is exercised by the National Assembly, which is composed of 50 deputies. Judicial power is held by the Supreme Court, which also constitutes a final court of appeal. A state secretariat, a general commissariat for information, and a general commissariat for development are also attached to the presidency.

Niger is divided into seven *départements* (provinces)—Agadez, Diffa, Dosso, Maradi, Niamey, Tahoua, and Zinder—each of which is administered by a prefect assisted by a commission of specialists concerned with development. Each *département* is divided into *arrondissements* (administrative subdivisions), 33 in all, each of which is administered by a subprefect.

Political organization. According to the constitution, suffrage is "universal, equal, and direct." The president is elected for five years by direct universal suffrage; he may be re-elected. The deputies of the national assembly are elected in the same way from a national roll. Niger is one of the few African states of French expression that has not decreed the constitutionality of a single party system. Niger, nevertheless, has a single party, the Parti Progressiste Nigérien-Rassemblement Démocratique Africain. The president of the Republic, Diori Hamani, heads the secretariat-general of the party, and the president of the National Assembly, Boubou Hama, heads the Parti Progressiste Nigérien. The participation of the population in national life is maintained by rural political campaigns; radio programs on political themes are also broadcast.

Education. Out of a school-age population that in 1970 numbered about 750,000, about 90,000 children are in school; most of these are boys rather than girls. Of the total, about 84,000 are in primary school, and 6,000 in secondary school. In addition, there are facilities for teacher training. Primary and secondary schools and teacher-training colleges are the responsibility of the Ministry of National Education. Other ministries are responsible for technical education. Educational television so far reaches only an estimated 800 children. Literacy programs are conducted in the five principal African languages. In October 1971 a centre for advanced scientific studies is due to open, and will constitute the first institution for higher learning in the country. Niger devotes about 10 percent of its budget to education.

Health and welfare. Health services are organized on a mass basis and concentrate on the eradication of certain diseases in rural areas, as well as on health education. Campaigns have been successfully waged against sleeping sickness and meningitis, while vaccinations against smallpox and measles are also conducted. Other diseases, however, such as tuberculosis, malaria, and leprosy remain endemic. There are about 2,500 hospital beds available in Niger's hospitals. Anti-tuberculosis centres are established at Niamey, Zinder, and Tahoua. Lack of finances and shortage of trained personnel remain the principal obstacles to the improvement of health conditions.

Cultural life and institutions. Niger forms part of the vast Sahelian cultural region of West Africa. Although the influence of Islām is predominant, pre-Islāmic cultural traditions are also strong and omnipresent. Paradoxically, the numerous ethnic strains to be found in Niger have resulted in a strengthening of the fabric of national life. Since independence, greater interest has been shown in the country's cultural heritage, particularly with respect to traditional architecture, handicrafts, dances, and music. With the assistance of the United Nations Educational, Scientific, and Cultural Organization a regional centre for the collection of oral traditions has been established at Niamey.

An institution predominant in cultural life is the National Museum at Niamey, and a cultural celebration is National Youth Week, held annually in December. Among youth the most popular sports are football, basketball, boxing, and cycling.

The press includes official publications, such as the *Journal Officiel de la République du Niger, Le Temps du Niger* (a daily), and *Le Niger* (a weekly), and specialized publications put out by the Chamber of Commerce and the Centre Régional de Recherche et de Documentation pour la Tradition Orale. There are eight newspapers published in the national languages—five in Hausa, one in Songhai, one in Tamashek, and a bilingual publication in Hausa and Songhai. There are several radio transmitters which between them cover the entire country. Television is so far uniquely educational. Broadcasts are in French and in the five principal languages.

Prospects for the future. Niger can lay claim to a share in a proud historic heritage—that of the region known as the West African Sudan. But, because of its geographical situation, it has remained in the shadow for much of the time that has elapsed since its creation as a separate political unit—that is to say, since the beginning of the colonial era. Endowed with unfavourable natural conditions, including a severe climate and a reputed paucity of resources, Niger has suffered above all from the circumstance that distance makes it difficult of access, and from the fact that its transportation system has been insufficiently developed. It is hardly to be wondered at, therefore, that people of other continents frequently confuse it with its larger southern neighbour, Nigeria.

The need for improved communications

Several favourable factors may, nevertheless, be distinguished. It does not suffer from a demographic pressure on the land, or from a rate of population increase that

robs it of the benefits of economic progress. Exports, few as they are, are sufficiently diversified not to make the economy dependent upon, or vulnerable to, price fluctuations or market speculations. There is, moreover, reason to hope that the projected exploitation of mineral resources, particularly uranium, will substantially increase national revenues and permit the construction of roads and other facilities—social no less than logistic—that will later open the way to further economic development.

Niger may thus be said to hold a privileged place among the developing countries. Having for long had to learn to moderate its ambitions and to rely upon its own efforts, having experienced a poverty, which is not to be confused with misery, and having neither made, nor had to resist, any territorial claim, Niger has remained apart from ideological and other conflicts, and has thus found itself able to play a moderating role in international relations. For these reasons, as well as others already mentioned, the future prospects opening before Niger would appear relatively bright.

BIBLIOGRAPHY. EDMOND SERE DE RIVIERES, *Histoire du Niger* (1965), a complete historical synthesis; PIERRE DONAINT, *Le Niger* (1965), geography textbook for secondary schools (useful for reference); SUZANNE BERNUS, *Particularismes ethniques en milieu urbain: l'exemple de Niamey* (1969), a recent study of the development of the urban centre of Niamey; SERVICE DE LA STATISTIQUE, NIAMEY, *Annuaire Statistique . . .* (annual), climatological, social, and economic statistics.

(Di.L.)

Nigeria

Nigeria—in full, Federal Republic of Nigeria—is the largest of the West African coastal states. Its population of 55,074,000 (1970) is the largest of any country in Africa. With an area of 356,699 square miles (923,773 square kilometres), it is the 13th largest state on the continent. Located approximately between 4° and 14° N, and 3° and 14° E, its territory extends about 650 miles (1,050 kilometres) from north to south, and 700 miles east to west. It is bordered on the south by the Gulf of Guinea, on the west by the Republic of Dahomey, on the north by the Republic of the Niger, and on the east by the republics of Chad and Cameroon. Part of the eastern boundary runs along the crest of the Adamawa Plateau.

Modern Nigeria dates from 1914, when the two British protectorates of Northern and Southern Nigeria were joined. The country became independent on October 1, 1960, and three years later adopted a republican constitution but elected to remain a member of the Commonwealth of Nations. Relics of British rule are still to be seen in various aspects of Nigerian life. The official language, English, is likely to remain unchanged, since there are more than 200 different languages spoken by the many national groups living in the country. Trade and cultural contacts with the more distant English-speaking countries of Ghana and Sierra Leone remain stronger than those with the adjacent French-speaking Dahomey, Niger, and Cameroon. (For associated physical features, see CHAD, LAKE; GUINEA, GULF OF; and NIGER RIVER; for details of cities, see IBADAN; LAGOS; for historical aspects, see WEST AFRICA, HISTORY OF.)

THE LANDSCAPE

Relief features. Nigeria is on the lower part of the great African continental plateau, which slopes slowly downward from south and east to north and west. Nigeria itself consists of several eroded surfaces, occurring as plateaus, at elevations of 2,000 feet (610 metres), 3,000 feet, and 4,000 feet above sea level. The coastal areas, including the Niger Delta, are covered with young soft rocks, commonly found in the Lake Chad Basin, and the western parts of the Sokoto region. Gently undulating plains, which become waterlogged during the rainy season, are found in these areas. In most parts of the Western State, and in the central part of the six northern states, the underlying rocks are old and hard, and the characteristic landforms consist of high plains with broad shallow valleys, dotted with numerous hills or inselbergs (steep-sided residual masses of rock, left after erosion).

The Udi Hills

The Udi Hills, with their scarp faces turned to the east, are perhaps the country's most prominent relief feature. Other prominent relief forms include the Jos Plateau and the Biu Plateau, both of which are dotted with many extinct volcanic cones. The craters of these volcanic hills are well preserved; several of them contain crater lakes.

Drainage. There are three major drainage areas—the Niger–Benue Basin; the Lake Chad Basin; and the coastal, or Gulf of Guinea, basin. The Niger River (*q.v.*), after which the country is named, and the Benue, its largest tributary, are the principal rivers. Both have their sources outside the country. The Niger has numerous rapids and waterfalls, but the Benue (whose valley, in its Nigerian course, is cut through young sedimentary rocks) is not interrupted by waterfalls and is navigable throughout its length whenever the water level is high enough. All the rivers draining the area north of the Niger–Benue trough rise on the Jos Plateau. These include the Sokoto, the Kaduna, and the Gongola as well as the rivers draining into Lake Chad. The coastal areas are drained by short rivers, which flow from north to south into the Gulf of Guinea.

Navigation is restricted to river stretches unhampered by rapids or falls. During the months of the dry season, the low water level renders navigation impossible, even along the Benue, which is free of rapids. At this season the smaller streams may dry up completely.

Only a small part of Lake Chad lies within Nigerian territory.

Climate. Nigeria has a tropical climate with wet and dry seasons. It is hot and wet throughout the year in the southeast but markedly dry in the southwest and further inland. The duration of the seasons depends on the relation of the area to the sea or to the Sahara. Three climatic patterns are distinguished: (1) a tropical wet climate in the southeast with uniformly high temperatures and heavy rainfall distributed throughout the year; (2) a tropical wet and dry, or savanna, climate in the north and west; and (3) the dry, or steppe, climate in the far north.

Two air masses, the equatorial maritime and the tropical continental, dominate the climate. The former is associated with the rain-bearing southwest monsoon, which blows from the ocean; the latter is associated with the harmattan, a dry and dusty wind from the Sahara. In general, the length of the rainy and dry seasons decreases from south to north. In the south, the rainy season lasts from March to November. In the far north, however, it lasts only from mid-May to September. This pattern is interrupted in the south, where rainfall reaches a peak twice a year and where there is a break in the rains in August. There are thus four seasons in the south: the long rainy season (March to early August), the short dry season (August), the short rainy season (September to early November), and the long dry season (mid-November to February).

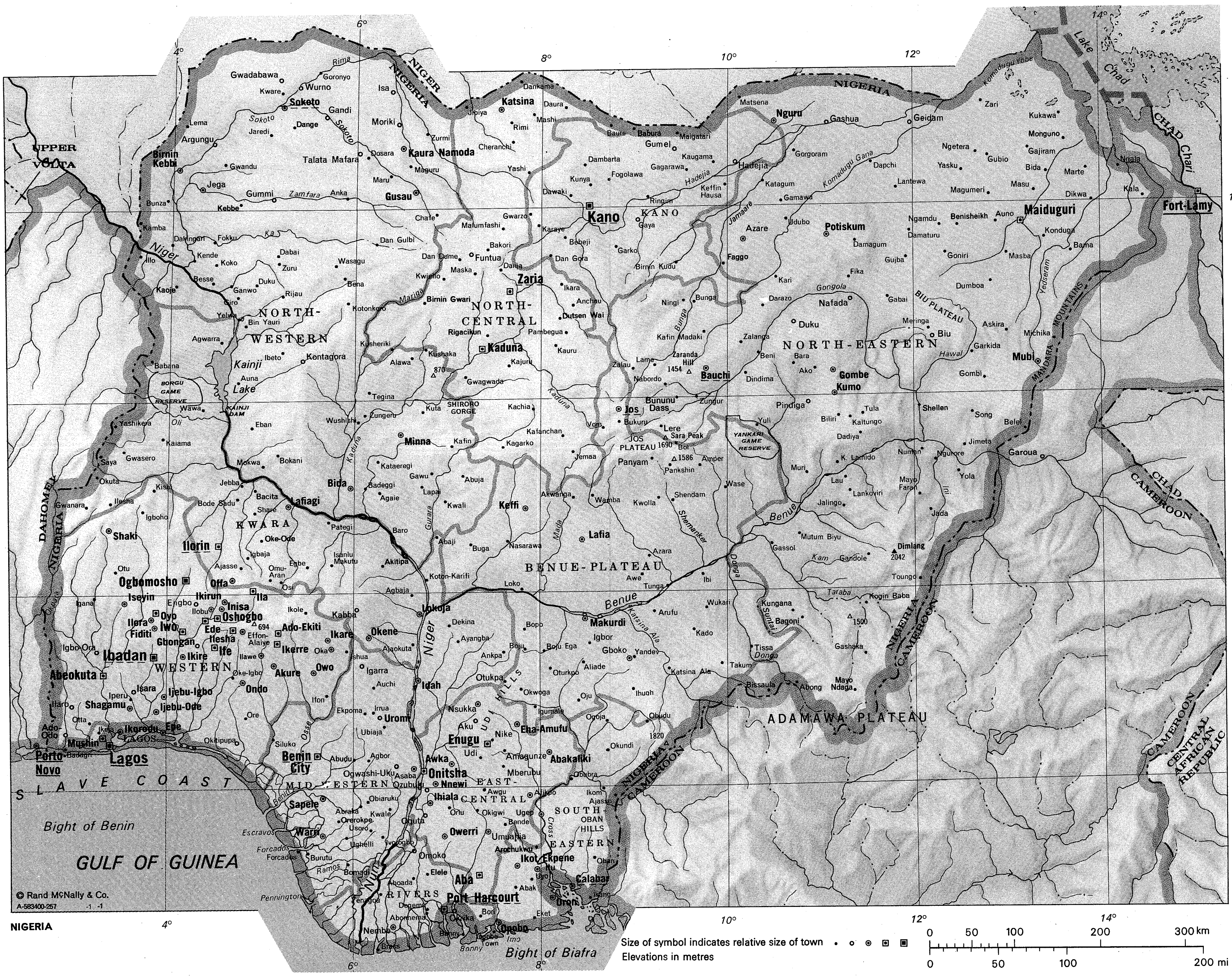
Rainfall is heavier and more reliable in the south, particularly in the southeast, which has more than 120 inches (3,050 millimetres) a year, as compared with 70 inches in the southwest. The annual rainfall decreases as one moves farther from the coast; in the far north it is not more than 20 inches. The rainy season is preceded by intense heat, after which the drought is broken by sharp thunderstorms accompanied by lightning, during which as much as one and a half inches of rain may fall in less than an hour.

Temperature and humidity remain relatively constant throughout the year in the south. In the north, however, considerable seasonal changes occur, and the daily temperature range is wide during the dry season. On the coast, the mean monthly maximum temperatures are steady throughout the year, remaining, for example, constant at 95° F (35° C) at Lagos and at about 85° F (29° C) at Port Harcourt; the mean monthly minimum temperatures remain approximately at 70° F (21° C) for Lagos, and at 73° F (23° C) for Port Harcourt. In the northeastern city of Maiduguri, on the other hand, the mean monthly maximum temperature may exceed 100° F (38° C) during the hot months of April and May, while in the same season frosts can also occur at night.

MAP INDEX

Political subdivisions

Cities and towns


NIGERIA
NIGER
UPPER VOLTA
DAHOMEY
CHAD
CAMEROON
CENTRAL AFRICAN REPUBLIC
Fort-Lamy
Lake Chad
Chari
NORTH-WESTERN
NORTH-CENTRAL
KANO
NORTH-EASTERN
KWARA
BENUE-PLATEAU
WESTERN
MID-WESTERN
EAST-CENTRAL
SOUTH-EASTERN
RIVERS
LAGOS
ADAMAWA PLATEAU
JOS PLATEAU
BIU PLATEAU
MANDARA MOUNTAINS
OBAN HILLS
UDI HILLS
SHIRORO GORGE
BORGU GAME RESERVE
YANKARI GAME RESERVE
KAINJI DAM
Kainji Lake
Niger
Benue
Kaduna
Sokoto
Rima
Zamfara
Ka
Gongola
Hadejia
Komadugu Gana
Komadugu Yobe
Jamaare
Katsina Ala
Donga
Taraba
Cross
Osse
Benin
Escravos
Forcados
Ramos
Nun
Bonny
SLAVE COAST
Bight of Benin
GULF OF GUINEA
Bight of Biafra
Sokoto
Gwadabawa
Wurno
Goronyo
Gandi
Dange
Argungu
Birnin Kebbi
Jega
Gummi
Kebbe
Bunza
Kamba
Gusau
Kaura Namoda
Talata Mafara
Katsina
Daura
Funtua
Zaria
Kano
Kaduna
Nguru
Gashua
Geidam
Hadejia
Potiskum
Maiduguri
Azare
Faggo
Bauchi
Jos
Nafada
Gombe
Kumo
Biu
Mubi
Yola
Jimeta
Garoua
Minna
Bida
Keffi
Lafia
Makurdi
Lokoja
Ilorin
Offa
Ogbomosho
Oyo
Iwo
Ede
Oshogbo
Ilesha
Ife
Ibadan
Abeokuta
Ijebu-Ode
Ijebu-Igbo
Shagamu
Ikorodu
Epe
Mushin
Lagos
Porto Novo
Ondo
Akure
Owo
Ado-Ekiti
Ikare
Ikerre
Okene
Idah
Benin City
Sapele
Warri
Onitsha
Nnewi
Awka
Enugu
Abakaliki
Eha-Amufu
Owerri
Umuahia
Aba
Port Harcourt
Opobo
Ikot Ekpene
Calabar
Oron
Uromi
Kontagora
Lafiagi
Ilorin
Dimlang 2042
870
1820
1500
Size of symbol indicates relative size of town
Elevations in metres
0 50 100 200 300 km
0 50 100 200 mi
© Rand McNally & Co.
A-583400-257

southeast—the original forest vegetation has consequently been completely replaced by open palm bush. In the Western and Mid-Western states, large areas of forest have also been replaced by cocoa and rubber farms.

Tree-studded savanna (tropical grassland) occupies more than half the area north of the forest belt. Trees characteristic of this area are the baobab, the tamarind, and the locust bean. The savanna landscape becomes more open in the far north and is characterized by scattered stunted trees and short grass. Semidesert conditions appear in the Lake Chad region, where common trees include various species of acacia (of which one is the source of gum arabic) and the doum species of palm. Gallery forests (narrow forest zones occurring along rivers) are also characteristic of the open type of savanna landscape encountered in the north.

Animal life. Camels, several species of antelope, hyenas, lions, and giraffes are found in the grassland to the north, while the red riverhog, the forest elephant, the chimpanzee, and some varieties of birds and snakes are confined to the rain forest belt. Animals found all over the country include leopards, the golden cats, monkeys, gorillas, and wild pigs. Rodents, or gnawing animals, constitute the largest family of mammals and are also ubiquitous. Among common rodents are various squirrels, the porcupine, and the cane rat, known locally as "Cutting-Grass." The northern grasslands also abound in Guinea fowl. Other common birds include quails, vultures, kites, bustards, and gray parrots. The rivers contain a great variety of fish, crocodiles, and hippopotamuses. There are many butterflies, moths, and insects. Large scorpions are to be found, as also are goliath beetles.

Soils. The four main soil groups correspond closely with the main climatic and vegetation zones, which comprise the coastal swamp and alluvial soils, the rainforest soils, the lateritic soils (red soils, leached of silica and containing iron and aluminum hydroxides), and the sandy soils of the north.

Along the coast, the soils are either sandy or swampy and, like the soils of the forest belt, are heavily leached. In the rain forest belt, soils derived from old hard rocks, complex in structure, which pre-date the sedimentary rocks found elsewhere, support cocoa trees, while those derived from sandstones do not. Under cultivation, forest soils soon lose their fertility, which is concentrated in a thin top layer. Lateritic soils, which form along gentle slopes in areas with a markedly dry season, are widespread. Rich in iron compounds, and sometimes so hard as to appear to be rocks, they are difficult to cultivate.

Soil erosion

Soil erosion is most obvious in those densely populated areas of northern and eastern Nigeria in which overcultivation and overgrazing have exposed the soil to erosion by wind and running water. The areas most affected include the scarplands of Iboland in the east, where the threat posed by advancing gullies has resulted in the abandonment of some villages; the Jos Plateau in the centre; and the Kano–Katsina region and parts of Sokoto region, in the north. In the extreme north, wind erosion is particularly noticeable toward the end of the dry season, when the storms preceding the onset of the rains blow away much soil.

The landscape under human settlement. Marked differences exist between north and south not only in physical landscape, climate, and vegetation but also with respect to social organization, religion, literacy, and agricultural practices. These differences, due in part to the fact that the north is landlocked and the south is not, and in part to historical antecedents, form the basis of the division of Nigeria under human settlement into three main regions—the south, or Guinea coastlands; the middle belt; and the north, or Nigerian Sudan ("Sudan," as a historic region and as distinct from the state of that name, here signifies the open savanna belt which runs east to west from the Nile to the Atlantic).

The south is the most developed part of Nigeria. Its forest resources are intensively exploited, and its tree crops harvested on peasant farms as well as in commercial plantations. All the major industrial centres and oil fields, as well as the seaports, are concentrated in this

In general, as one proceeds northward, mean maximum temperatures increase, while mean minimum temperatures decrease. Owing to the blanketing effect of clouds during the rainy season, and of dust haze during the harmattan, the heat of the sun is not as fierce as might be expected. The climate of Nigeria is rendered trying not so much by high temperatures as by the high relative humidity, which, for example, at 6 AM averages 95 percent at Port Harcourt, and 85 percent at Lagos.

The harmattan season

The relative humidity falls considerably during the harmattan, which blows for over three months in the north but rarely for more than two weeks along the coast. During the harmattan period, the climate is invigorating, dust is pervasive, and the climate is excessively dry, causing lips to split and furniture to crack.

Vegetation. Vegetation in Nigeria is governed by the south to north decrease in rainfall, and the main vegetation belts run, therefore, in broad east to west belts, parallel to the Equator. Mangrove and freshwater swamps occur along the coast and in the Niger Delta. A few miles inland, swamps give way to dense tropical rain forests, in which the most important economic species of tree include such hardwoods as mahogany, iroko (a tree with mottled wood), and obeche, which has whitish wood. The oil palm tree, which is economically valuable, grows wild in the forest and is usually preserved when the forest is cleared for cultivation. In the more densely populated parts of Iboland and Ibibioland—areas in the

region. The south itself also consists of several cultural regions, of which the most important are the Yoruba areas in the west, the Benin area in the central part, and the Ibo–Ibibio area in the east.

The middle belt, stretching across the centre of the country, is the most sparsely settled and least developed part of Nigeria. The peoples inhabiting it are divided into more than 180 linguistic groups. Although a few highly capitalized schemes, such as the Kainji Dam and the Bacita sugar project, have been developed there, its future depends primarily upon its ability to increase its food production.

Historically and culturally, the Nigerian Sudan is a region of great interest. Until the beginning of the 20th century, when a new economic pattern was created by the construction of the railroad to the coastal ports, the region maintained regular trans-Saharan contacts with the Mediterranean and the Middle East. Islām is the predominant religion. It is a cattle zone, inhabited by the nomadic cattle-owning Fulani people and by the Hausas, who are settled cultivators. Except in the Lake Chad Basin, where the Kanuri people established the state of Bornu, the Nigerian Sudan is dominated by a blend of Hausa–Fulani culture.

Rural settlement. About 90 percent of the people live in rural settlements consisting mainly of small hamlets and villages. In parts of Iboland and the Anang region in the southeast and of Tivland in the central region, rural settlements consist of dispersed homesteads, called compounds. Each compound houses a man, his immediate family, and some relations. The compound is enclosed by a fence of matting and sticks or by a wall of mud or concrete. It is usually surrounded by a small garden area, called compoundland, which is cropped every year with corn, vegetables, and yams. A number of compounds make up the hamlet or village, which is usually inhabited by people claiming a common ancestor—often the founder of the village, after whom the settlement may be named.

Composition of compounds

Rural Nigerians do not use stone for building but depend on whatever materials are at hand. House types therefore change as one travels inland. Along the coastlands, where the soil is too sandy for making daub, the walls of houses consist of bamboos tied together with ropes, with the roofs being made of bamboo leaf mats. Bamboos, ropes, and mats are all made from the raffia palm, which abounds in the region. Rectangular mud houses with mat roofs are also found in the forest belt; the houses of the more prosperous are roofed with corrugated iron sheets. In the extensive savanna areas of the middle belt, and in parts of the north, houses consist of round mud buildings roofed with grass thatch; in the drier areas of the extreme north, flat roofs of mud replace the sloping grass thatch.

Each village has a chief, or headman, who usually is one of the oldest men in the community, and who usually rules by consent of the people. This is particularly true in the eastern states. In Yorubaland, and in most parts of the northern states, the chief is usually more powerful and is held in high esteem. A characteristic feature of village life is the existence of an age-grade system of social organization, which groups together people of the same age group. Communal jobs, such as road building or the raising of funds for a social purpose, such as building a school, are usually organized on the basis of the age-grade system.

Urban settlement. With the exception of the Yorubas, the Hausas, and the Kanuris, Nigerians were not town dwellers before the 20th century. The Yorubas, of whom perhaps half live in towns of over 5,000 persons, are the most urbanized people in tropical Africa. Their towns, most of which are several hundred years old, were originally administrative and trading centres, and most of them still are. The more important Yoruba towns, with their approximate populations, are Ibadan (*q.v.*), with about 758,000 people living in the municipality alone; Ogbomosho (387,000); Abeokuta (226,000); Ife (157,000); and Oyo (136,000).

The northern towns of the Nigerian Sudan, including Kano (with almost 357,000 inhabitants), Zaria (201,000), Sokoto (109,000), and Katsina (109,000), are much older than the Yoruba towns. Owing their growth to the trans-Saharan trade, as well as to the agricultural wealth of the Sudan, these ancient towns, like other pre-industrial cities, were unplanned, and they consist of an amorphous assemblage of mud buildings. Their inhabitants include traders, farmers, and administrators as well as craftsmen, musicians, and drummers.

During the period of British rule, many new towns grew up, and the older ones grew larger. Many towns originally were primarily administrative centres but—like the southern towns of Port Harcourt, Lagos, and Ibadan, as well as the Sudanese towns of Kano and Kaduna—have since become industrialized. Outside the walls of the ancient cities, the British established two customarily segregated towns or quarters—one, for white administrative officers and commercial agents, known as the Reservation and the other, for Nigerians from other regions, known as Sabon Gari (strangers' town). Such quarters exist in most cities in Yorubaland and Hausaland. Along the coast there are a few important pre-British trading towns, such as Calabar and Bonny. Other port towns, such as Port Harcourt and Sapele, were established during the colonial period.

Ancient and modern towns

Lagos (*q.v.*), the federal capital, is the only conurbation in Nigeria. Primarily a Yoruba town, it is built on an island. It is also the most industrialized city in the country and is one of the largest towns in tropical Africa.

PEOPLE AND POPULATION

The great diversity of peoples and cultures in Nigeria is largely a result of the location of the country at the meeting point of transcontinental migration routes from north to south, west to east and southeast to northwest. There are over 200 ethnic groups in the country, each of which has its own customs, traditions, and language. The larger groups include the Hausas (perhaps 6,000,000), the Fulanis (perhaps 5,000,000), the Yorubas (perhaps 10,000,000), and the Ibos (perhaps 7,000,000). Other prominent but less numerous groups include the Edos, of Benin City, as well as the Ibibios, in the forest belt; the Tivs and Nupes, in the middle belt; and the Kanuris, in the Chad Basin. The greatest concentration of smaller ethnic groups occurs—as already mentioned—in the middle belt, where there are over 180 linguistic groups. In the Niger Delta, the home of the Ijaws, social organization was altered radically during the period of the slave trade, in part because of the forced migration of peoples from the interior into the area and in part because of contact with European traders. The distinct cultural group which emerged stressed its cultural separation from other groups rather than its common descent. Jaja of Opobo, a 19th-century chief renowned for defying the British, was an Ibo who grew up among the Ijaw people of the Niger delta.

Racial and religious groups. Nigeria is a country of black-skinned peoples. The peoples of the savanna zone in the north tend to be taller than those of the forest belt of the south. Arab penetration into the Chad Basin has resulted in much racial mingling; the Shuwa Arabs and the Kanuris of this region are of mixed Negro and Arab origin. The cattle-owning Fulanis still maintain non-Negro features, but the town Fulanis have intermingled with Negroes. (The origin of the Fulani people—also known as Fulbes, Fulas, Fellatas, or Peuls—remains undetermined. Earlier migrations resulted in their establishing states throughout the Sudan region, from Fouta Toro in Senegal, to Macina in the Sudan itself. Speculation has attributed to them a West African, an East African, and a Middle Eastern origin.) Much intermingling has also occurred in the south, particularly in the coastal port towns of Calabar, Abonnema, and Warri, where many Syrians and European traders settled in the course of the last 70 years. Much of the country was originally animist in its beliefs, worshipping idols. At the time of the 1963 census, about 47 percent of the people were Muslims, and 35 percent professed Christianity. Though these figures might suggest that few people now worship idols, the fact is that many professing Muslims and Christians remain

Nigeria, Area and Population

	area		population	
	sq mi	sq km	1963 census	1971 estimate
States				
Benue-Plateau	39,204	101,538	4,009,000	4,856,000
East Central	11,548	29,909	7,228,000	8,754,000
Kano	16,630	43,072	5,775,000	6,995,000
Kwara	28,672	74,260	2,399,000	2,906,000
Lagos	1,381	3,577	1,444,000	1,847,000
Mid-Western	14,922	38,648	2,536,000	3,072,000
North-Central	27,108	70,210	4,098,000	4,964,000
North-Eastern	105,025	272,015	7,793,000	9,440,000
North-Western	65,143	168,720	5,733,000	6,945,000
Rivers	6,985	18,091	1,544,000	1,871,000
South-Eastern	10,951	28,363	3,623,000	4,388,000
Western	29,100	75,369	9,488,000	11,492,000
Total Nigeria	356,669	923,773*	55,670,000*†	67,529,000*†

*Figures do not add to total given because of rounding. †The 1963 census population is possibly overstated. ‡The 1971 estimate is possibly overstated; UN 1970 estimate, 55,074,000.
Source: Official government figures; UN.

idol worshippers. Religious freedom is entrenched in the constitution, and to some extent in all parts of the country; but especially in the Lagos and Western states, Muslims and Christians live and work together. The greatest concentration of Muslims is in the northern states, where 72 percent of the people profess Islām. In the Yoruba west, Christians are slightly more numerous, and in the eastern states they make up 77 percent of the population.

The main Christian groups are Roman Catholic, British Methodist, Anglican, and American Baptist. All of these church groups, as well as some Muslim sects, own and run schools and hospitals throughout the country. The development of education in the country has always primarily been the responsibility of religious groups, and about 90 percent of literate Nigerians have attended missionary institutions. This pattern is now being modified, however, because, since the civil war of the late 1960s, some states have taken over the control of schools from missionary groups.

Hausa as the lingua franca

Ethnic and linguistic groups. Hausa is the most widely spoken African language in Nigeria. It is spoken by the Hausas and the Fulanis but is also the lingua franca in the northern states. As a result of the Fulani conquest of Hausaland in the early 19th century, and the subsequent imposition of Fulani rule, the two groups live together in the same towns and villages. The religion of both groups is Islām. The town Fulanis, who are less orthodox in their religion, remain a distinctive aristocratic group. While they intermarry freely with Hausas and other groups, they nevertheless continue to control the administration of the Hausa towns. The cattle-owning Fulani, on the other hand, are not only more orthodox but also more disinclined to intermarry. Pure Fulani blood is therefore found more often amongst these nomadic herdsmen, than it is in the towns. The pastoral Fulani, while more ardent Muslims than those of the towns, are also, paradoxically, less subject to Islāmic influences. They also speak the Fulani language—Fulfulde—rather than Hausa. Unlike the Fulanis, the numerically dominant Hausas are settled cultivators as well as renowned traders. Cattle, however, including those owned by the settled Hausa farmers, are cared for by the Fulanis. At the time of the British conquest, the Fulani had established a rather unwieldy empire that extended beyond Hausaland to include vast areas occupied by the small groups of the middle belt, but before the Fulani conquest, which occurred in the early 19th century, the Hausas were organized into large states, of which the most prominent were Zaria, Kano, and Gobir.

Another important linguistic group consists of the 10,000,000 Yoruba-speaking peoples who, like the Hausas and the Fulanis, have ancient connections with the Middle East. The Yorubas, although they are farmers, often live in large pre-industrial cities. Each Yoruba subgroup is ruled by an influential paramount chief, or *oba,* who is usually supported by a council composed of chiefs of various ranks. The *oni* of Ife, who is the accepted spiritual leader of the Yorubas, and the *alafin* of Oyo, who is their traditional political leader, are the two most powerful rulers; their influence is acknowledged throughout Yorubaland. The various Yoruba subgroups also share a traditional religious system. It features the worship, through cults and secret societies, of gods such as Ogun, the god of war and iron; Shango, the god of thunder and lightning; and Orisha Oko, the goddess of farmland.

The Ibo-speaking peoples, whose leaders unsuccessfully attempted between 1967 and 1970 to establish the independent state of Biafra, are one of the largest linguistic groups in Nigeria. They live in small dispersed settlements and have never organized into large political units. Traditional Ibo society has rather been ultrademocratic; the largest political unit has been the village group, ruled by a council of elders rather than by a chief. The Ibos have a reputation throughout West Africa for energy and individualism. The relatively rapid progress of Iboland owes much to community efforts made at the village level, or through the extended family system.

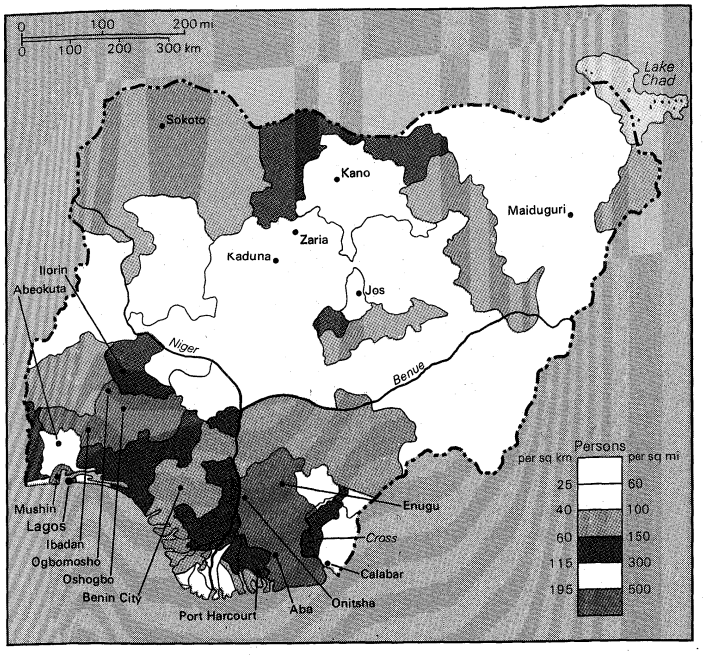
Other large linguistic groups include the Ibibios (who live near to the Ibos and share many common characteristics with them), and the Edo people of Benin City, whose culture has largely been influenced by their Yoruba neighbours. In the middle belt, the Tivs and the Nupes form the largest groups. Both are settled cultivators, but while Nupe society is hierarchical, that of the Tivs tends to be decentralized.

Densely settled areas

The distribution of population. The main concentrations of people are in the forest belt west of the Cross River and in the western half of the extreme north. In parts of Iboland and Ibibioland, the population density exceeds 800 persons per square mile (compared with the average national density of 154 persons per square mile). With the exception of Rwanda and Burundi, southeast Nigeria therefore constitutes the most densely settled area in Africa south of the Sahara. This concentration of agricultural people nevertheless occurs in a region which has heavily leached and impoverished soils, and a food-deficit consequently exists. Many migrants therefore leave the region to seek employment in the cities or in other rural parts of the country. The second region of dense population in the forest belt occurs in the cocoa-growing area of Yorubaland, which attracts many migrants from the congested districts of Iboland and Ibibioland. In the extreme north, there are also two areas of dense population—the Sokoto area and the Kano–Katsina area. The Kano concentration is based on intensive agriculture in an area of relatively fertile soils, but the densely settled areas around Sokoto and Katsina have somewhat impoverished soils and do not produce enough food for the local population. The average density in these northern areas is about 500 persons per square mile.

Smaller pockets of dense concentrations of people, averaging about 400 persons per square mile, occur in the tin fields of the Jos Plateau, in southern Tivland, and in the Okene district. The remaining, and by far the greater, part of the country is somewhat sparsely settled; vast areas of the middle belt, the Lake Chad Basin, and the Cross River district are virtually uninhabited. While most of these sparsely peopled areas suffered from extensive slave raids during the 19th century, there are some areas, such as the Niger Delta, the Cross River area, and parts of the middle Benue Valley, which, because of their difficult environment, have never been densely populated. The dense concentrations around Kano, Sokoto, and in parts of the cocoa belt occur, on the other hand, in areas that were protected by powerful chiefs and were therefore relatively peaceful during the period of the slave trade.

Demographic trends and migrations. Although the census figures are unreliable, there is sufficient evidence to show that demographic trends in Nigeria are similar to those in other developing nations. High birth and mortality rates are characteristic, although, during the last 30 years, there has been a decline in the rate of infant mortality and an increase in life expectation. There has consequently been a rapid growth in the population, which, reported as about 55,700,000 in 1963, was expected to increase to 70,000,000 in 1972. Except for the influx of a



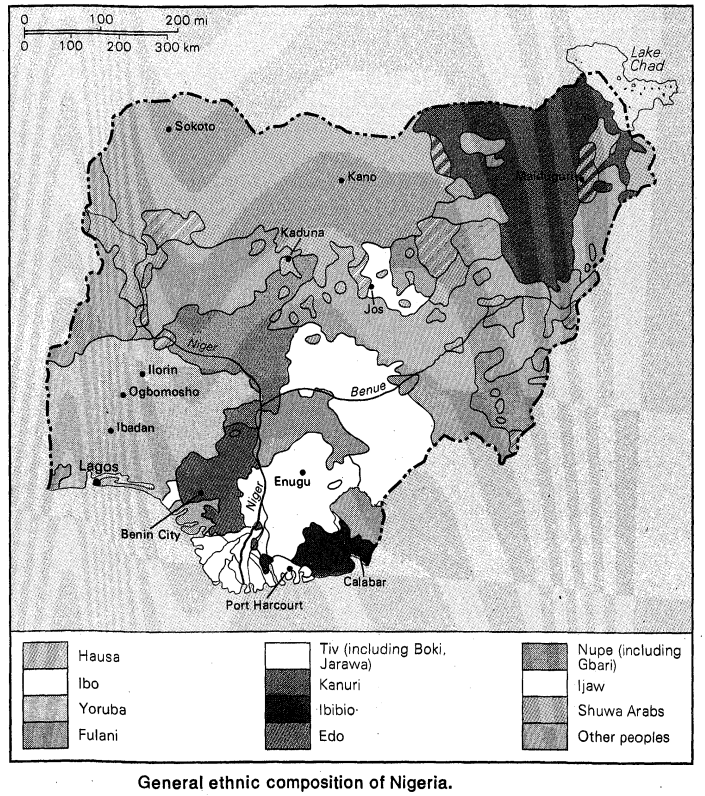
Population density of Nigeria.

small number of skilled workers from Europe or of traders from the Middle East, the growth in population has been by natural increase. A considerable movement of population nevertheless does take place within the country.

Internal migrations

Past census figures indicated that internal migrations took the form of a south to north movement of migrants who settled in the cities of the north as well as a north to south movement of seasonal migrants, from the Sokoto and Kano areas, who travelled to work in the cocoa-growing areas of Yorubaland. A larger number of people migrate westward from Iboland and Ibibioland as well as out of the Niger Delta. Most of these migrants work as labourers in the cocoa- or rubber-producing districts or as self-employed tenant farmers, cultivating food crops for sale to the neighbouring towns.

Many Nigerians also migrate to work in the neighbouring West African countries, such as Dahomey, Ghana, Equatorial Guinea, Cameroon, and even Sierra Leone. Before the deportation of many aliens from Ghana which began in 1969, about 1,000,000 Nigerians were living there. On the island of Fernando Po, however, which is part of Equatorial Guinea, Nigerians work on four-year contracts, at the expiration of which they normally return home. (R.K.U.)



General ethnic composition of Nigeria.

THE NATIONAL ECONOMY

From the economic point of view, Nigeria's most important characteristic is its size. Its population is the largest, and its gross domestic product the third largest (after that of South Africa and that of Egypt) in Africa. From 1967 until January 1970, the economy was disrupted by civil war. Until then, the 1960s had been a period of fairly rapid economic growth, with an annual average rate of increase in real gross domestic product of 4.5 percent in 1960–66—slightly faster than the average rate of increase for less developed countries as a whole. The growth of industrial production, from a small initial base, was particularly fast—an 8.5 percent annual average rate of increase, compared with about 6.5 percent for all less developed countries combined. Agricultural production, however, rose at a much slower rate—2 percent per year, or about the same as in other less developed countries. The rapid increase in exports, at about 9 percent a year, as against under 5 percent for less developed countries as a whole, was partly due to the expansion of production of crude oil, of which Nigeria was the world's 12th largest producer, being responsible for about 1 percent of the total output. Other world markets in which Nigerian exports are important are those for palm oil and palm kernels (of which Nigeria accounts for over a third of the total), cocoa, and groundnuts.

Despite a rate of growth that is rapid by the standards of other less developed countries, Nigeria still has one of the poorest economies in the world. Of the countries for which estimates were available in the mid-60s, only five —Ethiopia, Malawi, Tanzania, Upper Volta, and Burma —had a gross national product per capita that was lower than the $150 estimated for Nigeria.

Oil resources

Mineral resources. Oil, natural gas, coal, tin, and columbite are Nigeria's most important minerals. Proved reserves of crude oil are over 400,000,000 tons, with estimated reserves amounting to at least another 400,000,000 tons. The discovery of oil and its production in commercial quantities dates only from the late 1950s, although prospecting had been going on for many years before that. All of the reserves are in the southern part of the country (south of a line through Benin, Owerri and Calabar), with most of the output at the beginning of the 1970s coming from onshore fields in the Niger Delta area. Natural-gas reserves, both associated and unassociated with oil, are large, but most of the gas produced has to be flared for lack of a market. Coal reserves are estimated at over 350,000,000 tons, 110,000,000 of which are in the East-Central State around Enugu, the remainder being in the northern states. As a fuel, Nigerian coal is unlikely to be competitive with local oil in the long run. Extensive deposits of lignite (an imperfectly formed coal) also exist in the south, but technical exploitation difficulties exclude their economic use. In the northern and eastern states there are deposits, as yet unexploited, of tin as well as of columbite (a black crystalline mineral used to make niobium, which is used to make alloy steels).

Biological resources. Approximately a quarter of Nigeria's total area of over 357,000 square miles (924,000 square kilometres) is in use as arable land and under permanent crops, about 24 percent consists of permanent meadows and pastures, and about 34 percent is forested. Thus, in relation to the size of the population, there can hardly be said to be a shortage of land, although the prevalence of the shifting cultivation system means that such a shortage is emerging in some areas. This system has the merit, however, of protecting the soil against both erosion and loss of fertility. Most of the commercially exploited timber resources are in the Benin lowlands immediately to the northwest of the Niger Delta, the

main types being obeche (a whitish hardwood, used for furniture making), abura (a pinkish-brown wood, used for veneers), and mahogany.

Hydroelectric resources. There is a widespread network of rivers, many of which are potential sources of hydroelectric energy. At the end of 1966, total installed generating capacity was 417 megawatts, of which only 21 megawatts was hydroelectric, from stations in the Jos plateau. Since 1968, however, the Kainji power station has been in operation, with an initial capacity of 320 megawatts, eventually to be raised to 960 megawatts. With additional stations planned at Jebba and in the Shiroro Gorge, total hydroelectric capacity could exceed 1,800 megawatts during the 1970s.

Agriculture, forestry, and fishing. Agricultural production, including forestry and fishing, although still accounting for over half of the gross domestic product, has nevertheless represented a declining proportion of it in recent years. Since over 70 percent of the active population is still engaged in agricultural occupations, it is evident that output per man in agriculture is well below that in other sectors of the economy. Produce grown mainly for domestic consumption (most of it not marketed) is worth about five times that grown for export. The main food crops grown in the southern states are yam, cassava, and cocoyam, and in the northern states maize, guinea corn, rice, millet, cowpeas and cattle. The range of export crops is wide. The eastern states concentrate on producing palm oil and kernels, the midwest on rubber, the west on cocoa, and the north on groundnuts, cotton, and hides and skins. In the past, a marketing board in each of the regions has handled the main export commodity and guaranteed prices to producers, while a central marketing board has been responsible for exports—selling on behalf of the regional boards. With the creation of new states or of new state boundaries the marketing system had not been fully readjusted by the beginning of the 1970s. Interim arrangements, however, preserved the essentials of the old system. Output of all crops has tended to be variable, but, in general, output of export crops appears to have been growing faster than that of local food.

Crops for export

The output of cocoa, groundnuts, and rubber has expanded particularly fast in recent years. In 1966—the year preceding the civil war—the following quantities in short tons were exported: cocoa, 190,000; palm kernels, 394,000; palm oil, 143,000; groundnuts, 573,000; groundnut oil, 104,000; groundnut cake, 133,000; rubber, 70,000; and cotton (mostly for the local textile industry), 15,000. Of these, palm kernels, palm oil, and rubber were the only crops to be significantly affected by the civil war, and by the early 1970s output of these commodities was again substantial. The timber industry also suffered from the war, but even before that output had been falling away from the high level of production of the early 1960s, because of destruction of forests for the purpose of temporary cultivation. Forest products amount to about 4 percent of gross domestic product; another 3 percent is accounted for by fisheries. Lake Chad is the main source of fish, yielding about 25,000 tons per year. Over 80 percent of the domestic demand for fish is met by imports, mostly from Iceland and the United States.

Mining and quarrying. Mining represents the fastest growing sector of the Nigerian economy, accounting for over 5 percent of the gross domestic product, compared with less than 1 percent in the late 1950s. The most spectacular increase has been in oil, the output of which amounts to over 20,000,000 tons. The oil industry was seriously affected by the civil war, but recovery has been fast, with output exceeding 1,000,000 barrels a day early in 1970. The oil refinery near Port Harcourt, which has a capacity of 45,000 barrels per day and which was supplying nearly all the domestic demand before the civil war, came back into operation in 1970, after war damage had been repaired. Even before the civil war, the output of coal, used mainly by the railway and a cement company, and for electricity generation, had been declining. The tin industry, which was less affected by the war, produces ore, practically all of which is smelted locally and exported as metal. Production is about 13,000 tons. Columbite output, on the other hand, which reached a peak of 2,500 tons in 1965, has been declining because of a fall in world prices; production is about 1,000 tons a year. The mining industry is virtually entirely foreign owned.

Manufacturing. While the manufacturing sector has grown rapidly in recent years, it still accounts for only about 5 percent of Nigeria's gross domestic product. The largest production is of consumer goods for the home market; unlike the rest of the economy, manufacturing was stimulated by the civil war and by its restrictions on imports. Major industries in terms of value of output are textiles, beer, food, tobacco, and vegetable oils. Companies engaged in manufacturing are of all sizes; categories range from rural cottage industry to small-scale urban industry to large modern industrial plants. Over two-thirds of the last category is foreign owned. Most of the remainder is owned by public authorities; only a small proportion is privately owned by Nigerians.

Energy. With its oil, natural gas, and coal deposits, and hydroelectric resources, Nigeria has a wide range of potential sources of energy. Output of electricity in 1970 was 1,549,200,000 kilowatt-hours. The discovery and exploitation of the oil and gas reserves is still to be fully reflected in the pattern of energy consumption, which —before the war—was: coal, 21 percent; petroleum products, 69 percent; natural gas, 8 percent; and hydroelectricity, 1 percent.

Financial services. The Central Bank of Nigeria, founded in 1959, performs central banking functions. There are 11 commercial banks. The Nigerian Industrial Development Bank provides medium- and long-term finance to private firms. Since the war, during which the financial sector tended to be dominated by government borrowing requirements, prospects for normal operations have improved.

Foreign trade. The value of Nigeria's domestic exports more than doubled in the ten years before the war, reaching a total of more than $780,000,000 in 1966. In the same period the ratio of exports to gross domestic product rose from 15 percent to 17 percent, due to foreign exchange earnings from the growing oil industry. By 1966 this commodity, which earned nothing a decade earlier, had become the most important single export, contributing 33 percent of total export earnings. Next in importance were groundnuts (15 percent), followed by cocoa (10 percent), palm kernels (8 percent), rubber (4 percent), palm oil (4 percent), and groundnut oil (4 percent). This pattern of exports, disturbed by the war, was re-establishing itself in the early 1970s.

Imports grew more slowly in the ten years before the war, rising by about 67 percent to a total of $717,000,000. Nigeria thus moved in this period from a position of chronic trade deficit at the outset to one of a potentially regular surplus. The war destroyed this favourable position, but only temporarily. The war also changed—perhaps more permanently—the pattern of imports, with chemicals (over 10 percent of the total) and machinery (about 9 percent) displacing food and cotton piece goods at the top of the import list. The United Kingdom, still Nigeria's main trading partner, in 1970 bought 30 percent of Nigeria's exports, and supplied 29 percent of its imports.

Management of the economy. *The private sector.* The Nigerian private sector is numerically dominated by cash-crop farmers and small-scale local businessmen, operating in both urban and rural areas. So far, however, few Nigerian businessmen have been successful in building up large industrial firms, most of which are owned by foreign private investors or public authorities.

The public sector. Government accounts for about 5 percent of the gross domestic product, even when its directly economic activities, such as manufacturing, are excluded. Apart from manufacturing projects (which represent a total investment of over $98,000,000), the government owns a shipping line, an airline, sporting facilities, broadcasting stations, and hotels. Of these only the shipping line has operated at a profit. The public authorities also provide incentives to private firms, including tariff protection, import duty relief on materials,

The role of government

accelerated depreciation allowances, and some relief from income tax as well as assisting the growth of private industry by providing credit, industrial estates, and technical assistance.

Taxation. Taxation arrangements are complicated by the federal constitution, according to which part of the revenue collected by the federal government has to be redistributed to the states. The allocation of revenue between the states and the federal government, and between the states themselves, is a constant bone of contention; at the beginning of the 1970s a satisfactory accommodation still had not been arrived at. The most important source of all governmental revenue is import duties, which account for about a third of the total, followed by excise duties (under 20 percent) and mining royalties and rents (about 15 percent). Personal income tax accounts for less than 10 percent, and company tax for about 5 percent, of total revenue. Fiscal policy has been orthodox, with budgets being balanced in the early 1970s. Public finances were, however, severely strained during the civil war.

Trade unions and employer associations. The Nigerian trade-union movement is weak, in spite of having a membership equivalent to about 65 percent of all those employed in establishments with ten or more employees. Membership is spread throughout hundreds of small unions, and there is no single effective central organization. The largest employers of wage labour are the government and public corporations, which fix wage rates unilaterally. The larger private employers have formed the Nigerian Employers Consultative Association, in the hope of encouraging the growth of a collective-bargaining system.

Contemporary economic policies. Before the civil war, there were virtually no restrictions on imports or foreign exchange remittances, and generous incentives were offered to foreign investors. With the onset of the war, this situation inevitably changed. Import restrictions and exchange controls were imposed, severe legislation against strikes was enacted, and efforts were made to raise extra tax revenue from private companies as well as by imposing indirect tax increases. At the same time a new Companies Act required all foreign companies to place their local operations under subsidiaries incorporated in Nigeria, and a new petroleum decree gave the government wide discretionary powers in the oil industry. The change of emphasis in general economic policy engendered by the war is likely to be permanent.

Problems and prospects. Among Nigeria's current economic problems are a high rate of population growth, a growing problem of unemployment (particularly among students leaving school), a potential shortage of local food, and an accelerating rate of price inflation. Other problems stem directly from the three-year war that ended at the beginning of 1970. Among these, the relief and rehabilitation of the areas affected by the war are the most pressing. The huge internal debt built up during the war will also constitute a drain on public finances for several years. That the prospects for the Nigerian economy in the 1970s are nevertheless good is due partly to the large market it offers to potential investors, partly to the stimulation of economic development resulting from restrictions on imports during the war, and, most of all, to the expansion in the oil industry. Earnings from oil could end the foreign exchange shortage at a crucial stage in Nigerian development. This hopeful prospect assumes continuing political stability in a country in which many potential sources of tension remain. (E.I.U.)

TRANSPORTATION

The north–south traffic flows

Patterns of transport flow. The general pattern of transport flow is in a north–south direction, running from the interior to the southern seaports. Construction of main railroad lines started from the ports of Lagos (1898) and Port Harcourt (1916) and continued inland to the Sudanese commodity collection centres—reaching Kano in 1912, Nguru in 1930, Kaura Namoda in 1929, and Maiduguri in 1964. The major roads also run north–south. In the resulting competition between rail and road traffic, road traffic is gaining the advantage. The north–south pattern of flow is characteristic of the colonial type of economy, in which raw materials are produced in the interior and are then shipped out through the nearest port to the metropolitan country, which in turn supplies manufactured goods. The flow pattern also results from the fact that different agricultural products are grown in zones, or belts, running east to west, so that the general direction of internal exchange trade is from north to south. This pattern is modified in the south, however, where local staples flow from the eastern states and the Midwestern State to Ibadan and Lagos in the southwest. Feeder roads, running in an east–west direction, also link up districts to the main north–south railways and roads.

On several occasions the country's single-track railroads have proved incapable of transporting large quantities of groundnuts and cotton from the north, a circumstance which has stressed the growing importance of huge haulage, even over such long distances as from Kano to Lagos. The greater flexibility of road transport also appeals to importers, most of whom ship their goods by road. The proportion of imports moved from Lagos by road has, for example, increased from one-fifth to over a half in recent years, while the proportion of imports moved by rail has declined sharply.

The average daily traffic flow on the roads is greatest in the cocoa belt of southwestern Nigeria, after which it is greatest in the groundnut and cotton belt of the Kano–Katsina region, in the Jos Plateau tin fields, and in the palm belt of southeastern Nigeria. These are the four main areas of economic development and of population concentration in the country, and each is consequently served by a dense network of all-weather roads. By contrast, the relatively unproductive and sparsely settled areas of the middle belt, of the Cross River region, and of the Chad Basin have tenuous road links that carry only a few trucks a day.

Component systems of transportation. Although the development of the railway and of roads have overshadowed inland waterways, the creeks and rivers were the first means of communication in 20th-century Nigeria. The most important waterways are the Niger and Benue rivers, which still carry substantial quantities of goods, including transit cargo to and from Garoua in the Federal Republic of Cameroon. The Cross River is used to ship exports to Calabar, but, like other rivers in the country, it is not navigable during the dry season when the water level falls. Passenger and cargo boats also operate on the lagoons and on the numerous creeks that occur along the Nigerian coast from Lagos to the Cross River estuary.

The railroad system

Although the road is now the most important means of transportation, it was the railroad that formerly was most relied upon. The railroad system consists of two single-track main lines—the eastern line from Port Harcourt to Maiduguri and the western line from Lagos to Kano. Branch lines connect the western main line to Kaura Namoda, to Nguru, and to Baro on the Niger. The total route mileage amounts to about 2,680 miles (4,313 kilometres) of three-foot six-inch (1.067-metre) track. Since 1960, tracks have been relaid with heavier rail to permit heavier axle loads and higher speeds, rail movements have been speeded by improvements in signalling, and steam engines have been replaced by diesel locomotives, which can haul heavier loads. Diesels also require less maintenance and servicing and, since they do not need to pick up water, save running time; they are particularly useful in the drier north where services were often disrupted by water shortages during the dry season.

Nigerian roads fall into three categories; these are trunk A roads, which are maintained by the federal government and which link Lagos with the state capitals; trunk B roads, which are maintained by state governments and which connect provincial capitals and other large towns with the trunk A system; and other roads, which are maintained by local government, which carry local traffic, and which act as feeders to the trunk-road systems. All trunk A and most trunk B roads are surfaced; almost all other roads are earth roads.

The southern terminals of both the two main-line railways, as well as of the north–south trunk roads, are the

port towns of Lagos and Port Harcourt. These are Nigeria's main international seaports, the less important ones being Warri, Sapele, Koko, and Calabar. Lagos and Port Harcourt both have modern equipment for handling cargo. Both are administered by the Nigerian Ports Authority, which was established in 1954 and which has recently (1970) taken over responsibility for the port installations and for the administration of the other seaports. Shore labour in Nigerian ports is, however, supplied by private contractors.

All the 12 state capitals are served by air transport that is supplied by the Nigeria Airways Corporation. A few minor airfields also exist in some provincial cities, but only two international airports—those at Lagos and at Kano—handle transcontinental traffic. The creation of more states in 1967 resulted in an increase in air services beween Lagos and the state capitals. While air fares are relatively high, Nigerian Airways has nevertheless maintained flights from Lagos to cities such as Kano and Kaduna at fares that are less than the corresponding first class rail fare.

ADMINISTRATION AND SOCIAL CONDITIONS

The structure of government. *The constitutional framework.* Although the two British protectorates of Northern and Southern Nigeria were amalgamated in 1914, the government of the two territories remained essentially different until 1946, when the first legislative council for the entire country was established. This situation arose largely as a result of cultural differences as well as of wide variations in the degree of development of political institutions among the many ethnic groups in the country. Thus, the north, much of which had a Hausa—Fulani administration in pre-British days, was administered by indirect rule—a system whereby the British governed through established feudal overlords. The south was directly administered by the British. The 1946 constitution provided for a central legislature for the whole country as well as for three Regional Houses of Assembly —one for each group of provinces. This was in accordance with the thinking of Nigerian politicians, who became prominent after World War II, demanding independence and a federal form of government with strong powers vested in the states.

The 1951 federal constitution, however, gave only limited powers to the states, until, three years after its adoption and as a result of mounting pressures from the states, residual powers were transferred from the federal government to the states. By the modified federal constitution of 1954, the states became responsible for direct taxes (including income tax), school education, health, and most aspects of economic development. The weakened federal government retained responsibility for defense, external affairs, aviation, railways and some key roads, postal services, higher education, customs and excise, and banking. The federal government was, however, permitted to assume more powers in time of war or whenever any state threatened the continued existence of the federation.

When the army assumed power in January 1966, the legislative sections of the constitution were suspended, and the political structure of the country transformed by the creation in 1967 of twelve states, instead of four. With the end of the 30-month civil war in January 1970, a return to civilian rule is eventually expected. It is anticipated that a federal constitution will be maintained.

State and local governments. When, in 1960, the United Kingdom granted independence to Nigeria, Nigerian politicians inherited an unwieldy federation in which one state—the Northern Region—had a much greater size and population than the other two states combined. Before independence there was, therefore, a sustained demand by minority groups within the federation for more states. Although, however, the 1954 constitution provided for the creation of more states, none of the state governments was willing to "dismember" its territory. The creation of the Mid-Western State out of Western Nigerian territory in 1963 followed a major political crisis in that part of the country, during which the federal government invoked its emergency powers. It was not until May 1967, on the eve of the Biafran secession, that the federal military government created 12 states in place of the previous four.

Under the civilian regime that ended in 1966, the federal government consisted of a Council of Ministers presided over by the prime minister, a Senate, and a House of Representatives. Legislative powers were vested in the Senate and the House, while executive powers were vested in the Council of Ministers. The head of state was the president, who was elected for a five-year term, after which he could be re-elected. Each region, or state, had a similar constitution, in which there was a governor; an Upper House, or House of Chiefs; and a regional House of Assembly. At the local government level, there was considerable variation from one state to another. In the feudal Northern Region, the state was divided into emirates, in each of which the emir, a traditional ruler, presided over the local council. In the Western Region, where the Yoruba people had established centralized administrations in precolonial days, each local government unit was ruled by an *oba* and his chiefs. Administration in the Eastern Region was, however, of a totally different kind. As the sphere of influence of a chief rarely extended beyond the territory of a village, the region was constituted in provinces, each of which had an elected Provincial Assembly, presided over by a provincial commissioner appointed by the state government.

Since the assumption of power by the army in 1966, the country has been ruled by the Supreme Military Council, consisting of the commander in chief of the armed forces as head of state and chairman, 12 military governors, and the heads of the army, navy, police, and air force. There is also a Federal Executive Council, presided over by the commander-in-chief General Gowon and made up of civilian representatives, one from each of the 12 states. The representatives, who are designated commissioners, are appointed by the commander in chief. Each state is governed by an Executive Council, which also consists of civilian commissioners appointed by the state governor, who is also chairman of the council. Local council areas are each administered by a single administrator, appointed by the state governor.

The political process. *Elections.* Political activity at all levels was banned when the army came to power, but during the previous civilian regime, election to the federal and state legislatures, as well as to local councils, was on the basis of one man one vote. The franchise was extended to all adults, except in the Northern Region, where women were not allowed to vote. No government party ever lost in any election. To maintain itself in power many governing parties, both at the federal and at the state level, resorted to fraudulent practices and to the victimization of political opponents, many of whom were imprisoned. The life of the federal and state legislatures was five years.

Parties. At the time of independence in 1960, and until the end of civilian rule in 1966, there were three main political parties in the country. These were the National Convention of Nigerian Citizens (NCNC), founded by Nnamdi Azikiwe; the Action Group (AG), a Yoruba party founded by Chief Awolowo, a former lawyer and trade unionist; and the Northern Peoples Congress (NPC), founded by the late Sir Ahmadu Bello, the *sardauna* of Sokoto. Of all the parties, only the NCNC had some semblance of being a national party. In 1965 it controlled the governments of Eastern and Mid-Western Nigeria as well as the Lagos City Council in the federal capital. It had a strong following in the Yoruba west, but, like other Nigerian parties, was identified with an ethnic group—in this instance the Ibos of Eastern Nigeria. The most conservative party was the NPC, which, as the name implies, was a Northern Nigerian party. Thus, when in 1954 and 1959 the NCNC and the AG contested all seats in the federal legislature, the NPC restricted itself to Northern Nigeria, and—because of the large size of the north as well as because of fraudulent election practices—was able to control not only the north but also the federal legislature.

The former political parties

In each state there were a number of small parties, which were usually in opposition to the state government. Thus, in the Northern Region there were the Northern Elements Progressive Union (NEPU)—led by *mallam* (a title signifying "Muslim legal clerk") Aminu Kano, a Fulani teacher—and the United Middle Belt Congress; in the Western Region there was the Nigerian National Democratic Party, led by the late Chief Samuel Akintola, a lawyer and former Action Group leader; and in the Eastern Region there were the United Independence Party (UNIP) and the Dynamic Party.

Participation of citizens. Except in the north, where the influence of tradition was greater and education less widespread, Nigerians showed much interest in postindependence politics. Political consciousness was particularly marked in the towns and cities of the southern states. The large-scale involvement of citizens in politics resulted primarily from the fact that, since Nigeria is a developing country, people looked to the government for jobs as well as for basic services such as roads, water, schools, and hospitals.

Justice. The Nigerian legal and judicial system is subject to regional variations, particularly between the largely Muslim north and the Christian-influenced south. Three codes of law may be recognized in the country; customary law, Nigerian statute law, and English law. Customary laws are administered by native, or customary, courts. Such courts are usually presided over by persons with no formal legal education excepting in the case of the Alkali, or Muslim, Courts where the judges (*alkalis*) have to undergo formal training in Muslim law. Customary courts are of different grades, with Grade A courts serving as courts of appeal for cases from Grades B, C, and D courts. During the postindependence period of civilian rule, these courts fell into disrepute through their use as instruments of political coercion.

Three codes of law

Nigerian statute laws include much of the legislation enacted by the British colonial administration, most of which has since been revised. State legislatures may pass laws on certain matters not included in the Exclusive Legislative List, which is composed of such subjects as defense, external affairs, and mining—all of which can only be legislated upon by the federal government. It is also the practice that, whenever federal legislation on any matter is in conflict with state legislation, federal law prevails in the courts. In addition to Nigerian statutes, English law is used in the high courts and in magistrate's courts. There is a high court for each state, but one Supreme Court for the federation.

The armed forces. The armed forces of Nigeria consist of the army, the navy, and the air force. Before the civil war, Nigeria had a ceremonial army of about 7,000, most of whose officers were killed during the military coups that occurred in January and July 1966 respectively. The war to end the Biafran seccession led to the enlistment of thousands of people and to heavy expenditure on military equipment for all sections of the armed forces. When the war ended in January 1970, an estimated 200,000 persons were bearing arms. Since the government does not intend to demobilize the army, it appears that Nigeria will maintain the largest army in tropical Africa. Both the navy and the air force also grew considerably during the war, but further expansion is not envisaged.

Administration. *Educational services.* Education is largely the responsibility of the state governments although there are two federal universities at Lagos and Ibadan. The University of Nigeria at Nsukka and Enugu, the Ahmadu Bello University at Zaria and Kano, and the University of Ife are owned by state governments. Each of these universities has a medical school attached to it. Another state university was opened at Benin City in 1970. In the 1969–70 academic year there were about 10,000 students attending university institutions. About 20,000 Nigerian students are attending United States or European universities.

Primary education is free in some states, including the Western State and Lagos State as well as North-Western State. In the northern states, where the literacy rate is lowest, secondary education is also virtually free. In general, formal education is less sought after in the north. Teachers are trained not only in the universities but also in several advanced teachers colleges (maintaining higher standards) and in numerous other training colleges.

Until 1950, most schools were controlled by religious bodies, such as the American Baptist Mission, the Wesleyan Methodist Mission, the Church of Scotland Mission, and the Anglican Church's Church Missionary Society. A few government schools were also maintained, although government policy was to give grants to mission schools rather than to expand its own. After independence, many more secondary schools were established by local council authorities or by the legislatures. There is a shortage of teachers of all kinds, which the federal government is attempting to remedy by training teachers in the universities.

Health services. Medical and health services are the responsibility of the state governments, each of which maintains a hospital in the larger cities or towns. Specialized hospitals are located in Lagos, Ibadan, Jos, Port Harcourt, and Kano. Many maternity hospitals are run by religious organizations. Medical services are generally inadequate; many hospitals do not have enough medical personnel. Drugs are always scarce. Rural areas are the most deprived of services, with most communities being served only by dispensaries and maternity homes, staffed by junior personnel.

There are about 2,180 practicing doctors in Nigeria, an average of about one for every 29,000 people. It is hoped that when the five medical schools in the country are in full service, they will, from 1975 onward, produce a total of 400 doctors a year, as compared to about 120 in 1970.

Housing. Housing is inadequate in the cities, where overcrowding has led to the spread of slums. Since there are only a few housing corporations, most houses are built by individuals, who acquire plots to build houses either by direct or by contract labour. Each city has three types of residential districts; the local Nigerian town, with congested and poor quality housing; the Nigerian strangers' town (usually called the Sabon Gari), equally congested and occupied by Nigerians from other states or regions; and a district of modern and expensive housing, occupied by the Nigerian professional elite. In the villages, most people live in huts, but, especially in the cocoa belt, modern houses are now being built.

Police services. The Nigeria Police Force, established by the federal constitution, is headed by the inspector general, who is represented in each state by a police commissioner. On the eve of independence the organization of the police was a controversial issue between some politicians who wanted each state to control its police force and others who advocated federal control. The matter was settled by a compromise, whereby the police force remained united under the inspector general but with the state governments being allowed to appoint their police commissioner. Each commissioner is responsible to the inspector general but has to comply with the directions of the state government.

During the period of civilian rule, the Native Authority Police, in Northern Nigeria, and the Local Government Police, in Western Nigeria, were sometimes open to criticism because they were far below the national police force in training and standards. The Federal Military Government followed a policy of gradually absorbing local police forces into the Nigeria Police Force.

Social conditions. *Wages and cost of living.* The gap between the rich and the poor is far greater than in industrialized countries. Most wage earners in the country earn less than $700 a year, while lawyers, university teachers, engineers, and doctors earn over $3,000. There is a general sentiment that the wages of the poor are too low; in consequence, the federal government has established a wage review board. A price control board was also established to deal with the problem of sharp price increases that occurred during and after the war. Since the last general salary review in 1959, the cost of living has increased so much that many families find it difficult to eat adequately, particularly during the hungry-season

months from March to July before the first harvests are gathered, when local staple food prices are very high.

Health conditions. The concentration of people in the large pre-industrial cities such as Ibadan, Oshogbo, Zaria, and Kano has created enormous sanitary problems concerning sewage disposal, water shortages, and poor drainage. Medical science has brought about improvements in health conditions, but many people still die from malaria, cerebro-spinal meningitis, and other preventable diseases.

Rural communities in particular suffer greatly from inadequate or impure water supplies. Some villagers have to walk distances of up to six miles to the nearest water point—usually a stream. Since people wash clothes, bathe, and fish in the same stream, the water drawn by anyone living in villages further downstream is often polluted. During the dry season, wayside pits containing rain water are used until they dry up. Cattle are also often watered in such pools, a circumstance that contributes to the high incidence of intestinal diseases and guinea worm in many rural areas.

CULTURAL LIFE AND INSTITUTIONS

Nigeria has a rich and varied cultural heritage, deriving partly from the varied racial elements in the population and partly from the influence of Middle Eastern and western European cultures. The oldest works of art so far discovered consist of late Stone Age terra-cotta heads associated with the Nok culture, which flourished in the region of the Jos tin fields from about 500 BC to about AD 200. The heads are considered to be of a technical standard indicating an advanced agricultural culture connected with iron and tin. Nok culture is thought to have influenced the celebrated bronze and terra-cotta heads of Ife and of ancient Benin. Ancient art works of wood and bronze have also been excavated at Awka, near Onitsha, as well as in the coastal areas of Oron. During the last two centuries, the cultural life of the south has been influenced by Europe, while Arabic influence has been manifest in the north.

The Nok culture

Nigerian arts and traditions have revived since independence partly because of the realization of the desirability of preserving Nigerian culture and partly because of the patronage that Nigerian arts have received from abroad. Carved calabashes from Oyo, masks and ebony heads from Benin City, Awka, or Ikot Ekpene, or thorn carving from Shagamu are used to decorate the houses of the well-to-do, who also wear locally woven and locally dyed cloths, instead of, as in the past, using imported materials. Oil paintings of Nigerian subjects are also common.

The Institutes of African Studies at the Universities of Ibadan and Ife have done much to publicize and re-awaken interest in traditional folk dancing and in poetry, as also have the School of Fine Arts and the School of Drama at Zaria and Ibadan respectively. Vernacular radio and TV programs in at least ten languages include traditional music and dancing, folk operas, and story telling. Since, except in the Muslim north, writing became common only during the last 60 years, and since until recently few Nigerians showed any interest in folk traditions and local culture, it is believed that much of the country's culture has perished during the last two generations. Today, many ancient folk songs have been revived by popular singers who use modern musical instruments to produce sounds that villagers can hardly identify with the songs they inherited from their ancestors.

Cultural institutions. In Nigeria, where superstition still waxes strong, many cultural institutions touch upon various aspects of life. Secret societies, such as Ekpo and Ekpe among the peoples of the South-Eastern State, were formerly used as instruments of government, while other institutions were associated with matrimony. According to the Fulani custom of *sharo* (test of young manhood), rival suitors underwent the ordeal of caning as a means of eliminating the less persistent grooms-to-be, while in Ibibio territory, girls were confined for several years in bride-fattening rooms before they were handed over to their husbands. These and other customs were discouraged by colonial administrators and missionaries, who disapproved of them. Some of the more adaptable cultural institutions have, however, been revived since independence; these include, for example, the EKPO and EKONG societies for young boys in parts of the South-Eastern State.

Press and broadcasting. In addition to the radio networks of the Nigerian Broadcasting Corporation (NBC), owned by the federal government, there are several state-owned broadcasting stations, of which the most prominent is the Western Nigeria Broadcasting Service. The Lagos NBC studios broadcast news in English and in nine Nigerian languages, as well as music, while local NBC stations concentrate on local culture and languages. Television broadcast is from Lagos, Ibadan, Enugu, Port Harcourt, Aba, and Kaduna. Most programs are composed of films from the United States and western Europe, although local drama groups also feature prominently.

There are several national papers, all of which are published in English. The most widely read daily papers are the *Daily Times*, the *Nigerian Morning Post*, and the *New Nigerian*. Regional papers include the *Midwest Echo*, the *Daily Sketch*, and the *Daily Standard*.

CONCLUSION

The 30-month civil war over the attempted Biafran secession proved a serious, although temporary, setback to the development of tropical Africa's most populous country. It has encouraged Nigerian self-reliance and has stimulated the establishment and expansion of Nigerian manufacturing industries, as well as of raw-material processing. In the aftermath of the war, much still remains to be done to reconstruct war-damaged areas, as well as to rehabilitate millions of people. The greatest task, however, appears to be that of reconciling Ibos and other ethnic groups inhabiting the former secessionist territory. These other groups, now constituted into the Rivers and South-Eastern states, were subjected to pressures from the Ibos, who wished them to support the Biafran cause.

Economically, the outlook is bright. The war ended with little national debt, thanks to Nigeria's production of crude oil, which financed the prosecution of the war. Soon after the end of the war, oil production exceeded prewar figures.

BIBLIOGRAPHY. BRITISH COLONIAL OFFICE, *The Nigeria Handbook* (1953), useful information on the history of British Nigeria; K.M. BUCHANAN and J.C. PUGH, *Land and People in Nigeria* (1955), useful information on systems of farming, diseases, racial composition of the population, and climate, but outdated on other topics; M. CROWDER, *The Story of Nigeria* (1962), a history of Nigeria from the rise of the Sudanese states, and of Ife and Benin to 1960; K.O. DIKE, *Trade and Politics in the Niger Delta, 1830–1885* (1956), a detailed account concerning the trade in slaves and later in palm oil; B. FLOYD, "The Federal Republic of Nigeria," *Focus*, vol. 15, no. 2 (Oct. 1964), an informative report on various geographical aspects to 1962; D. FORDE and R. SCOTT, *The Native Economies of Nigeria* (1946), a detailed account of the economies of some of the main ethnic groups up to the end of World War II; G.K. HELLEINER, *Peasant Agriculture, Government, and Economic Growth in Nigeria* (1966), a fairly up-to-date account; A.L. MABOGUNJE, *Urbanization in Nigeria* (1968), detailed studies of Lagos and Ibadan, but little on the new towns of the east and north; *Nigeria Year Book* (annual), information on politics, trade, and business, with biographies of public figures included; NORTHERN NIGERIA MINISTRY OF TRADE AND INDUSTRY, *The Industrial Potentialities of Northern Nigeria* (1963), a guide for prospective investors, with useful information on infra-structure and development projects; H. ROBINSON *et al.*, *The Economic Coordination of Transport Development in Nigeria* (1961), a report prepared for the Joint Planning Committee of the National Economic Council of Nigeria by experts from Stanford Research Institute; S.S. RICHARDSON, "The Courts and the Legal System," in L.F. BLITZ (ed.), *The Politics and Administration of Nigerian Government* (1965), an explanation of differences in the character of the native or customary courts of the four pre-civil war states; R.K. UDO, *Geographical Regions of Nigeria* (1970), a current work on Nigeria's geography, "Disintegration of Nucleated Settlement in Eastern Nigeria," *Geogrl. Rev.*, 55:53–67 (1965), a study of the changing pattern of settlement and land use in the densely populated districts of southeastern Nigeria.

(R.K.U.)

Niger River

The Niger, with a length of nearly 2,600 miles (4,200 kilometres), is the principal river of West Africa, and the third longest in Africa. It is believed to have been named by the Greeks. Along its course it is known by several names. These include the Joliba (a Malinke [Mandingo] word meaning "great river") in its upper course; the Mayo Balleo and the Isa Eghirren in its central reach; and the Kwarra, Kworra, or Quorra, in its lower stretch.

Courses of the Niger and its tributaries

Physiography. The Niger rises in Guinea at 9°05′ N and 10°47′ W on the landward side of the Fouta Djallon highlands, only 150 miles from the Atlantic Ocean. Issuing as the Tembi from a deep ravine 2,800 feet above sea level, it flows due north over the first 100 miles. It then takes a northeast course, during which it receives its upper tributaries, the Mafou, the Niandan, the Milo, and the Sankarani on the right bank and the Tinkisso on the left bank. Seven or eight miles below Bamako, in Mali, the Sotuba rocks mark the end of what may be considered the upper river. Over the next 40 miles the Niger drops more than 1,000 feet into a valley formed by a tectonic subsidence. The Sotuba and Kénié Rapids both occur in the course of this descent.

A little lower down is Koulikoro, from which point the river takes a more east-northeasterly direction and its bed becomes fairly free from impediments for the next 1,000 miles. At Mopti it is joined by the Bani, its largest tributary on the right bank, after which it enters a region of lakes, creeks, and backwaters. These lakes are chiefly on the left bank and are connected to the river by channels that experience seasonal changes in the direction of flow. At high water most of the lakes become part of a general inundation. The largest of the lakes is Lac Faguibine, which is nearly 75 miles long and 15 miles wide, and more than 160 feet deep in places.

The labyrinth of lakes, creeks, and backwaters comes to an end at Kabara, port of Timbuktu (Tombouctou). Here, the river turns almost due east, passing its most northern point at latitude 17°5′ N, where it is bordered by the Sahara on its left bank. Near Taoussa, 250 miles downstream from Timbuktu, a rocky ridge that obstructs the course of the river is pierced by a defile (narrow gorge) more than a mile long, with an average width of about 800 feet, and a depth of over 100 feet in places. At low water the strong current here endangers navigation. Further downstream the river widens considerably, flowing down to Gao in a southeasterly direction across a floodplain three to six miles wide.

Except for the stretch from Bamako to Koulikoro, which is an area of rapids and steep descents, the middle course of the Niger River is navigable to small craft during high water as far as Ansongo—1,057 miles in all. Below Ansongo, 430 miles downstream from Timbuktu, navigability is interrupted by a series of defiles and rapids. The river becomes navigable to small vessels again at Labbezenga, Republic of the Niger, and continues navigable to the Atlantic Ocean. Navigation is seasonal, due to water level fluctuation in the rainy and dry seasons.

Downstream from Jebba, the Niger enters its lower course, flowing east–southeast through a broad and shallow valley five to ten miles wide. About 70 miles from Jebba it is joined by the Kaduna River, an important tributary that contributes about 25 percent to the annual discharge of the river below the Niger–Kaduna confluence. At Lokoja, in Kwara state, Nigeria, the river receives the water of its greatest tributary, the Benue, thereby approximately doubling the volume of its annual discharge. At their confluence the Niger is about three-quarters of a mile wide, and the Benue over a mile. Together they form a lake-like stretch of water about two miles wide and dotted with islands and sandbanks. From Lokoja downstream to the town of Idah, the Niger runs between hills, and the channel is narrow and rocky. Between Idah and Onitsha the banks are lower, and the country flatter. At Onitsha, the largest town on the Niger's banks, the valley narrows as the river flows south through what is probably a fault in the area's sandstone. It emerges at Aboh, separating into many branches before reaching the ocean via Africa's largest delta.

The Niger Delta

The Niger Delta, which stretches for nearly 150 miles from north to south and spreads along the coast for about 200 miles, extends over an area of 14,000 square miles. Within the delta the river breaks up into an intricate network of channels called rivers. The Nun River is regarded as the direct continuation of the river, but some of the other important channels include the Forcados, the Brass, the Sombreiro, and the Bonny. The mouths of these channels are almost all obstructed by sandbars. The Forcados, for instance, which supplanted the Nun as the most travelled channel in the early 20th century, was in turn displaced by the Escravos River in 1964.

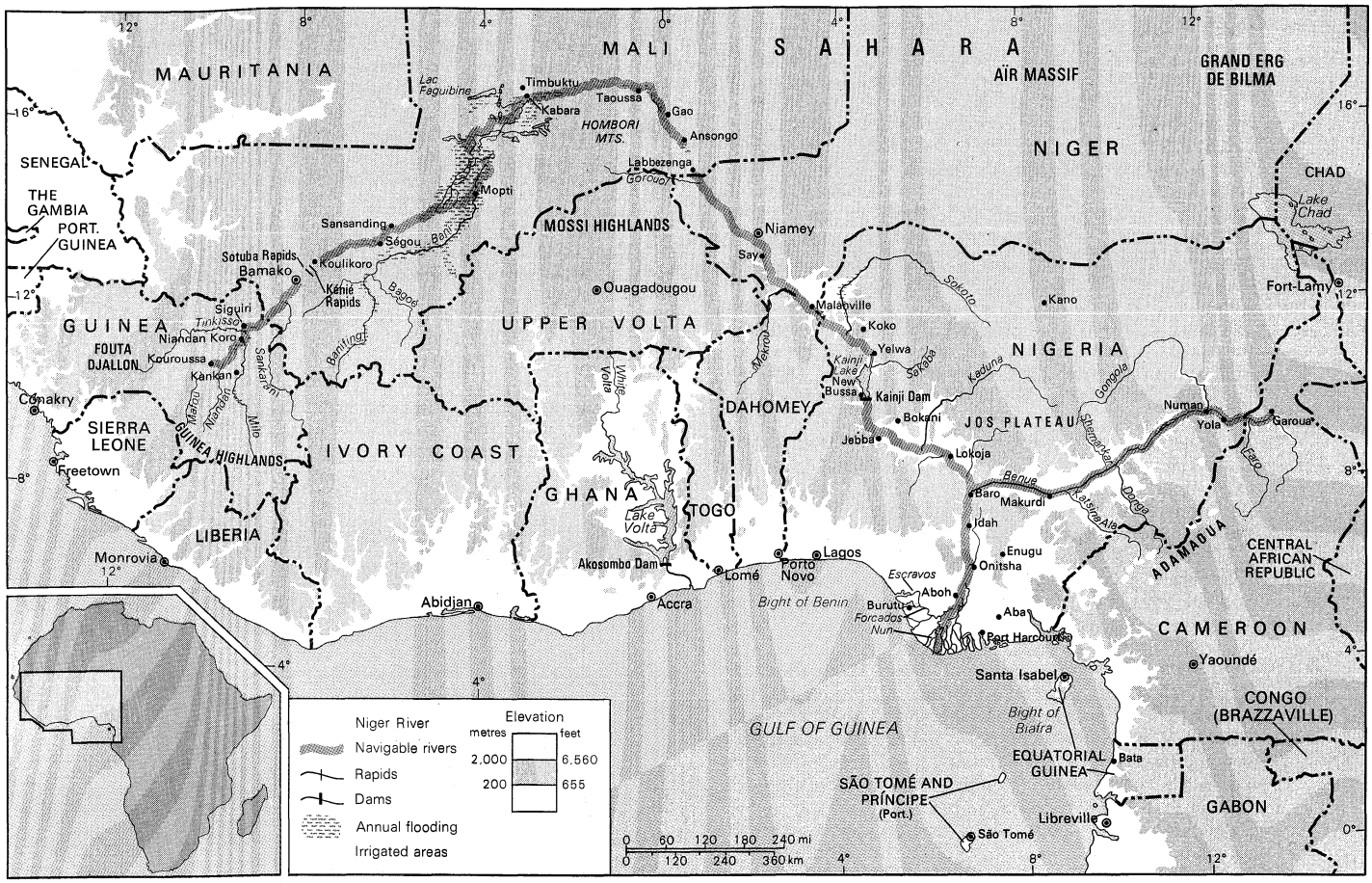
The Benue, (signifying "Mother of Water" in the Batta tongue), rises at 4,400 feet above sea level on the Adamawa (Adamaoua) Plateau in northern Cameroon at about 7°40′ N and 13°15′ E. In its upper course, which extends north–northwest to its confluence with the Mayo Kebi, close to the town of Garoua, it is a mountain torrent, falling more than 2,000 feet over a distance of 110 miles. The river then turns westward and, for the greater part of its course, flows over a broad and fertile floodplain. At Yola (Jimeta), a town in northeastern Nigeria 600 feet above sea level, some 850 miles inland, the width of the river in flood is from 1,000 to 1,500 yards. Near Numan, some 40 miles downstream from Yola, the Benue is joined on its north bank by its most important tributary, the Gongola. Other important tributaries include the Shemankar, the Faro, the Donga, and the Katsina Ala.

Together with its tributaries, the Niger drains a total area of more than 730,000 square miles. The Niger drainage system is bounded in the south by such highlands as the Fouta Djallon, the Kong Mountains, the Yoruba Hills, and the Cameroon Highlands. This southern rampart forms a watershed separating the rivers of the Niger system from others that flow directly southward to the Atlantic Ocean. With the exception of such highlands as the Jos Plateau, and the Adrar des Iforas, the Aïr, and the Haggar Mountains to the northeast, the northern edge of the Niger Basin is, however, less clearly defined than the southern edge.

Climate and hydrology. Within the Niger Basin, climate shows great variability. Mean annual rainfall decreases northward from over 160 inches in the delta area to under 10 inches in Timbuktu. Both the upper and the lower stretches of the river, however, drain areas with more than 50 inches of rain per annum. The middle Niger is an area where rainfall decreases and is also the sector where the greatest amount of evaporation takes place. It is estimated that in the lake region the Niger loses nearly 65 percent of the annual volume of discharge that flows past Mopti.

Because of climatic variations, the annual river flood does not occur at the same time in different parts of the basin. In the upper Niger the high-water discharge occurs in June, and the low-water season is in December. In the middle Niger, a first high-water discharge—the white flood (so-called because of the light sediment content of the water)—occurs soon after the rainy season between July and October; a second rise—the black flood (so-called because of the greater sediment content)—begins in December with the arrival of floodwaters from upstream. May and June are the low-water months in the middle stretch. On the Benue, there is only one high-water season. Due to the more southerly location, this normally occurs from May to October—earlier than on the middle Niger. The lower Niger below its confluence with the Benue consequently has a high-water period that begins in May or June—about a month earlier than on the middle Niger—and a low-water period at least a month shorter, as rains in the south start earlier. In January there also occurs a slight rise of water due to the arrival of floodwaters from the upper Niger. The difference between high and low water often measures as much as 35 feet.

Vegetation and fauna. Along its course, the Niger traverses virtually all the vegetational zones in West Africa. The Fouta Djallon plateau, where the Niger rises, is covered by a sedge type of vegetation consisting of fine, wire-like tufts interspersed with bare rock surfaces. From the



Niger and Benue rivers.

Fouta Djallon to well below the Niger's confluence with the Benue, the river flows mainly through savannah grassland country. In the north of the grassland region, tall, tussocky grass is interspersed with fairly dense wooded vegetation. In the south, the grass becomes short and discontinuous, and thorny shrub and acacia wood occur. About the latitude of Onitsha, the river enters the high rain-forest belt, which merges below Aboh with the mangrove swamp vegetation of the Delta.

River fauna

Many fish are found in the Niger and its tributaries; the chief food species are catfish, carp, and Nile perch.

Other Niger fauna include the hippopotamus, at least three different types of crocodile (including the much-feared Nile crocodile), and a variety of lizards. There is a rich variety of bird life. Geese are found in the lake region; and heron, egrets, and storks are found both on the river and the lakes. The striking crown bird is found wherever there is open ground in the grassland zone, and pelicans and flamingos are particularly associated with the upper Benue area. Smaller riverain species include white-headed plover, crocodile bird, sandpiper, curlew, and green-red shank.

Human ecology. The Niger Valley is sparsely settled, although there are population concentrations in the lake region and in the Nupe area of Nigeria. In medieval times, the valley was the heartland of the Mali and Songhai empires, and some of the river towns date from this period. The ethnic pattern along the course of the river shows larger groups, such as the Bambara, the Malinke, the Songhai, and the Zerma, occupying both sides of the river until the Nigerian boundary is reached, after which many small ethnic groups are encountered. Fishing is an important activity along the length of the river system, especially during the dry season when the deep-sea and coastal fish catch is smallest. River fishing is a specialized occupation for certain ethnic groups such as the Bozo and Somono in the lake region, the Sorkawa on the middle Niger, the Kede and the Kakanda between Jebba and Lokoja, and the Wurbo and the Jukun on the Benue.

History of the mapping of the Niger. It was not until the late 18th century that systematic European attempts to find the source, direction, and outlet of the Niger were made. In 1795 Mungo Park, a Scottish explorer sent out by the African Association of London, travelled overland from the Gambia region and reached the Niger near Ségou where, on July 20, 1796, he established that the river flowed eastward. In 1805 he sailed over 1,500 miles down the river, seeking to reach its mouth, but lost his life at Bussa. In 1822 another Scottish explorer, Major Alexander G. Laing located but did not visit the source of the river. In 1830 two English explorers, Richard and John Lander, established the lower course of the Niger by canoeing down the river from Yauri, 60 miles above Bussa, to the Atlantic Ocean, via the Nun River passage. In the second half of the 19th century, two German explorers—Heinrich Barth, and Edward R. Flegel—in separate travels established the course of the Benue from its source to its confluence with the Niger.

Mungo Park's expedition

Irrigation and navigation. The irrigation of the Niger Valley—for the purpose of transforming it into a densely populated, agricultural corridor running through the interior of West Africa—has long been a goal of planners. The French colonial administration, for example, set up the Office du Niger in 1932 to study and develop the Niger Valley, and to plan the irrigation of large areas in the lake region. A barrage to raise the level of the Niger by some 14 feet was completed at Sansanding in 1947. Feeder canals were constructed, and by 1969 some 500,000 hectares had been irrigated for rice, cotton, sugarcane, and vegetable production. The British colonial administration, for its part, also encouraged irrigated rice cultivation in the Bida region. In 1964 a large irrigated sugar plantation was started at Bacita, below Jebba.

Hydroelectric development

The Niger is also a source of hydroelectricity. The largest project is the Kainji Dam in Nigeria, completed in 1969. A new lake 500 square miles in extent has been created upstream, offering opportunities for fishing and irrigation. It is proposed that another dam should be built

at Jebba in 1982 to add a further 500 megawatts to the installed capacity.

Most of the Niger River—more than 75 percent of its total length—is used by commercial shipping. From the Atlantic Ocean to Onitsha—232 miles—the river is navigable to oceangoing vessels the year around. From Onitsha to Lokojo, at the confluence of the Benue and the Niger, oceangoing vessels can move for ten months of the year, from June to March. Navigation in this stretch is made possible by the influx of water from the Benue River, which is at high level in June. From Lokojo to Jebba the Niger is navigable to all craft only from October to mid-November. Thus, Jebba is in effect the head of navigation of the Niger waterway, although extreme fluctuation in water level at times constitutes a major handicap to vessels plying beyond Lokojo. Above Jebba the Niger is navigable only to smaller craft. The Bussa Rapids, formerly an obstacle to navigation between Jebba and Yelwa, may have been effectively eliminated by the construction of the Kainji Dam and Reservoir.

Seasonal limits to navigation

Rail and road routes cross the river at many points. Two railway bridges span the river at Kouroussa and Jebba, and a third crosses the Benue at Makurdi. Road bridges have been built at Ségou, Malanville, Kainji, and Onitsha. Ferries carry traffic across the Niger at Bamako, Gao, Niamey, Yelwa, Lokoja, and Idah; and across the Benue at Garoua, Yola and Numan. Among ports used for riverboat traffic are Koulikoro, Timbuktu, Baro, Onitsha, Burutu, and Koko.

The coordination of multinational efforts to develop the Niger and its tributaries is the responsibility of the Niger River Commission, formed in 1963 by Nigeria, Niger, Mali, Guinea, Chad, Cameroon, the Ivory Coast, Dahomey, and Upper Volta. The Commission has sponsored a study of the navigational possibilities of the middle Niger from Gao (Mali) to Yelwa (Nigeria). There are also plans to develop more irrigation and fishing projects, especially in the Kainji Lake area in Nigeria.

BIBLIOGRAPHY. MUNGO PARK, *Travels in the Interior Districts of Africa: Performed in the Years 1795, 1796 and 1797* (1799), the journal of the first European to reach the Niger, containing information on the middle stretch of the Niger especially between Koulikoro and Sansanding; RICHARD and JOHN LANDER, *Journal of an Expedition to Explore the Course and Termination of the Niger*, 3 vol. (1832), the first identification of the Benue (then known as the Tshadda); H. BARTH, *Travels and Discoveries in North and Central Africa*, vol. 4 and 5 (1857–58), a detailed account of the inland lake region of the Niger and of the area of the river between Timbuktu and Say; E.R. FLEGEL, *Vom Niger-Benue: Briefe aus Afrika* (1890), private correspondence describing events in the Niger-Benue area during the period of intense European colonial rivalry; E.A.L. HOURST, *Sur le Niger et pays des Touaregs* (1898), an illustrated description of conditions on the Niger from Bamako to the sea; Y.F.M. URVOY, *Les Bassins du Niger* (1942), a standard work on the geology and physical geography of the Niger Basin from its source to the Nigerian border; J. RICHARD-MOLARD, *Afrique Occidentale Française* (1949), an account of human and economic characteristics of the Niger Basin in French-speaking territories; R.J.H. CHURCH, *West Africa*, 6th ed. (1968), geography and economic significance on a state by state basis; NETHERLANDS ENGINEERING CONSULTANTS, *River Studies and Recommendations on Improvement of Niger and Benue* (1959), a detailed study of the hydrographic and physical characteristics of the Niger and Benue rivers within Nigeria; W.B. MORGAN and J.C. PUGH, *West Africa* (1969), with a useful section on fishing activities; R.K. UDO, *Geographical Regions of Nigeria* (1970), a detailed regional description of the Niger and its tributary, the Benue, within Nigeria.

(A.L.M.)

Nightingale, Florence

Florence Nightingale, for her pioneering development of military and civilian nursing and of hospital care, is universally regarded as the founder of trained nursing as a profession for women. Her work made her a legend in her own time. She also made important contributions to the welfare of the soldier in peace and war, to military and civil hygiene, and to the foundation of British district nursing and the training of midwives.

Florence Nightingale, *c.* 1885.
By courtesy of the Gernsheim Collection, the University of Texas at Austin

The second daughter of William Edward Nightingale and Frances (Fanny) Smith, Florence Nightingale was born on May 12, 1820, at Florence, Italy, where her well-to-do parents were temporarily resident. She grew up in Derbyshire, Hampshire, and London, where her family maintained comfortable homes. She was educated largely by her father, who taught her Greek, Latin, French, German, Italian, history, philosophy, and mathematics. Throughout her life she read widely in many languages. Social life was generally unsatisfying for Miss Nightingale. On February 7, 1837, she believed that she had heard the voice of God informing her that she had a mission, but it was not until nine years later that she realized what that mission was. Meanwhile, she strove to escape to a life of her own. Her proposal to study nursing at a hospital was scotched. She was then persuaded to study Parliamentary reports, and in three years she was regarded by influential friends as an expert on public health and hospitals. In 1846 a friend sent Miss Nightingale the Year Book of the Institution of Protestant Deaconesses at Kaiserswerth, Germany, which trained country girls of good character to nurse the sick. Four years later she entered the institution and went through the full course of training as a nurse. In 1853 she was appointed superintendent of the Institution for the Care of Sick Gentlewomen, in London. The changes that she made and her administration were very successful. But she yearned for a wider field; by January 1854 she was referring to the institution as "this little molehill."

The beginning of her nursing career

The Crimean War broke out in March 1854, and the allied British and French armies landed on the Crimea in September. Almost at once the British conscience was dismayed by published graphic reports of the disgraceful conditions suffered by sick and wounded British soldiers. Women were urged to serve as nurses like the French Sisters of Charity. Miss Nightingale volunteered at once to leave in three days for Constantinople, taking three nurses with her. Meanwhile, she was officially approached by her old friend, the then secretary of state at war, Sidney Herbert (later Lord Herbert of Lea), to take out a much larger party of nurses. She was to have complete charge of the nursing in the military hospitals in Turkey (*i.e.*, at Scutari). The party left England on October 21, 1854, and entered the Barrack Hospital at Scutari on November 5.

On her party's arrival she found that they had no decent facilities whatever. Their quarters were flea and rat infested, and the water allowance was one pint per head per day for all purposes. She had to use the provisions brought with her. The doctors were hostile, and at first the nurses were not allowed in the wards. After the Battle of Inkerman (fought on the very day of her arrival) the hospital was soon grossly overcrowded with sick and wounded. Furniture, clothing, and bedding were deficient, and in the corridors men lay on straw palliasses amidst filth caused by inadequate sanitation. Miss Nightingale was then asked to help, and one of her first requisitions was for 200 scrubbing brushes. She next arranged for the patients' filthy clothes to be washed outside the hospital.

All supplies had completely broken down, but Miss Nightingale had authority to purchase outside the hospital; she had brought £30,000 with her. By the end of the year she was purveying the hospital. She was harrassed by the cares of administration, a vast correspondence, the writing of numerous official and private reports, as well as by the insubordination of her nurses, some of whom had to be sent home because of drunkenness or immorality. She spent many hours a day in the wards, and there was scarcely a man whom she had not personally attended. After 8:00 PM she would allow no woman in the wards except herself. The night nursing—such as it was—was done by convalescent orderlies. Each night, however, she made her rounds, giving comfort and advice and establishing the wounded soldiers' conception of "The Lady with the Lamp."

By May 1855 nursing the sick had become her secondary interest, and her prime concern now was the welfare of the British Army. She now transferred herself and some of her nurses to the Crimea, and on landing at Balaklava she was very ill with Crimean fever. Then her opponent, the inspector general of hospitals, contended that she had authority only at Scutari and none in the Crimea. It was not until March 16, 1856, that her position as general superintendent of the Female Nursing Establishment of the Military Hospitals of the Army was confirmed in general orders.

Shortly after the last patient left the Barrack Hospital, Miss Nightingale sailed for England, where she had long been a national hero. But she refused official transport home and every kind of public reception. Miss Nightingale returned home determined to destroy her popular image and to inaugurate official action to improve the health, living conditions, and food of the British soldier. In the first she succeeded extraordinarily well. In the second she encountered difficulties, as the important men regarded her scheme tolerantly but without enthusiasm. In October 1856, however, she had a long interview with Queen Victoria, the Prince Consort, and Lord Panmure, Herbert's successor. She later had a private interview with the Queen, and Panmure promised a royal commission.

Soldiers' welfare and training of nurses

The Royal Commission on the Health of the Army was appointed in May 1857. Miss Nightingale gave extensive evidence and compiled an immense confidential report, covering the whole field of army medical and hospital administration, which was later privately printed as her *Notes on Matters Affecting the Health, Efficiency and Hospital Administration of the British Army* (1858). One consequence of the commission's activities was the foundation of the Army Medical School in 1857. The Indian Mutiny in the same year turned Miss Nightingale's interest to the health of the Army in India, and for that purpose another royal commission was appointed in 1859. This resulted in 1868 in the establishment of a Sanitary Department in the India Office with supreme authority in India.

Meanwhile, Miss Nightingale had been engaged in other pioneering activities. In 1860 she used the Nightingale Fund of £45,000, subscribed by the public to commemorate her Crimean work, to establish at St. Thomas's Hospital the Nightingale School for Nurses—the first of its kind in the world. Within a few years she was largely instrumental in inaugurating training for midwives and for nurses in workhouse infirmaries, and she played a part in the reform of workhouses. All these works were accomplished by a woman generally supposed to have died. From 1857 Miss Nightingale had lived, mainly in London, as an invalid. Her correspondence was enormous. Lying on her couch year after year, she received innumerable visitors, from the highest to the humblest, and few came who did not give information or receive it. Although she had never been to India, she was an acknowledged master of most things Indian, and successive viceroys consulted her before assuming their offices. She drove her influential friends to obtain for her those things that she felt her cause needed. When Sidney Herbert, a dying man, was forced to discontinue his active cooperation in their work, she sent him a very cruel letter.

It has never been shown that Florence Nightingale had any organic illness; her invalidism may have been partly neurotic and partly intentional. By this apparent stratagem she was able to devote herself night and day to the task at hand. Her sight gradually failed, until in 1901 she became completely blind. In 1907 the king conferred on her the Order of Merit—the first woman ever to receive it. Florence Nightingale died on August 13, 1910. The offer of a national funeral and burial in Westminster Abbey was, by her wish, declined. Her coffin was borne to the family grave in the country churchyard of East Wellow, Hampshire, by six sergeants of the British Army.

BIBLIOGRAPHY. SIR EDWARD COOK, *The Life of Florence Nightingale*, 2 vol. (1913), still the most comprehensive and completely documented biography; CECIL WOODHAM-SMITH, *Florence Nightingale, 1820–1910* (1950, reprinted 1964), a full and well-written modern biography; SIR ZACHARY COPE, *Florence Nightingale and the Doctors* (1958), very valuable in its field; LYTTON STRACHEY, essay in *Eminent Victorians* (1918, reprinted 1969), a brilliant sketch, but must be read with attention to subtle bias.

(E.A.U.)

Nijinsky, Vaslav

The leading dancer of the Mariinsky Theatre of St. Petersburg (now Leningrad) and of Sergey Diaghilev's great Ballets Russes, Vaslav Nijinsky created a new epoch in the dance and won almost legendary fame in doing so. His choreographies are noted for their rejection of the conventional forms of classical ballet and of the even then old patterns and structure of both the Classical and Romantic styles of dancing. Nijinsky was in search of pure line in movement and free form of expression.

The second son of Thomas Laurentiyevich Nijinsky and Eleonora Bereda, Vaslav was born in Kiev, Ukraine, on March 12 (February 28, old style), 1890. Both his parents were celebrated dancers, and his father in particular was famous for his virtuosity and enormous leaps. The Nijinskys had their own dance company and performed throughout the Russian Empire. Nijinsky's childhood was mostly spent in the Caucasus, where he danced as a small child with his brother Stanislav and his little sister Bronisława. His father, noticing the child's great disposition for

Nijinsky in *Spectre de la rose*.

dancing, gave him his first lessons, and at the age of nine, at the end of August 1898, Nijinsky entered the Imperial School of Dancing in St. Petersburg, where his teachers, the foremost of the time, soon discovered his extraordinary talent. When he was 16 years old, they urged him to graduate and enter the Mariinsky Theatre. Nijinsky declined, preferring to fulfill the customary period of study. At the time he already had been heralded as the "eighth wonder of the world" and the "Vestris of the North" (in reference to Auguste Vestris, a famous French dancer of the 18th century). During his school years he appeared at the Mariinsky Theatre, first as a member of the *corps de ballet*, later in small parts. He danced in St. Petersburg before the Tsar at the Chinese Theatre of Tsarskoe Selo and the Hermitage Theatre of the Winter Palace.

Association with the Mariinsky Theatre

Nijinsky graduated in the spring of 1907 and on July 14, 1907, joined the Mariinsky Theatre as a soloist. His first appearance was in the ballet *La Source* with the Russian ballerina Julia Sedova as his partner; the public and the ballet critics burst out immediately in wild enthusiasm. Among his Mariinsky partners were three great ballerinas, Mathilde Kschessinskaya, Anna Pavlovna Pavlova, and Tamara Platonovna Karsavina. As *danseur noble*, he danced the leading parts in many ballets, including *Ivanotschka, Giselle, Swan Lake, The Sleeping Beauty*, and *Chopiniana*. From 1907 to 1911 Nijinsky danced all of the leading parts at the Mariinsky Theatre and at the Bolshoi Theatre in Moscow, where he was a guest performer. His success was phenomenal.

In 1909 Sergey Diaghilev, former assistant to the administrator of the Imperial Theatres, was commissioned by the grand duke Vladimir to organize a ballet company of the members of the Mariinsky and Bolshoi theatres. Diaghilev decided to take the company to Paris in the spring and asked Nijinsky to join as principal dancer. Its first performance was on May 17, 1909, at the Théâtre du Châtelet. Nijinsky took Paris by storm. The expression and beauty of his body, his featherweight lightness and steellike strength, his great elevation and incredible gift of rising and seeming to remain in the air, and his extraordinary virtuosity and dramatic acting made him a genius of the ballet. From 1907 to 1912 he worked with the company's choreographer, Michel Fokine. With his phenomenal talent for characterization, he created some of his most renowned roles in Fokine's *Le Carnaval, Les Sylphides* (a revision of *Chopiniana*), *Le Spectre de la rose, Shéhérazade, Petrushka, Le Dieu bleu, Daphnis et Chloé*, and *Narcisse*. His later ballets were *Mephisto Valse, Variations on the Music of Johann Sebastian Bach, Les Papillons de nuit*, and *The Minstrel*. Until 1917 Nijinsky appeared all over Europe, in the United States, and in South America. He was called *le dieu de la danse*.

Principal dancer with Diaghilev's Ballets Russes

In 1912 he began his career as a choreographer. He created for Diaghilev's Ballets Russes the ballets *L'Après-midi d'un faune, Jeux*, and *Le Sacre du printemps. Tyl Eulenspiegel* was produced in the United States without Diaghilev's personal supervision. His work in the field of choreography was generally considered daringly original.

Nijinsky married Romola, countess de Pulszky-Lubocy-Cselfalva, in Buenos Aires on September 10, 1913. During part of World War I and again in World War II, he was interned in Hungary as a Russian subject. In 1919, at the age of 29, he retired from the stage, owing to a nervous breakdown, which was diagnosed as schizophrenia. He lived from 1919 until 1950 in Switzerland, France, and England, and died in London on April 8, 1950. Nijinsky is buried next to Auguste Vestris in the cemetery of Montmartre in Paris.

BIBLIOGRAPHY. The standard biographies, both by his wife, ROMOLA NIJINSKY, are *Nijinsky* (1933), and *The Last Years of Nijinsky* (1952). Other important studies are GEOFFREY WITHWORTH, *Nijinsky* (1911); CYRIL W. BEAUMONT, *Nijinsky* (1932); and COLIN WILSON, *The Outsider* (1956). Nijinsky's diary was published as *The Diary of Vaslav Nijinsky* (1936). *Nijinsky As We Knew Him* (1972), is a collection of reminiscences of the dancer by such prominent writers, artists, musicians, and dancers as Paul Claudel, Jean Cocteau, Auguste Rodin, Marcel Proust, Sergei Prokofiev, and Tamara Karsavina.

(R.Ni.)

Nikon

Russian patriarch and the leader of a reform movement that caused a schism in the Russian Orthodox Church, Nikon (Nikita) was born in 1605 in the village of Veldemanovo, near Nizhny Novgorod (now Gorky), the son of a peasant of Finnic stock. After acquiring the rudiments of an education in a nearby monastery, Nikon married, entered the clergy, and settled in Moscow, until the death of all three of his children moved him to seek repentance and solitude. For the next 12 years, from 1634 to 1646, he lived as a monk, as a hermit, and finally as an abbot in several northern localities. In 1646 he went on monastic business to Moscow, where he made so favourable an impression on the young tsar Alexis and on Patriarch Joseph that they appointed him abbot of the Novospassky monastery in Moscow, the burial place of the Romanov family. During his stay there, Nikon became closely associated with the circle led by the Tsar's confessor, Stefan Vonifatyev, and the priests Ivan Neronov and Avvakum Petrovich (all, like him, natives of the Nizhny Novgorod region), which strove to revitalize the church by bringing about closer contact with the mass of the faithful and to purify religious books and rituals from accidental errors and Roman Catholic influences. With their backing, Nikon became first metropolitan of Novgorod (1648) and then patriarch of Moscow and all Russia (1652).

Tass—Sovfoto

Nikon, portrait by an unknown artist, 1687.

Nikon accepted the highest post in the Russian Church only on condition that he be given full authority in matters of dogma and ritual. In 1654, when the Tsar departed for the campaign against Poland, he asked Nikon to supervise the country's administration as well as watch over the safety of the Tsar's family; and in 1657, with the outbreak of the new war with Poland, he invested him with full sovereign powers. Enjoying the friendship of the Tsar, the backing of the reformers, and the sympathy of the population of Moscow, Nikon stood at the pinnacle of his career.

It was not long, however, before Nikon alienated his friends and infuriated his opponents by his brutal treatment of all those who disagreed with him. On assuming the patriarchate, he consulted Greek scholars employed in Moscow as well as the books in the patriarchal library and concluded not only that many Russian books and practices were badly corrupted but also that the revisions of the circle of vonifatyev had introduced new corruptions. He

Nikon's reforms

then undertook a thorough revision of Russian books and rituals in accord with their Greek models to bring them in line with the rest of the Orthodox Church. When his onetime friends questioned his reforms, Nikon had them exiled. Assisted by Greek and Kievan monks and supported by the Greek hierarchy, he next carried out several reforms of his own: he altered the form of bowing in the church, replaced a two-fingered manner of crossing oneself with a three-fingered one, and ordered that three alleluias be sung where Moscow tradition called for two. A council of the Russian clergy that he convened in 1654 authorized him to proceed with the revision of liturgical books. He next began to remove from churches and homes icons that he considered incorrectly rendered. To quell mounting opposition to these moves, he called in 1656 another council, which excommunicated those who failed to adopt the reforms.

Though all the changes introduced by Nikon affected only the outward forms of religion, some of which were not even very old, the population and much of the clergy resisted him from the beginning. The uneducated Muscovite clergy refused to relearn prayers and rituals, while the mass of the faithful was deeply troubled by Nikon's contempt for practices regarded as holy and essential to Russia's salvation. This was the origin of the Raskol, or great schism within the Russian Orthodox Church. Yet what really brought about Nikon's downfall was the hostility of the Tsar's family and the powerful boyar (aristocratic) families, who resented the high handed manner in which he exercised authority in the Tsar's absence. They also objected to his claims that the church could interfere in affairs of state but was itself immune to state interference.

When Alexis returned to Moscow in 1658, relations between Tsar and Patriarch were no longer what they had been. Grown in self-confidence and incited by relatives and courtiers, Alexis ceased to consult the Patriarch, though he avoided an open break with him. Nikon finally struck back after several boyars had insulted him with impunity and the Tsar failed to appear at two consecutive services at which Nikon officiated. On July 20 (July 10, old style), 1658, in characteristically impetuous fashion he announced his resignation to the congregation in the Church of the Assumption in the Kremlin, and shortly afterward he retired to the Voskresensky monastery.

Nikon had apparently hoped by this act to compel the Tsar, whose piety was well known, to recall him and to restore his previous influence. This did not happen. After several months in self-imposed exile, Nikon began to regret his decision and attempted a reconciliation, but the Tsar either refused to answer his letters or urged him to formalize his resignation. Nikon refused to do so on the ground that he had resigned merely from the Moscow see, not from the patriarchate as such. For eight years, during which Russia was effectively without a patriarch, Nikon stubbornly held on to his post, while Alexis, troubled by lack of clear precedent and by the fear of damnation, could not decide on a formal deposition. Finally, in November 1666, Alexis convened a council attended by the patriarchs of Antioch and Alexandria to settle the dispute. The charges against Nikon were presented by the Tsar himself. They concerned largely his behaviour in the period of the Tsar's absence from Moscow, including his alleged arrogation of the title of "grand sovereign." Many of the charges were entirely without foundation. The Greek hierarchy now turned against Nikon and decided in favour of the monarchy whose favours it needed. A Greek adventurer, Paisios Ligaridis (now known to have been in collusion with Rome), was particularly active in bringing about Nikon's downfall. The council deprived Nikon of all his sacerdotal functions and on December 23 exiled him as a monk to Beloozero, about 350 miles directly north of Moscow. It retained, however, the reforms he had introduced and anathemized those who opposed them and who were henceforth known as Old Ritualists (or Old Believers). In his last years, Nikon's relations with Alexis improved. The succeeding tsar, Fyodor III, recalled Nikon from exile; but he died on August 27, 1681, en route to Moscow.

Exile

Nikon was one of the outstanding leaders of the Russian Orthodox Church, an able administrator, and a man of principle. His ultimate failure was due to two main factors: (1) his insistence on the hegemony of church over state had no precedent in Byzantine or Russian traditions and could not be enforced in any event; and (2) his uncontrollable temper and autocratic disposition alienated all who came in contact with him and enabled his opponents first to disgrace and then to defeat him.

BIBLIOGRAPHY. W. PALMER, *The Patriarch and the Tsar*, 6 vol. (1871–76); P. PASCAL, *Avvakum et les débuts du Raskol* (1938); W.K. MEDLIN, *Moscow and East Rome: A Political Study of the Relations of Church and State in Muscovite Russia* (1952).

(R.E.Pi.)

Nile River

The Nile, the father of African rivers and the longest river in the world, rises south of the Equator and flows northward through northeastern Africa to drain into the Mediterranean Sea. It has a length of about 4,132 miles (6,648 kilometres) and drains an area estimated at 1,293,000 square miles (3,349,000 square kilometres). Its basin includes parts of Tanzania, Burundi, Rwanda, Zaire, Kenya, Uganda, most of The Sudan, Ethiopia, and the cultivated part of Egypt. It supports a population estimated at about 50,000,000 people in the mid-1960s. Its most distant source is the Kagera River in Burundi.

The Nile is formed by three principal streams, the Blue Nile and the Atbara, which flow from the Highlands of Ethiopia, and the White Nile, the headstreams of which flow into lakes Victoria and Albert.

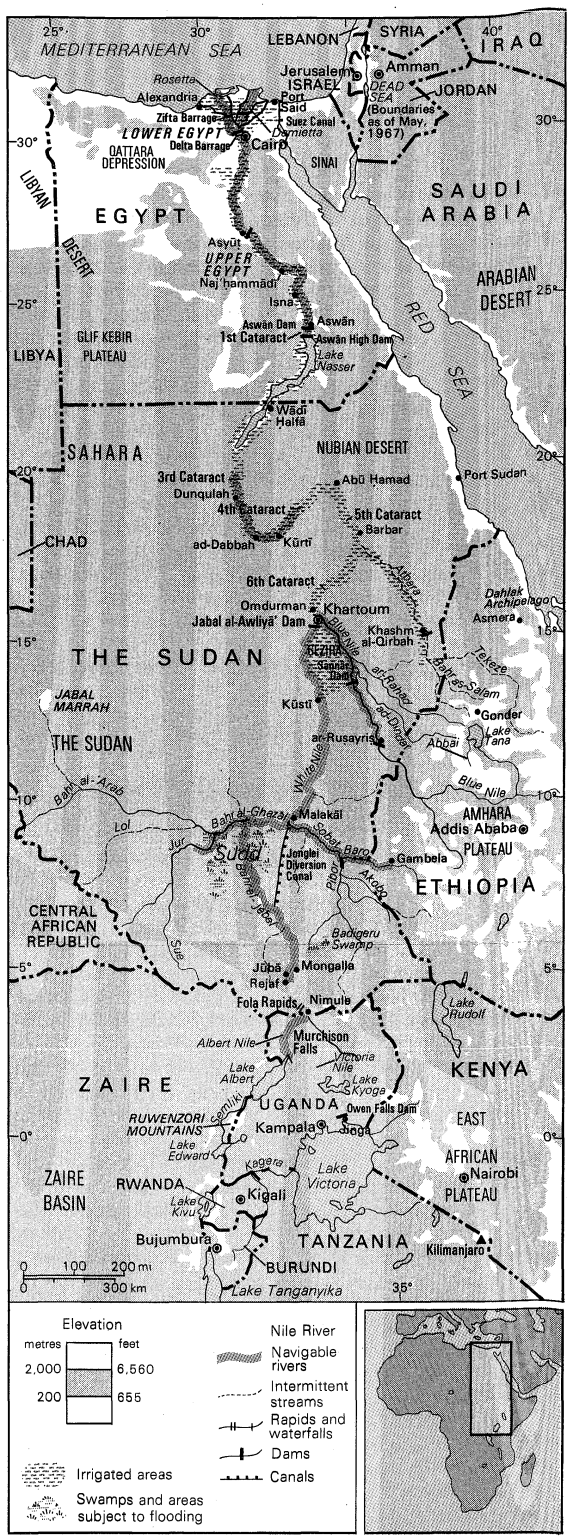
Origin of the name

The name Nile comes from the Greek Neilos (Latin Nilus), which is probably derived from the Semitic root *naḥal*, meaning a valley or river valley, and hence, by an extension of the meaning, a river. The fact that the Nile, unlike other great rivers known to them, flowed from the south northward, was an unsolved mystery to the ancient Egyptians and Greeks. The ancient Egyptians called the river Ar or Aur (Coptic Iaro), or "black," in allusion to the colour of the sediments carried by the river when it is in flood. Nile mud is black enough to have given the land itself its oldest name, Kem or Kemi, which also means "black" and signifies darkness. In *The Odyssey*, the epic poem written by the Greek poet Homer (7th century BC), Aigyptos is the name of the Nile (masculine) as well as the country of Egypt (feminine) through which it flows. The Nile in Egypt and northern Sudan is now called an-Nīl, al-Baḥr (the sea), and Baḥr an-Nīl or Nahr an-Nīl (the River Nile). (For physical features, see EAST AFRICAN LAKES; EAST AFRICAN MOUNTAINS; SAHARA.)

THE NATURAL ENVIRONMENT

The Nile Basin. The Nile River Basin, which covers about one-tenth of the area of the continent, served as the stage for the evolution and decay of advanced civilizations in the ancient world. On the banks of the river dwelled men who were the first to cultivate the arts of agriculture and the first to use the plow. The basin is bordered on the east by the Red Sea and the Ethiopian Highlands; on the west by the less well-defined watershed between the Nile, Chad, and Congo basins, extending northwest to include the Jabal Marrah (Marra Mountains) of The Sudan, the Glif Kebir Plateau, and the Libyan Desert (part of the Sahara); on the south, by the East African Highlands north of Lake Victoria, a Nile source; and on the north, by the Mediterranean.

The availability of water from the Nile throughout the year, combined with the area's unusually high temperature, makes possible intensive cultivation along its banks. Even in some of the regions in which the average rainfall is sufficient for cultivation, such as The Sudan, marked year to year variations in precipitation often make cultivation without irrigation hazardous. The Nile is also a vital waterway for transport, especially when motor transport is not feasible; *e.g.*, during the flood season. Recent improvements in air, rail, and highway facilities, however, have reduced dependency on the waterway.



The Nile River Basin and its drainage network.

Physiography of the Nile. *Origin and sources.* It is thought that in the mid-Tertiary Period (approximately 30,000,000 years ago), the early Nile, then a much shorter stream, had its sources about latitude 18° to 20° N. Its main headstream may then have been the present Atbara River. To the south lay the vast enclosed drainage system containing the large Lake Sudd. In a later stage (about 20,000 to 25,000 years ago), the East African drainage system—that of Lake Victoria—developed a new outlet, which sent its water northward into Lake Sudd. With the accumulation of sediments over a long period, the water level of this lake rose gradually; as a result of the overflow, the Sudd was drained, spilling over to the north. The overflow waters of Lake Sudd, rapidly forming a riverbed, linked the two major parts of the Nile system, thus unifying the drainage from Lake Victoria to the Mediterranean Sea.

The basin of the present-day Nile falls naturally into seven major regions.

The Lake Plateau of East Africa. The Lake Plateau region produces a number of headstreams and lakes that feed the White Nile. It is generally agreed that the Nile has several sources rather than one. The furthest headstream may be regarded as the Kagera River, which rises in the highlands of Burundi near the northern tip of Lake Tanganyika and then flows into Lake Victoria. The Nile proper, however, rises from Lake Victoria, the third largest lake in the world, which has an area over 26,800 square miles and forms a huge but shallow lake. The Nile begins near Jinja, Uganda, on the north shore of the lake, flowing northward over Ripon Falls, which has been submerged since the construction of the Owen Falls Dam in 1954. The northward stretch of the river, known as the Victoria Nile, enters the shallow Lake Kyoga (Kioga) and, passing through its swamp vegetation, flows out in a westerly direction, descending into the East African Rift Valley over Murchison Falls before entering the northern end of Lake Albert. Unlike Lake Victoria, Lake Albert is a deep, narrow lake with mountainous sides. Here the waters of the Victoria Nile unite with the lake waters, passing northward as the Albert Nile, a portion of the river, somewhat wider and slower, that is fringed with swamps and navigable for steamers.

Sources of the Nile

Bahr el-Jebel. At Nimule the Nile enters The Sudan rapidly over steep slopes and gorges and is called the Bahr el-Jebel (the River of Mountain). About four miles below Nimule lie the Fola Rapids, where the river is confined between rocky walls. It is not until Rejaf, about 100 miles downstream, that the river flows smoothly onto the huge and extremely flat clay plain, which extends through a narrow valley with hill country on either side, lying some 1,200 to 1,500 feet above sea level, and through the centre of which flows the mainstream. As the gradient of the Nile is here only 1:13,000, the great volume of additional water that arrives during the rainy season cannot be accommodated by the river, with the result that almost the entire plain becomes inundated. This circumstance promotes the growth of enormous quantities of swamp vegetation, including tall grasses and a tall reedlike papyrus known as sudd, literally meaning "barrier." The general term sudd is applied to the area because the great masses of vegetation, the growth of which is helped by the gentle flow of the water, break off and float downstream, choking the mainstream and locking the navigable channels. Channels have become further choked since 1957 by the rapid spread of the South American water hyacinth.

This basin receives drainage from numerous other streams. The Baḥr el-Ghazāl (River of Gazelles) flows in from the southwestern Sudan, joining the Bahr el-Jebel at Lake No, a large lagoon where the mainstream takes an easterly direction. The Baḥr el-Ghazāl looks impressive on a map, but its waters undergo such extensive loss through evaporation that only a small proportion of them ever reach the Nile. From the Ethiopian mountains to the southeast, the mainstream is joined by the Sobat—which from that point is called the White Nile—a short distance above Malakāl. The Sobat, the regime of which is quite different from that of Bahr el-Jebel, is an important source of water that compensates for the water lost through evaporation in the marshes.

The White Nile (al-Baḥr al-Abyaḍ). The White Nile, about 500 miles in length, is the longest branch of the Nile and supplies two-sevenths of its volume. It begins at Malakāl and joins the Blue Nile at Khartoum, receiving no tributaries of importance. Throughout this stretch it is a wide, placid stream with a very small slope (the gradient is no more than 1:19,000), often having a narrow fringe of swamps; the swamps have an average width of about 400 yards, but they are much wider in some places. The valley is wide and shallow, thus causing a considerable loss of water both by evaporation and seepage.

The Blue Nile (al-Baḥr al-Azraq). The Blue Nile drains from the lofty Ethiopian mountains north-northwestward, where it descends from a height of 6,000 feet above sea level. Its reputed source is a small spring, considered holy by the Orthodox Church of Ethiopia, from which a small stream, the Abbai, flows down to Lake Tana, a fairly shallow lake (with an area of about 1,400 square miles) that lies 6,000 feet above sea level. The river leaves Lake Tana in a southeasterly direction, flowing through a series of rapids and plunging through a deep gorge. It is estimated that the lake supplies the river with only one-fourteenth of its total flow, but this is important since it is silt-free. The river then flows west and northwest through The Sudan to join the White Nile at Khartoum. In the greater part of its course from Lake Tana down to The Sudan plains, it runs in a canyon that in places is 4,000 feet below the general level of the plateau. All its tributaries also run in deep ravines. While the White Nile at Khartoum is a river of almost constant volume, it is the Blue Nile that contributes to the Nile floods, which occur in Egypt in September as a result of the seasonal rains the river receives from its torrential tributary streams.

Source of the Blue Nile

The Atbara. The Atbara, the last tributary of the Nile, flows into the mainstream nearly 200 miles north of Khartoum. It rises in Ethiopia at heights of 6,000 to 10,000 feet above sea level, not far from Gonder, to the north of Lake Tana. The two principal tributaries that feed the Atbara are the Baḥr as-Salam and the Satīt, or Tekeze (the Terrible). The Satīt is the most important of these, having a basin more than double the area of the Atbara itself. It rises among the high peaks of Amhara and flows north through a spectacular gorge to join the Atbara in The Sudan. For most of its course in The Sudan, the Atbara is well below the general level of the plains. Between the plains and the river, the ground is eroded and cut into by gullies formed by water running off the plains after rainfall. The Atbara rises and falls rapidly, like the Blue Nile. In flood it becomes a large muddy river, and in the dry season it is a string of pools.

The Nile north of Khartoum. Along this stretch, which is sometimes called the United Nile, two parts can be distinguished. The first part, which stretches from Khartoum to Wādī Ḥalfā', is about 930 miles in length; the river is never far from the desert, and the desert often forms the riverbank. The second part consists of the stretch from Wādī Ḥalfā' (where the river enters Egypt) to Cairo, where the Nile Delta begins—a stretch along which cultivation is feasible.

Below Khartoum, the Nile flows 50 miles northward until it reaches Sablūkah (Sababka), the site of the sixth and highest cataract. Here the river cuts through hills for a distance of eight miles. Flowing northward at Barbar, the river takes an S-bend, in the middle of which, from Abū Ḥamad to Kūrtī and ad-Dabbah (Debba), the river flows southwestward for about 170 miles. At the end of this bend, at Dunqulah, it starts taking a northeasterly direction, crossing the second cataract below Wādī Ḥalfā'.

The cataracts

For the 1,200 miles from the sixth cataract to Aswān, the riverbed alternates between gentle stretches and series of rapids. Outcropping crystalline rocks that cross the course of the Nile cause the five famous cataracts. Because of these cataracts, the river is not completely navigable, although sections between the cataracts are navigable by sailing vessels and by river steamers. Since the land on both sides of the river in this stretch is not suitable for cultivation, the desert course of the Nile serves mainly to transport the water to Egypt. The Nile is navigable throughout Egypt from as far south as Aswān.

From the Sudanese border to Cairo, the Nile flows in a relatively narrow, flat-bottomed groove, sinuous in outline and generally incised into underlying sandstone and limestone rock surface; the gradient is approximately 1:14,000 over a distance of about 700 miles. For the first 250 miles it flows over sandstone in a narrow valley that is, on the average, less than two miles wide (although narrowing in some parts to a width of only 220 yards), except at Aswān, where resistant igneous and metamorphic rocks (*i.e.*, rocks formed by heat and pressure) appear in the rapids of the first cataract. The river then flows northward for about 500 miles through a valley that consists of a level-floored groove in a limestone plateau, averaging ten to 14 miles in width and enclosed by scarps that rise in places to heights of 1,500 feet above the river level. For the last 200 miles of its course before reaching Cairo, the Nile shows a strong tendency to hug the eastern edge of the valley floor, so that the greater part of the cultivated land is found on the left bank.

The Nile Delta. North of Cairo, the head of the alluvial plain of Lower Egypt, the walls of the plateau that form the Nile Valley become ill defined and diverge from each other, leaving between them the triangular lowland that forms the Nile Delta. In the 1st century AD, the Greek geographer Strabo recorded the Nile as having seven delta distributaries. The flow has since been controlled, so that the river now flows across the delta to the sea through two main distributaries, the Rosetta and the Damietta branches.

The Nile Delta, the prototype of all deltas, comprises a gulf of the prehistoric Mediterranean Sea that has been filled in; it is composed of silt brought mainly from the Ethiopian Highlands. The silt varies in its thickness from 50 to 75 feet and comprises the most fertile soil in Africa. It forms a monotonous plain that extends 100 miles from north to south, its greatest width being 155 miles between Alexandria and Port Said; altogether it covers an area twice that of the Nile Valley in Upper Egypt. The land surface slopes gently to the sea, falling some 52 feet from Cairo in a gentle gradient of 1:10,000. In the north, on the seaward border, are a number of shallow brackish lagoons and salt marshes: Buḥayrat Maryūṭ (Lake Marout), Buḥayrat Idkū (Edku), Buḥayrat al-Burullus (Burullus), and Buḥayrat al-Manzilah (Menzalah).

Climate. There is hardly an area within the Nile Basin that experiences a true Equatorial or a true Mediterranean type of climate. While almost the whole of the basin is rainless during the northern winter, its southern parts and the Highlands of Ethiopia experience heavy rain, amounting to over 60 inches, during the northern summer. The greater part of the region falls under the influence of the trade winds, which flow from the northeast in the Northern Hemisphere and from the southeast in the Southern Hemisphere, and account for the prevailing aridity of most of the basin.

Rainfall

Tropical climate with well-distributed rainfall is found in parts of the Lake Plateau and southwest Ethiopia. On the Lake Plateau there is little variation throughout the year in the mean temperature, which ranges from 60° F to 80° F (16° C to 27° C) depending on locality and height. Relative humidity, which varies similarly, is about 80 percent on the average. Similar climatic conditions prevail over the extreme southern parts of The Sudan, which receive as much as 50 inches of rain spread over a nine-month period (March–November), with the maximum occurring in August. At the beginning and end of the rainy season, strong and sudden windstorms may occur. The humidity reaches its highest at the peak of the rainy season and reaches its low level between January and March. Maximum temperatures are recorded during the dry season (December to February) with the minimum occurring in July and August.

Northward, the rainy season gets shorter, and the amount of rainfall decreases. The rainy season, which occurs in the south from April to October, is confined to July and August in the northern part of central Sudan, where three seasons may be distinguished. The first of these is the pleasant, cool, dry winter, which occurs in December and January; this is followed by hot and very dry weather from March to June; this is followed, in turn, by a hot rainy period from July to October. The minimum temperature occurs in January and the maximum in May or June, when it rises to 94° F (34° C) in Khartoum. Only about ten inches of rainfall occur in the Gezira area, as compared with over 21 inches at Dakar, in Senegal, which is in the same latitude, 15° N. North of Khartoum less than five inches of rain falls annually, an amount insufficient for permanent settlement. In June and July the central parts of The Sudan are frequently visited by storms during which strong winds carry large

quantities of sand and dust. These storms, which are of three to four hours duration, are called haboobs.

A desert-type climate exists over most of the remainder of the area north to the Mediterranean. The principal characteristics of the northern Sudan and the desert of Egypt are aridity, a dry atmosphere, and a considerable seasonal, as well as diurnal, temperature range in Upper Egypt. Temperatures can surpass 100° F (38° C); in Aswān, for example, the mean daily maximum in June is 107° F (42° C). While no low temperatures are recorded anywhere in The Sudan or Egypt, winter temperatures decrease to the north. Thus, only Egypt has what could be called a winter season, which occurs from November to April, when the average temperature in Cairo is 52° F (11° C). In the hot summer season from May to October, temperatures average about 81° F (27° C). The rainfall in Egypt is of Mediterranean origin and falls mostly in the winter, with the amount decreasing toward the south. From eight inches on the coast, it falls gradually to a little over an inch in Cairo and to less than an inch in Upper Egypt. During the spring, from March to June, depressions from the Sahara or along the coast travel east, causing dry southerly winds, which sometimes results in a condition called khamsin. These are sandstorms or dust storms during which the atmosphere becomes hazy; on occasion they may persist for three or four days, at the end of which the phenomenon of a "blue" sun may be observed.

Plant and animal life *Vegetation.* In the areas where no irrigation is practiced, different zones of plant life may be roughly divided according to the amount of rainfall.

The tropical forest zone

Tropical rain forest is found along the Nile–Congo divide, in parts of the Lake Plateau, and in southwest Ethiopia. Heat and copious rainfall produce thick forests with a great variety of tropical trees and plants, including ebony, banana, rubber, bamboo, and coffee shrub. Savanna (grassland) forest, characterized by a sparse growth of thinly foliaged trees of medium height and a ground covering of grass and perennial herbs, occurs in large parts of the Lake Plateau, in parts of the Ethiopian plateau, in the area that fringes the Blue Nile near ar-Ruṣayris and in the southern Baḥr el-Ghazāl region.

On The Sudan plains, a mixture of thin bush, thorny trees, and open grassland, or true savanna, prevails. This area is swampy during the rainy season. This is particularly true of the Sudd region of the south central Sudan, an area of nearly 100,000 square miles. Here the vegetation includes papyrus; tall bamboo-like grasses; reed mace; ambatch, or turor; water lettuce; a species of convolvulus; and the South American water hyacinth.

North of latitude 10° there occurs a belt of thorny savanna or orchard shrub country characterized by small scattered tree stands, thornbush, and—after rain—grass and herbs. North of this, however, rainfall decreases and the vegetation thins out, so that the countryside is dotted with small thorny shrubs, mostly acacias; this area is known as shrub steppe. From Khartoum northward there is true desert, with scanty and irregular rainfall and no permanent vegetation at all except for a few stunted shrubs. Grasses and small herbs may be scattered along drainage lines after rainfall, but these die away in a few weeks. In Egypt, the vegetation near the Nile is almost entirely the result of irrigation and cultivation.

Animal life. Many varieties of fish are found in the Nile system. Notable among those found in the Lower Nile system are the Nile perch (which may weigh over 200 pounds), the bolti (tilapia), the barbel, several species of catfish, the elephant-snout fish, and the tiger fish, or water leopard. Most of these species and the sardine-like *Haplochromis*, the lungfish, and mudfish are found as far upstream as Lake Victoria. The common eel penetrates as far south as Khartoum, and the spiny eel is found in Lake Victoria.

The Nile crocodile, found in most parts of the river, has not yet penetrated the lakes of the Upper Nile Basin. Other reptiles found in the Nile Basin include the soft-shelled turtle, three species of monitor lizard, and some 30 species of snakes, of which more than half are venomous. The hippopotamus, once common throughout the Nile system, is now found only in the Sudd region of Sudan and to the south.

Fish

Many schools of fish that come to feed on the reddish-gray waters of the Nile, in Egypt, during the flood season may be reduced or disappear after the floodwaters are controlled by the High Dam at Aswān. Most of the species of the Nile fish are migrants; the dam may prevent such species from migrating to Lake Nasser. The diminution in the number of anchovies in the eastern Mediterranean has also been attributed to the reduction in the outflow of waterborne nutrients due to the construction of the dam.

MAN AND THE NILE

Human ecology. Nilote is the general term used to describe the people of the Upper Nile Basin. Much of southern Sudan is populated by Nilotic and Negro people consisting of such tribes as the Shilluk, the Dinka, and the Nuer, who are a mixture of Negro and Hamitic stock. The Hamites are lean, slender, brown people with aquiline features and frizzy hair; they include Nubians, Bejas, Gallas, and Somali, who also live on the banks of the Nile. The Dinka, a typical group living in the Sudd region, are a pastoral people who live a seminomadic life, grazing their cattle on the grasslands between rivers in the dry season and moving to the uplands when the lower land is flooded. There are also the Nilo-Hamites, who include the Bari, a people who resemble the Dinka in appearance and live in the country of Bahr el-Jebel south of Mongalla. The Azande, a Nigritic people who are of median height and copper coloured, live in the southern Baḥr al-Ghazāl and in Zaire (formerly Congo [Kinshasa]). The central zone of Sudan (between latitudes 10° and 15° N) is inhabited by pastoral Semitic people, but Muslim Arabs and Nubians, as well as Hamitic people, such as the Beja, predominate in the northern parts. While many different people are found in Ethiopia, most are classified as Hamitic. The ruling peoples of the country, the Amhara and the Galla, though both of Hamitic stock, have been subjected to strong Semitic influence. The remainder of the population is made up of many tribes of Negro people known as Shangalla. The modern Egyptian is still primarily of Hamitic stock intermixed with Semitic and Caucasian elements to the north and Nubian–Negro elements to the south. The Egyptian peasant is stocky; his facial features show some resemblance to those of the Arabs. The people of the purest Arab descent are the Bedouins of the desert.

The average population density in the cultivated parts of the Nile floodplain is more than 1,000 per square mile. This great population, composed mostly of peasant farmers (fellahin), can survive only by making the most careful use of the available land and water.

The large quantities of silt washed down from the rich Highlands of Ethiopia are deposited by the floodwaters in Egypt, where the fertility of the riverine lands has continued over the centuries, despite intensive cultivation. A vital feature in the life of the Egyptian people is the river's behaviour, since a good harvest follows a good flood, and a poor flood may mean a later food shortage.

Exploration and mapping of the Nile. The ancient Egyptians were probably familiar with the Nile as far as Khartoum and with the Blue Nile as far as its source in Lake Tana, but they showed little or no interest in exploring the White Nile. The source of the Nile was unknown to them, but the river was associated with the worship of the god Apis (Hapi), the bull of Memphis, since it provided them with water to irrigate their crops. The Greek historian Herodotus, who visited Egypt in 457 BC, travelled down the Nile as far as the first cataract (Aswān). About the second century BC, the Greek scientific writer Eratosthenes sketched a nearly correct route of the Nile to Khartoum, showing the two Ethiopian affluents, and suggested lakes as the source of the river.

Early exploration

About 30 BC the Greek explorers Bion, Dalion, and Simonides and the merchant Diogenes, explored the Nile above the first cataract and perhaps south of Khartoum. In 25 BC the Greek geographer Strabo and a Roman governor of Egypt, Aelius Gallas, also explored the Nile as far as the first cataract. A Roman expedition to find the

source of the Nile in AD 66, during the reign of the emperor Nero, was impeded by the Sudd, and the attempt was abandoned. Claudius Ptolemy, the Greek astronomer and geographer who lived in Alexandria, wrote in AD 150 that the White Nile originated in the high snow-covered Mountains of the Moon (since identified with the Ruwenzori Mountain Range).

From the 17th century onward, several attempts were made to explore the Nile. In 1613, Pedro Páez, a Spanish Jesuit priest, located the source of the Blue Nile. In 1786, the Scottish explorer James Bruce visited Lake Tana as well as the source of the Blue Nile. He was followed by a number of other 19th-century explorers of various nationalities. After Mohammed Ali, the Ottoman viceroy of Egypt, conquered the Sudan in 1821, he sent expeditions (between 1839 and 1842) to attempt to find the source of the Nile.

After the English explorers Richard Burton and John Speke made an expedition, during which Speke reached the southern end of Lake Victoria in 1858, Speke returned in 1860 with another Englishman, James Grant, and after more than a year reached the western end of Lake Victoria, discovered the Kagera, and reached the Ripon Falls in 1862. The last unexplored stretch of the Nile was identified when yet another English explorer, Sir Samuel White Baker, discovered Lake Albert in 1864. Although Speke and Baker between them had solved the mystery of the Nile's headstreams, doubts persisted and were only laid to rest when the Scottish soldier General Charles George Gordon and his officers followed the river and mapped part of it and Lake Albert. Later, further features of the region were identified when the Welsh explorer Sir Henry Morton Stanley and the German traveller Mehmed Emin Paşa travelled through the Semliki Valley and by Lake Edward in 1889, and Stanley discovered the Ruwenzori Mountain Range. Exploration and mapping has continued over the years: it was not until the 1960s, for example, that a detailed study of the upper gorges of the Blue Nile was completed.

WATER RESOURCES AND THEIR USE

Hydrology. The periodic rise of the Nile, which occurred as regularly as the revolution of heavenly bodies, remained an unsolved mystery until the discovery of the role of the tropical regions in its regime. In effect, there was little detailed knowledge about the hydrology of the Nile before the 20th century except for early records of the river level that the ancient Egyptians made with the aid of nilometers (gauges formed by graduated scales cut in natural rocks or in stone walls), some of which still remain. Today, however, no other river of comparable size has a regime that is so well known. The discharge of the mainstream, as well as the tributaries, is regularly measured at all points critical for irrigation.

The Nile flood

The Nile swells in the summer, the floods rising as a result of the heavy tropical rains of the Upper Nile Basin, both in Ethiopia and on the East African Plateau. In southern Sudan the flood begins in April, but the effect is not felt at Aswān, in Egypt, until July. The water then starts to rise and continues to do so throughout August and September, with the maximum occurring in mid-September. At Cairo, the maximum is delayed until October. The level of the river then falls rapidly through November and December and more rapidly through the succeeding months. From March to May the level of the river is at its lowest. Although the flood is a fairly regular phenomenon, it occasionally varies in volume and date.

Following the river from its sources, an estimate can be made of the contribution of the various lakes and tributaries in the Nile flood.

Lake Victoria, in effect, forms a plateau reservoir. Most of its water (86 percent) is from rainfall, while the rest is contributed by its main tributary, the Kagera, as well as other much smaller ones. Much of the water is lost by evaporation. Only 18 percent of the lake's water flows into the Victoria Nile to feed Lake Albert, which forms a natural reservoir with a total annual inflow of about 1,000,000,000 cubic feet. The Victoria Nile contributes about two-thirds of the total inflow in Lake Albert, the rest coming from tributaries (of which the Semliki is the most important) as well as from rainfall. The loss by evaporation is not as high as it is in Lake Victoria, thus permitting a large annual outflow down the Bahr el-Jebel of 780,000,000,000 cubic feet. In addition to the water it receives from the great lakes, the torrential tributaries of the Bahr el-Jebel supply it with about 17 percent of its water. The discharge of Bahr el-Jebel varies little throughout the year because of the regulatory effect of the large swamps and lagoons of the Sudd region. About half of its water is lost in this stage by seepage and evaporation. The average flow of water is estimated at 950,000,000,000 cubic feet per annum as it enters the Sudd region at Mongalla, but only an average of 600,000,000,000 cubic feet per annum leaves at Malakāl. In order to reduce water loss, it has been proposed that a canal be cut through the basin and maintained as an open waterway.

The White Nile provides a regular supply of water throughout the year. During April and May, when the mainstream is at its lowest level, 83 percent of its water comes from the White Nile. The White Nile obtains its water in roughly equal amounts from two main sources. The first source is the rainfall on the East African Plateau of the previous summer. This long delay in reaching the lower parts of the Nile is caused partly by the retention of water in Lake Victoria and partly by the slow flow in the flat Sudd region. The second source is the drainage of southwestern Ethiopia through the Sobat (contributed mainly by its two tributaries, the Baro and the Pibor) that enters the mainstream below the Sudd. The annual flood of the Sobat, a consequence of the Ethiopian summer rains, is to a great extent responsible for the variations in the level of the White Nile. The rains that swell its upper valley begin in April and cause widespread inundation over the 200 miles of plains through which the river passes, thus delaying the arrival of the rainwater in its lower reaches until November–December. Relatively small amounts of the mud carried by the Sobat's flood reach the White Nile.

The regime of the Blue Nile

The Blue Nile, the most important of the three great Ethiopian affluents, plays an overwhelming part in bringing the Nile flood to Egypt. It receives two tributaries in The Sudan—the Rahad and the Dindar—both of which also originate in Ethiopia. The regime of the Blue Nile is distinguished from that of the White Nile by the more rapid passage of its floodwater into the mainstream. The river level begins to rise in June, reaching a maximum level at Khartoum in about the first week in September.

The Atbara River draws its floodwater from the rains on the northern part of the Ethiopian Plateau, as does the Blue Nile. While the floods of the two streams occur at the same time, the Blue Nile is a perennial stream, while the Atbara, as mentioned, shrinks to a series of pools in the dry season.

The swelling of the Blue Nile causes the first floodwaters to reach the central Sudan in May. The maximum is reached in August, after which the level falls again. The rise at Khartoum averages more than 20 feet. When the Blue Nile is in flood it holds back the White Nile water, turning it into an extensive lake and delaying its flow. The Jabal al-Awliyā', about 45 miles south of Khartoum, increases this ponding effect.

The effect of the flood is not felt at Wādī Ḥalfā' (on the Egyptian border) until mid-June. At Aswān the maximum is delayed until September, when the total discharge amounts to a little over 24,700,000,000 cubic feet per day. Out of this amount the Blue Nile accounts for 68 percent, the Atbara 22 percent, and the White Nile 10 percent. In early May, the Nile discharge at Aswān drops to its minimum. At this time the total discharge of 1,600,000,000 cubic feet comes mainly from the White Nile, which contributes 83 percent, while the Blue Nile contributes 17 percent. On the average, 84 percent of the Nile water at Aswān comes from the Ethiopian Highlands, while 16 percent is contributed by the East African Lake Plateau system.

Irrigation. As an aid to cultivation, irrigation almost certainly originated in Egypt. A particular phenomenon

that makes irrigation from the Nile feasible is the slope of the land from south to north—which amounts to about five inches to the mile—as well as the slightly greater slope downward from the riverbanks to the desert on either side.

Basin irrigation

The first use of the Nile for irrigation in Egypt began when seeds were sown in the mud left after the annual floodwater had subsided. With the passing of time, these practices were refined until a traditional method emerged, which is known as basin irrigation, and is still practiced on less than 10 percent of the agricultural land, mainly in Upper Egypt. The fields on the flat floodplain are divided by earth banks into a series of large basins, which vary in size but may be as large as 50,000 acres. At the time of the flood, the Nile waters, rich in reddish-gray silt, are drained off through carefully constructed flood canals. As the flood rises, the water spills into the various basins, where it is allowed to stand from six to eight weeks before it is permitted to drain away, leaving the rich silt behind it. Autumn and winter crops are then sown. The disadvantage of this method is that it allows only one crop per year—a crop, moreover, that is always at the mercy of annual fluctuations in the size of the flood.

In the basins situated above the flood level, water is elevated by such means as the shadoof (a counterbalanced lever device that uses a long pole); the sakieh, or Persian waterwheel; and the Archimedean screw. It should be noted that water thus lifted is largely free from sediments and thus does not fertilize the fields. Modern mechanical pumps are also in wide use today.

Because of the limitations of the basin method of irrigation, perennial irrigation, which controls the water so that it can be run into the land at regular intervals throughout the year, has been introduced. Perennial irrigation was made possible by the completion of several barrages and waterworks before the end of the 19th century. By the beginning of the 20th century, the canal system had been remodelled and the first Aswān Dam completed (see below, *Dams and reservoirs*). Over four-fifths of Egyptian cultivated land now benefits from perennial irrigation; after the completion of the Aswān High Dam it is likely that cultivation in Egypt will be almost entirely perennial.

While the people of The Sudan make use of the Nile for irrigation, reliance on the river is not absolute, as a fair amount of rainfall occurs in the southern parts. Basin irrigation from the Nile floods is used to a small extent, but it is less satisfactory because the surface is more uneven, with less deposition of silt; the area inundated also varies from year to year. Since about 1950, these traditional methods of irrigation have been largely displaced by diesel-engined pumps, which are used on about 2,500,000 acres on the banks of either the main Nile or the White Nile.

Perennial irrigation in The Sudan began with the completion of the combined dam and barrage near Sannār on the Blue Nile in 1925. This made possible the irrigation of the area of the clay plain called Gezira between the two Niles south of Khartoum. The success of this attempt encouraged the construction of more dams and barrages for large-scale irrigation schemes.

Dams and reservoirs. In 1843 it was decided to build a series of diversion dams (barrages or weirs) across the Nile at the head of the Delta about 12 miles north of Cairo, so as to raise the level of water upstream to supply the irrigation canals and to regulate navigation. This Delta barrage scheme was not fully completed until 1861, after which it was extended and improved; it may be regarded as marking the beginning of modern irrigation in the Nile Valley. The Zifta Barrage, nearly half way along the Damietta branch of the deltaic Nile, was added to this system in 1901. In 1902 the Asyūṭ Dam, more than 200 miles upstream from Cairo, was completed. This was followed in 1909 by the barrage at Isnā (Esna), about 159 miles above Asyūṭ, and in 1930 by the dam at Naj' Hammādī, 150 miles above Asyūṭ.

The Aswān High Dam

The first Aswān Dam was constructed between 1899 and 1902; it has a series of four locks to allow navigation. It has twice been enlarged—first between 1908 and 1911 and again between 1929 and 1934—thus raising the water level 120 feet and increasing the dam's capacity to 4,000,000 acre feet. It is equipped with a hydroelectric plant with an installed power of over 345,000 kilowatts.

The Aswān High Dam (as-Sadd al-'Ālī) is located about 600 miles upstream from Cairo and four miles upstream from the first Aswān Dam. It is built at a place where the river is 1,800 feet wide and has steep banks of granite. The dam is designed to control the Nile water for the expansion of cultivation and for the generation of hydroelectric power and to provide protection against unusually high floods. The work began in 1959 and was completed in 1971.

The High Dam has a storage capacity of 133,000,000,000 acre feet. It is 12,565 feet long at crest level and 3,280 feet wide at the base, with a height of 364 feet above the river bed. A hydroelectric plant was installed, which in 1966 generated 5,895,000,000 kilowatt hours of electricity. This will double when the plant's 12 turbine power stations are in full operation. Lake Nasser, the reservoir formed upstream as a result of the construction of the dam, will be over 300 miles long and will form the largest man-made lake in the world. The Aswān Reservoir will stretch 311 miles upstream from the dam site, thus extending 125 miles into The Sudan, where it will inundate an area inhabited by 50,000 Sudanese who will have to be resettled. The reservoir will have a storage capacity of 5,790,000,000,000 cubic feet, of which 2,500,000,000,000 is to store water for perennial irrigation, 1,000,000,000,000 for the disposition of silt and 1,000,000,000,000 for flood protection and for the regulation of flood flows for downstream use. By 1970 there were indications that the reduction in the flow of the Nile resulting from the construction of the dam was permitting the inundation of the lower reaches of the river by saltwater from the Mediterranean Sea with resulting deposition of salt in the delta soils.

In The Sudan the Sannār Dam on the Blue Nile, which was completed in 1925, provides water for the Gezira Plain at the time of year when the water level of the Blue Nile is low. It has a storage capacity of 33,000,000,000 cubic feet and also produces hydroelectric power. Another dam, at Jabal al-Awliyā' on the White Nile, was completed in 1937; it has a storage capacity of 88,000,000,000 cubic feet. In 1964, a dam was built on the Atbara at Khashm al-Qirbah with a storage capacity of 46,000,000,000 cubic feet; it also produces hydrolectric power. The ar-Ruṣayriṣ Dam on the Blue Nile was completed in 1966; it has a storage capacity of 102,000,000,000 cubic feet.

In Uganda, Lake Victoria was made into a reservoir by the completion in 1954 of the Owen Falls Dam; the dam is situated on the Victoria Nile just below the point where the lake waters flow into the river. This permits the storage of surplus water in high-flood years to meet the deficit in years when the waters are low. The fall from the lake is harnessed by a hydroelectric plant that provides power for industries in Uganda.

Navigation. As already mentioned, the Nile River is still a vital waterway for the transportation of people and goods, especially in the flood season when motor transport is not feasible; river steamers still provide the only means of transport facilities in most of the area, especially in The Sudan south of latitude 15° N, where motor transport is not usually possible from May to November. Most of the towns in Egypt and The Sudan are situated on or near riverbanks.

In The Sudan, steamer service on the Nile and its tributaries extends for about 2,400 miles. Until 1962 the only link between the northern and southern parts of The Sudan was by stern-wheel river steamers of shallow draft. The main service is from Kūstī to Jūbā. There are also seasonal and subsidiary services on the Dunqulah reaches of the main Nile, on the Blue Nile, up the Sobat to Gambela in Ethiopia, and up the Baḥr al-Ghazāl in the high-water season. The Blue Nile is navigable only during the high-water season and then only as far as ar-Ruṣayriṣ.

Navigable stretches

Because of the presence of the cataracts north of Khartoum, the river is navigable only in three stretches. The

first of these is from Wādī Ḥalfā' down to the first cataract south of Aswān. The second is the stretch between the third and the fourth cataract. The third and most important stretch extends from Khartoum southward to Jūbā.

In Egypt, the Nile is navigable by sailing vessels and shallow-draft river steamers as far south as Aswān; thousands of small boats ply the Nile and Delta waterways.

Prospects for the future. The Nile is a geographical unit. Despite the great diversity in the religious, ethnic, and political backgrounds of the people who live within its basin, all depend on the Nile. Much has been accomplished in the control and use of the Nile water by means of such projects as the Owen Falls Dam, the Sannār Dam, the Aswān High Dam, and many others. These projects not only provide water for irrigation and land reclamation but also affect every aspect of the livelihood of the people who dwell on the Nile's banks. The completion of the High Dam, for example, doubled the per capita consumption of electricity in Egypt, providing power for many basic industries. Plans are already completed to exploit the fish resources from Lake Nasser, which was formed as a result of the construction of the dam.

Further conservation projects on the Upper Nile are possible, and the tasks of accomplishing them become less difficult as understanding and cooperation increase among the Nile countries. In 1970 Egypt was planning to include in its development budget the financing of a project located in The Sudan—the Jonglei Diversion Canal (to run from Jonglei to the junction of the White Nile and Sobat near Malakāl), which would hasten the flow of water northward and prevent excessive losses in the Sudd region.

Present indications are that future developments may eventually lead to an integrated plan for the entire Nile Basin.

BIBLIOGRAPHY. S.W. BAKER, *Exploration of the Nile Tributaries of Abyssinia* (1868), is a description of the author's exploration of the Ethiopian tributaries of the Nile; a later edition is entitled *The Nile Tributaries of Abyssinia.* Geographical and other aspects of the Nile in The Sudan are covered in K.M. BARBOUR, *The Republic of Sudan: A Regional Geography* (1961); P.M. HOLT, *A Modern History of Sudan* (1961); J.H.G. LEBON, "Land Use Mapping in Sudan," *Econ. Geogr.*, 35:60–70 (1959); and J.D. TOTHILL (ed.), *Agriculture in the Sudan* (1948). F.V.M. BEKER, *Le Général Desaix, étude historique* (1852), is an account of the French expedition up the Nile, a member of which, D.V. DENON, wrote *Travels in Upper and Lower Egypt During the Campaigns of General Bonaparte,* 2 vol. (1802). J. BRUCE, *Travels to Discover the Source of the Nile,* abridged ed. (1964), was first published in 1790. Criticized as fiction because of exaggeration and the absence of witness, it is a work of major originality and importance; it contains descriptions of Bruce's travels in 1769, 1770, 1771, 1772, and 1773. R.F. BURTON, *The Nile Basin,* was first published in 1864; the 1967 edition contains a reprint of Burton's paper that introduced his theory of the Nile arising from Lake Tanganyika. A second essay by JAMES MACQUEEN is a critical review of *Captain Speke's Journal of the Discovery of the Source of the Nile* (1864, reprinted 1969), which gives a full account of his discovery of the sources of the Nile. A general study of the river from its source and a discussion of the utilization of its water is contained in H.E. HURST, *The Nile* (1952), and "Progress in the Study of the Hydrology of the Nile in the Last Twenty Years," *Geogrl J.*, 70:440–463 (1927). Experimental studies on the methods of long-term storage in the Nile are H.E. HURST, R.P. BLACK, and Y.M. SIMAIKA, *Long-Term Storage* (1965); and E. LOMBARDINI, *Essai sur l'hydrologie du Nil* (1865). Other important works in this area are J. LOZACH, *Le Delta du Nile* (1935); E. LUDWIG, *Der Nil* (1935; Eng. trans., 1939); H.G. LYONS, *The Physiography of the River Nile and Its Basin* (1906); and M. MACDONALD, *Nile Control,* 2 vol. (1920). H.H. JOHNSTON, *The Nile Quest* (1903), is a good introduction to the river's history. P. JOLLOIS, *Journal d'un ingénieur attaché à l'expédition d'Egypte, 1798–1802* (1904), presents an account of the difficulties encountered by the French scientists in their explorations up the Nile. A. MOORHEAD wrote two interesting books: *The White Nile* (1960) and *The Blue Nile* (1962), which provide a comprehensive study of the exploration and political history of the Nile. S. RAPPOPORT, *History of Egypt,* 3 vol. (1904), gives a historical account of the early expeditions of discovery with a detailed description of the land and people at the time.

(M.M.El-K.)

Nilotic Sudan, History of the

The Sudan generally is a zone that extends across the African continent south of the Sahara and north of the equatorial rain forest. The Nilotic Sudan, in this article, is understood to be the area along the Nile River south of Egypt, roughly equivalent to ancient Nubia and to the northern and central parts of the present Democratic Republic of The Sudan. Historically, this region has been one of the main points of contact between sub-Saharan Africa and the areas to the north.

This article is divided into the following sections:

I. The Nilotic Sudan to 1821

ANCIENT NUBIA TO THE 4TH CENTURY AD

The origins of Nubian culture. The earliest inhabitants of the Nilotic Sudan can be traced to Negroid peoples who lived in the vicinity of Khartoum, the Sudan, in Mesolithic (Middle Stone Age) times (30,000–20,000 BC). They were hunters and gatherers who made pottery and (later) objects of ground sandstone. Toward the end of the Neolithic (Late Stone Age; 10,000–3,000 BC) they had domesticated animals. These Negroid peoples were clearly in contact with predynastic civilizations (before *c.* 3100 BC) to the north in Egypt, but the arid uplands separating them from Nubia appear to have discouraged the Egyptians from settling there. At the end of the 4th millennium BC, kings of Egypt's 1st dynasty conquered upper Nubia beyond Aswān, introducing Egyptian cultural influence to a non-Negroid people who were scattered along the riverbank. In subsequent centuries, Nubia was subjected to successive military expeditions from Egypt in search of slaves or building materials for royal tombs, which destroyed much of the Egyptian-Nubian culture that had sprung from the initial conquests of the 1st dynasty. Throughout these five centuries (2800–2300 BC), the descendants of the Nubians continued to eke out an existence on the Nile, an easy prey to Egyptian military expeditions.

Migrations from Libya

Sometime after *c.* 2181, in the period known to Egyptologists as the First Intermediate Period (*c.* 2160–*c.* 2040), a new wave of immigrants entered Nubia from Libya, in the west, where the increasing desiccation of the Sahara drove them to settle along the Nile as cattle farmers. Other branches of these people seem to have gone beyond the Nile to the Red Sea Hills, while still others pushed south and west to Wadai and Darfur. These newcomers were able to settle on the Nile and assimilate the existing Nubians without opposition from Egypt. After the fall of the 6th dynasty (*c.* 2181), Egypt experienced a century and a half of weakness and internal strife, giving the immigrants in Nubia time to develop their own distinct civilization with unique crafts, architecture, and social structure, virtually unhindered by the potentially more dynamic civilization to the north. With the advent of the 11th dynasty (2133), however, Egypt recovered its strength and pressed southward into Nubia, at first sending only sporadic expeditions to exact tribute, but by the 12th dynasty (1991–1786) effectively occupying Nubia

as far south as Semna. The Nubians resisted the Egyptian occupation which was only maintained by a chain of forts erected along the Nile. Egyptian military and trading expeditions, of course, penetrated beyond Semna, and Egyptian fortified trading posts were actually established to the south at Karmah against frequent attacks upon Egyptian trading vessels by Nubian tribesmen beyond the southern frontier.

Egyptianization and the Kingdom of Kush. Despite the Egyptian presence in upper Nubia, the indigenous culture of the region continued to flourish, little changed by the proximity of Egyptian garrisons or the imports of luxury articles by Egyptian traders. Indeed, the Egyptianization of Nubia appears to have been enhanced during the decline in Egypt's political control over Nubia in the Second Intermediate Period (*c.* 1786–1567 BC), when Nubians were employed in large numbers as mercenaries against the Asian Hyksos invaders of Egypt. This experience did more to introduce Egyptian culture, which the mercenaries absorbed while fighting in Egyptian armies, than did the preceding centuries of Egyptian military occupation. The defeat of the Hyksos was the result of a national rising of the Egyptians who, once they had expelled the Hyksos from the Nile Valley, turned their energies southward to re-establish the military occupation of Nubia that the Hyksos invasion had disrupted. Under Thutmose I (1525–*c.* 1512 BC) the Egyptian conquest of the Northern Sudan was completed as far as Kurqus, 50 miles south of Abū Ḥamad, and subsequent Egyptian military expeditions penetrated even farther up the Nile. This third Egyptian occupation was the most complete and the most enduring, for despite sporadic rebellions against Egyptian control Nubia was deeply influenced by Egyptian culture. Nubia was divided into two administrative units: Wawat in the north, with its provincial capital at Aswān, and Kush (Cush) in the south, with its headquarters at Napata (Marawī). It was governed by a viceroy, usually a member of the royal entourage, who was responsible to the Egyptian king. Under him were two deputies, one for Wawat and one for Kush, and a hierarchy of lesser officials. The bureaucracy was staffed chiefly by Egyptians, but Egyptianized Nubians were not uncommon. Colonies of Egyptian officials, traders, and priests surrounded the administrative centres, but beyond these outposts the Nubians continued to preserve their own distinct traditions, customs, and crafts. To the Egyptians, Nubia remained a foreign land.

From J. Fage, *An Atlas of African History*; Edward Arnold (Publishers) Ltd.

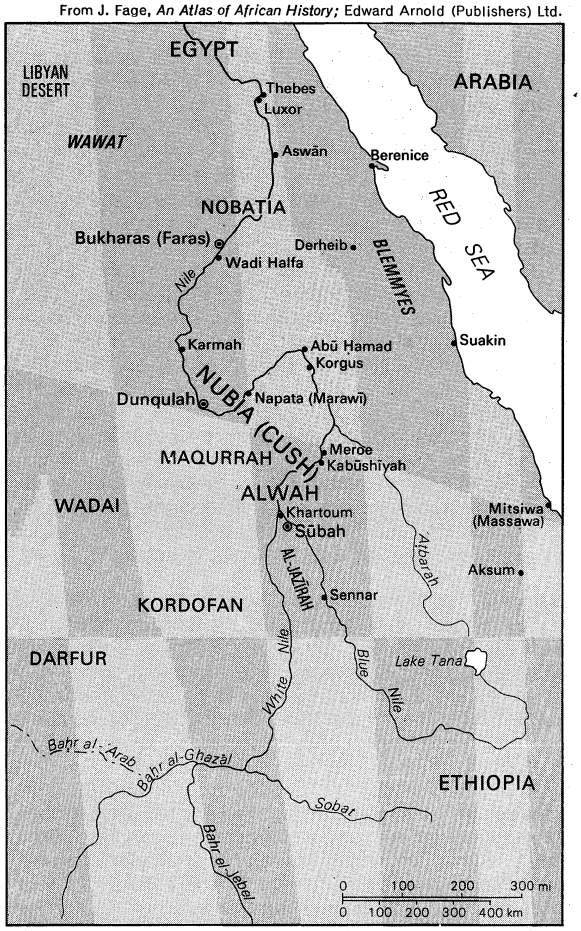


Figure 1: The Nilotic Sudan in ancient and medieval times.

But if it were a foreign land, it was also a rich one. Its position athwart the trade routes from Egypt to the Red Sea, and from the Nile to the south and west, brought great wealth from far-off places. Moreover, its cultivations along the Nile were rich, and in the hills the gold and emerald mines produced bullion and jewels for Egypt. The Nubians were also highly valued as soldiers. Nubia's wealth

As Egypt slipped once again into decline at the close of the New Kingdom (11th century BC), the viceroys of Kush, supported by their Nubian armies, became virtually independent kings, free of Egyptian control. By the 8th century BC, the kings of Kush came from hereditary ruling families of Egyptianized Nubian chiefs who possessed neither political nor family ties with Egypt. Under one such king, Kashta, Kush acquired control of Upper Egypt, and under his son Piankhi (*c.* 751–716 BC), the whole of Egypt to the shores of the Mediterranean was brought under the administration of Kush. As a world power, however, Kush was not to last. Just when the kings of Kush had established their rule from Abū Ḥamad to the Delta, the Assyrians invaded Egypt (671 BC) and with their superior iron-forged weapons defeated the armies of Kush under the redoubtable Taharqa; by 654 the Kushites had been driven back to Nubia and the safety of their capital, Napata.

Although reduced from a great power to an isolated kingdom behind the barren hills that blocked the southward advance from Aswān, Kush continued to rule over the Middle Nile for another thousand years, its unique Egyptian-Nubian culture preserved, while that of Egypt came under Persian, Greek, and Roman influences. Although Egyptianized in many ways, the culture of Kush was not simply Egyptian civilization in a Nubian environment. The Kushites developed their own language, expressed first by Egyptian hieroglyphs, then their own, and finally by a cursive script. They worshipped Egyptian gods but did not abandon their own. They buried their kings in pyramids but not in the Egyptian fashion. Their wealth continued to flow from the mines and to grow with their control of the trade routes. Soon after the retreat from Egypt the capital was moved from Napata southward to Meroe near Shandī, where the kingdom was increasingly exposed to the Negroid, African cultures farther south at the very time when its ties with Egypt were rapidly disappearing. The subsequent history of Kush is one of gradual decay, ending with inglorious extinction in AD 350 by the king of Aksum, who marched down from the Ethiopian highlands, destroyed Meroe, and sacked the decrepit towns along the river.

CHRISTIAN AND ISLAMIC INFLUENCE

Medieval Christian kingdoms. The two hundred years from the fall of Kush to the middle of the 6th century is an unknown age in the Sudan. Nubia was inhabited by a people called the Nobatae by the ancient geographers and the X-Group by modern archaeologists, who are still at a loss to explain their origins. The X-Group were clearly, however, the heirs of Kush, for their whole cultural life was dominated by Meroitic crafts and customs, and occasionally they even felt themselves sufficiently strong, in alliance with the nomadic Blemmyes (the Beja of the Eastern Sudan), to attack the Romans in Upper Egypt. When this happened, the Romans retaliated, defeating the Nobatae and Blemmyes and driving them into obscurity once again. When the Sudan was once more brought into the orbit of the Mediterranean world by the arrival of Christian missionaries in the 6th century, the Middle Nile was divided into three kingdoms: Nobatia, with its capital at Bukharas (modern Faras); Maqurrah, with its capital at Old Dunqulah; and the kingdom of ʿAlwah in the Nobatia, Maqurrah, and ʿAlwah

south, with its capital at Sūbah near Khartoum. Between 543 and 575 these three kingdoms were converted to Christianity by the work of Julian, a missionary who proselytized among the Nobatia (543–45), and his successor Longinus, who between 569 and 575 consolidated the work of Julian in Nobatia and even carried Christianity to ʿAlwah in the south. The new religion appears to have been adopted with enthusiasm. Christian churches sprang up along the Nile, and ancient temples were refurbished to accommodate Christian worshippers. After the retirement of Longinus, however, the Sudan once again receded into a period about which little is known, and it did not re-emerge into the stream of recorded history until the coming of the Arabs in the middle of the 7th century.

After the death of the Prophet Muḥammad in AD 632, the Arabs erupted from the desert steppes of Arabia and overran the lands to the east and to the west. Egypt was invaded in 639, and small groups of Arab raiders penetrated up the Nile and pillaged along the frontier of the Kingdom of Maqurrah, which by the 7th century had absorbed the state of Nobatia. Raid and counter-raid between the Arabs and the Nubians followed until a well-equipped Arab expedition under ʿAbd Allāh ibn Saʿd ibn Abī Sarḥ was sent south to punish the Nubians. The Arabs marched as far as Dunqulah, laid siege to the town, and destroyed the Christian cathedral. They suffered heavy casualties, however, so that when the King of Maqurrah sought an armistice, ʿAbd Allāh ibn Saʿd agreed to peace, happy to extricate his battered forces from a precarious position. Arab-Nubian relations were subsequently regularized by an annual exchange of gifts, trade relations, and the mutual understanding that no Muslims were to settle in Nubia and no Nubians were to take up residence in Egypt. With but few interruptions this peaceful, commercial relationship lasted for nearly six centuries, its very success undoubtedly the result of the mutual advantage that both the Arabs and the Nubians derived from it. The Arabs had a stable frontier; they appear to have had no designs to occupy the Sudan and were probably discouraged from doing so by the arid plains south of Aswān. Peace on the frontier was their object, and this the treaty guaranteed. In return, the Kingdom of Maqurrah gained another six hundred years of life.

Islāmic encroachments. When non-Arab Muslims acquired control of the Delta, friction arose in Upper Egypt. In the 9th century the Turkish Ṭūlūnid rulers of Egypt, wishing to rid themselves of the unruly nomadic Arab tribes in their domain, encouraged them to migrate southward. Lured by the prospects of gold in the Nubian Desert, the nomads pressed into Nubia, raiding and pillaging along borders, but the heartland of Maqurrah remained free from direct hostilities until the Mamlūks established their control over Egypt (1250). In the late 13th and early 14th centuries, the Mamlūk sultans sent regular military expeditions against Maqurrah, as much to rid Egypt of uncontrollable Arab Bedouins as to capture Nubia. The Mamlūks never succeeded in actually occupying Maqurrah, but they devastated the country, draining its political and economic vitality and plunging it into chaos and depression. By the 15th century Dunqulah was no longer strong enough to withstand Arab encroachment, and the country was open to Arab immigration. Once the Arab nomads, particularly the Juhaynah people, learned that the land beyond the Aswān reach could support their herds and that no political authority had the power to turn them back, they began to migrate southward, intermarrying with the Nubians and introducing Arabic, Muslim culture to the Christian inhabitants. The Arabs, who inherited through the male line, soon acquired control from the Nubians, who inherited through the female line, intermarriage resulting in Nubian inheritances passing from Nubian women to their half-Arab sons, but the Arabs replaced political authority in Maqurrah only with their own nomadic institutions. From Dunqulah the Juhaynah and others wandered east and west of the Nile with their herds; in the south, the Kingdom of ʿAlwah stood as the last indigenous Christian barrier to Arab occupation of the Sudan.

The Ṭūlūnids and Mamlūks

ʿAlwah extended from Kabūshīyah as far south as Sennar. Beyond, from the Ethiopian escarpment to the White Nile, lived Nilotic peoples about which little is known. ʿAlwah appears to have been much more prosperous and stronger than Maqurrah. It preserved the ironworking techniques of Kush, and its capital at Sūbah possessed many impressive buildings, churches, and gardens. Christianity remained the state religion, but its long isolation from the Christian world had probably resulted in bizarre and syncretistic accretions to liturgy and ritual. ʿAlwah was able to maintain its integrity so long as the Arabs failed to combine against it, but the continuous and corrosive raids of the Bedouins throughout the 15th century clearly weakened its power to resist. Thus, when an Arab confederation led by ʿAbd Allāh Jammāʿ was at last brought together to assault the Christian kingdom, ʿAlwah collapsed (*c.* 1500). Sūbah and the Blue Nile region were abandoned, left to the Funj, who suddenly appeared, seemingly from nowhere, to establish their authority from Sennar to the main Nile.

The fall of ʿAlwah

The Funj. The Funj were a strange and mysterious people. They were neither Arabs nor Muslims, and their homeland was probably on the upper Blue Nile in the borderlands between Ethiopia and The Sudan. Under their leader, ʿAmārah Dunqas, the Funj founded their capital at Sennar and throughout the 16th century struggled for control of the al-Jazīrah region against the Arab tribes who had settled around the confluence of the Blue and the White Niles. The Funj appear to have firmly established their supremacy by 1607–08. By the mid-17th century the Funj dynasty had reached its golden age under one of its greatest kings, Bādī II Abū Daqn (reigned 1644/45–80), who extended Funj authority across the White Nile into Kordofan and reduced the tribal chieftaincies scattered northward along the main Nile to tribute-paying feudatories. But as Bādī expanded Funj power, he also planted the seeds of its decline. During his conquests, slaves were captured and taken to Sennar, where, as they grew in numbers and influence, they formed a military caste. Loyal to the monarch alone, the slaves soon came to compete with the Funj aristocracy for control of the offices of state. Intrigue and hostility between these two rival groups soon led to open rebellion that undermined the position of the traditional ruling class. Under Bādī IV Abū Shulūkh (reigned 1724–62), the ruling aristocracy was finally broken, and the king assumed arbitrary power, supported by his slave troops. So long as Bādī IV could command the loyalty of his army, his position was secure and the kingdom enjoyed respite from internal strife, but at the end of his long reign he could no longer control the army. Under the leadership of his viceroy in Kordofan, Abū Likaylik, the military turned against the king and exiled him to Sūbah. Abū Likaylik probably represented a resurgence of older indigenous elements who had been Arabized and Islāmized but were neither Arab nor Funj. Henceforward, the Funj kings were but puppets of their viziers (chief ministers), whose struggles to win and to keep control precipitated the kingdom into steady decline, interrupted only by infrequent periods of peace and stability established by a strong vizier who was able to overcome his rivals. During its last half century the Funj kingdom was a spent state, kept intact only through want of a rival, but gradually disintegrating through wars, intrigue, and conspiracy, until the Egyptians advanced on Sennar in 1821 and pushed the Funj empire into oblivion.

Islāmization. The Funj were originally pagans, but the aristocracy soon adopted Islām and, although they retained many pagan customs, remained nominal Muslims. The conversion was largely the work of a handful of Islāmic missionaries who came to the Sudan from the larger Muslim world. The great success of these missionaries, however, was not among the Funj themselves but among the Arabized Nubian population settled along the Nile. Among these villagers the missionaries instilled a deep devotion to Islām that appears to have been conspicuously absent among the nomadic Arabs who first reached the Sudan after the collapse of the Kingdom of Maqurrah. One early missionary was Ghulām Allāh ibn

Early missionaries

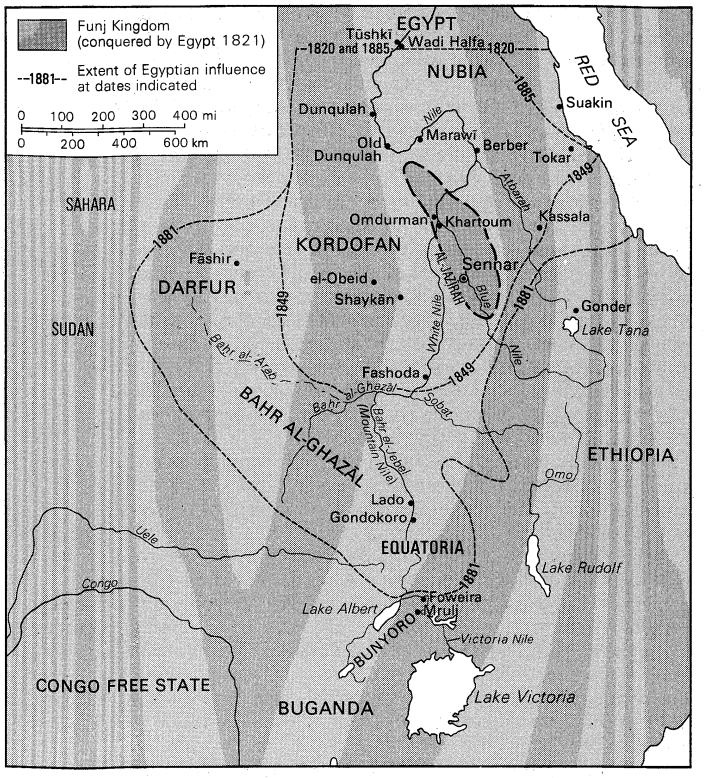


Figure 2: The Nilotic Sudan from the 17th to the 19th century.
From J. Fage, *An Atlas of African History;* Edward Arnold (Publishers) Ltd.

ʿĀʾid from the Yemen, who settled at Dunqulah in the 14th century. He was followed in the 15th century by Ḥamad Abū Danana, who appears to have emphasized the way to God through mystical exercises rather than through the more orthodox interpretations of the Qur'ān taught by Ghulām Allāh.

The spread of Islām was advanced in the 16th century, when the hegemony of the Funj enhanced security. In the 16th and 17th centuries, numerous schools of religious learning were founded along the White Nile and the Shāyqīyah were converted. Many of the more famous Sudanese missionaries who followed them were Ṣūfī holy men, members of influential religious brotherhoods who sought the way to God through mystical contemplation. Although the fervour of Sudanese Islām waned after 1700, the great reform movements that shook the Muslim world in the late 18th and early 19th centuries produced a revivalist spirit among the Ṣūfī brotherhoods, giving rise to a new order, the Mirghanīyah or Khatmīyah, later one of the strongest in the modern Sudan.

These men, called *fakīs*, attracted a following by their teachings and piety, and laid the foundations for a long line of indigenous Sudanese holy men who passed on the way to God taught them by their masters, or who founded their own religious schools, or who, if extraordinarily successful, gathered their own following into a religious order. The *fakīs* held a religious monopoly until the introduction, under Egyptian-Ottoman rule (see below), of an official hierarchy of jurists and scholars, the *ʿulamāʾ*, whose orthodox legalistic conception of Islām was as alien to the Sudanese as were their origins. This disparity between the mystical, traditional *fakīs*, close to the Sudanese, if not of them, and the orthodox, Islāmic jurists, aloof, if not actually part of the government bureaucracy, created a rivalry that in the past produced open hostility in times of trouble and sullen suspicion in times of peace. Recently, this schism has diminished; the *fakī* continues his customary practices unmolested, while the Sudanese have acknowledged the position of the *ʿulamāʾ* in society.

II. The Sudan since 1821

EGYPTIAN-OTTOMAN RULE

The administration of Muhammad ʿAli and his successors. In July 1820, Muḥammad ʿAlī, viceroy of Egypt under the Ottoman Turks, sent an army under his son Ismāʿīl to conquer the Sudan. Muḥammad ʿAli was interested in the gold and slaves that the Sudan could provide and wished to control the vast hinterland south of Egypt. By 1821 the Funj and the Sultan of Darfur had surrendered to his forces, and the Nilotic Sudan from Nubia to the Ethiopian foothills and from the ʿAṭbarah River to Darfur became part of his expanding empire.

Resistance to Egyptian rule

Collection of taxes under Muḥammad ʿAlī's regime amounted to virtual confiscation of gold, livestock, and slaves, and opposition to his rule became intense, eventually erupting into rebellion and the murder of Ismāʿīl and his bodyguard. But the rebels lacked leadership and coordination, and their revolt was brutally suppressed. A sullen hostility in the Sudanese was met by continued repression until the appointment of ʿAlī Khūrshīd Agha as governor general in 1826. His administration marked a new era in Egyptian-Sudanese relations. He reduced taxes and consulted the Sudanese through the respected Sudanese leader ʿAbd al-Qādir wad az-Zayn. Letters of amnesty were granted to fugitives. A more equitable system of taxation was implemented, and the support of the powerful class of holy men and *shaykhs* (tribal chiefs) for the administration was obtained by exempting them from taxation. But ʿAlī Khūrshīd was not content merely to restore the Sudan to its previous condition. Under his initiative trade routes were protected and expanded, Khartoum was developed as the administrative capital, and a host of agricultural and technical improvements were undertaken. When he retired to Cairo in 1838, Khūrshīd left a prosperous and contented country behind him.

His successor, Aḥmad Pasha abū Widān, with but few exceptions continued his policies, and made it his primary concern to root out official corruption. Abū Widān dealt ruthlessly with offenders or those who sought to thwart his schemes to reorganize taxation. He was particularly fond of the army, which reaped the benefits of regular pay and tolerable conditions in return for the brunt of the expansion and consolidation of Egyptian administration in Kassala and among the Baqqārah of southern Kordofan. Muḥammad ʿAlī, suspecting Abū Widān of disloyalty, recalled him to Cairo in the autumn of 1843, but he died mysteriously of poisoning before he left the Sudan.

During the next two decades the country stagnated because of ineffective government at Khartoum and vacillation by the viceroys at Cairo. If the successors of Abū Widān possessed administrative talent, they were seldom able to demonstrate it. Alarmed by the independent attitude of Abū Widān, Muḥammad ʿAlī first abolished the office of governor general and then, just as suddenly, revived it but sought to control its incumbents by limiting their tenure. Thus, no governor general held office long enough to introduce his own plans, let alone carry on those of his predecessor. New schemes were never begun, and old projects were allowed to languish. Without direction the army and the bureaucracy became demoralized and indifferent, while the Sudanese became disgruntled with the government. This state of affairs persisted until the more dynamic viceroy Ismāʿīl took over the guidance of Egyptian and Sudanese affairs in 1862.

Attempts to end the slave trade

During these quiescent decades, however, two ominous developments began that presaged future problems. Reacting to pressure from the Western powers, particularly Great Britain, the governor general of the Sudan was ordered to halt the slave trade. But not even the viceroy himself could overcome established custom with the stroke of a pen and the erection of a few police posts. If the restriction of the slave trade precipitated resistance among the Sudanese, the appointment of Christian officials to the administration and the expansion of the European Christian community caused open resentment. European merchants, mostly of Mediterranean origin, were either ignored or tolerated by the Sudanese and confined their contacts to compatriots within their own community and to the Turko-Egyptian officials whose manners and dress they frequently adopted. They became a powerful and influential group, whose lasting contribution to the Sudan was their lead taken in opening the

Southern Sudan to navigation and commerce, thereby bringing the vast Negroid, equatorial regions of the Upper Nile into the orbit of Sudanese history.

Ismaʿīl Pasha and the growth of European influence. In 1863, Ismāʿīl Pasha became viceroy of Egypt. Educated in Egypt, Vienna, and Paris, Ismāʿīl had absorbed the European interest in overseas adventures as well as Muḥammad ʿAlī's desire for imperial expansion, and had imaginative schemes for transforming Egypt and the Sudan into a modern state by employing Western technology. First he hoped to acquire the rest of the Nile Basin, including the Southern Sudan and the Bantu states by the great lakes of Central Africa. To finance this vast undertaking, and his projects for the modernization of Egypt itself, Ismāʿīl turned to those capital surplus nations of Europe, where investors were willing to risk their savings at high rates of interest in the cause of Egyptian and African development. But such funds would be attracted only as long as Ismāʿīl demonstrated his interest in reform by intensifying the campaign against the slave trade in the Sudan. Ismāʿīl needed no encouragement. He was genuinely opposed to the slave trade and made sincere efforts to suppress it and to cooperate with the European powers toward that end. Thus, these two major themes of Ismāʿīl's rule of the Nilotic Sudan—imperial expansion and the suppression of the slave trade—became intertwined, culminating in a third major development, the introduction of an ever-increasing number of European Christians to carry out the task of modernization.

Samuel Baker's expedition

In 1869 Ismāʿīl commissioned the Englishman Samuel Baker to lead an expedition up the White Nile to establish Egyptian hegemony over the equatorial regions of Central Africa and to curtail the slave trade on the Upper Nile. Baker remained in Equatorial Africa until 1873, where he established the Equatoria Province as part of Egyptian Sudan. He had extended Egyptian power and curbed the slave traders on the Nile, but he had also alienated certain African tribes, and being a rather tactless Christian, Ismāʿīl's Muslim administrators as well. Moreover, Baker had struck only at the Nilotic slave trade. To the west, on the vast plains of the Baḥr al-Ghazāl (now a province of the Democratic Republic of The Sudan), slave merchants had established enormous empires with stations garrisoned by slave soldiers. From these stations the long lines of human chattels were sent overland through Darfur and Kordofan to the slave markets of the Northern Sudan, Egypt, and Arabia. Not only did the firearms of the Khartoumers (as the traders were called) establish their supremacy over the peoples of the interior but those merchants with the strongest resources gradually swallowed up lesser traders until virtually the whole of the Baḥr al-Ghazāl was controlled by the greatest slaver of them all, az-Zubayr Raḥma Manṣūr, more commonly known as Zobeir Pasha. So powerful had he become that in 1873, the year Baker retired from the Sudan, the viceroy (now called the khedive) appointed az-Zubayr governor of the Baḥr al-Ghazāl. Ismāʿīl's officials had failed to destroy az-Zubayr as Baker had crushed the slavers on the Victoria Nile, and to elevate az-Zubayr to the governorship was the only way to establish at least the nominal sovereignty of Cairo over that enormous province. Thus, the agents of az-Zubayr continued to pillage the Baḥr al-Ghazāl under the Egyptian flag, while officially Egypt extended its dominion to the tropical rain forests of the Congo.

Gordon's administration

Ismāʿīl next offered the governorship of the Equatoria Province to another Englishman, Charles George Gordon, who in China had won fame and the sobriquet Chinese Gordon. Gordon arrived in Equatoria in 1874. His object was the same as Baker's—to consolidate Egyptian authority in Equatoria and to establish Egyptian sovereignty over the kingdoms of the great lakes—but his means were considerably more pacific. He reasserted government control over the stations that had been reoccupied by the slave traders since Baker's departure and stopped the slavers' raids against the local tribes. But Gordon's goal was the lakes, not the river, and he also sought to make the kingdoms of Bunyoro and Buganda (in present Uganda) recognize Egyptian sovereignty. Like Baker, he failed. Although he established stations beyond the Victoria Nile in Bunyoro proper at Foweira and Mruli, his resources were never sufficient to accomplish by force what he could not achieve by peaceful negotiation. When Gordon retired from Equatoria, the lake kingdoms remained stubbornly independent.

In 1877 Ismāʿīl appointed Gordon governor general of the Sudan. Gordon was the first European, Christian governor general of the Sudan. He returned there intending to lead a crusade against the slave trade, and to assist him in this humanitarian enterprise, he surrounded himself with a cadre of European and American Christian officials. In 1877 Ismāʿīl had signed the Anglo-Egyptian Slave Trade Convention, which provided for the termination of the sale and purchase of slaves in the Sudan by 1880. Gordon set out to fulfill the terms of this treaty and in whirlwind tours through the country broke up the markets and imprisoned the traders. His European subordinates did the same in the provinces.

The effects of the anti-slavery campaign

Gordon's crusading zeal blinded him to his invidious position as a Christian in a Muslim land and obscured from him the social and economic effects of arbitrary repression. Not only did his campaign create a crisis in the Sudan's economy but the Sudanese soon came to believe that the crusade, led by European Christians, violated the principles and traditions of Islām. By 1879 a strong current of reaction against Gordon's reforms was running through the country. The powerful slave-trading interests had, of course, turned against the administration, while the ordinary villagers and nomads, who habitually blamed the government for any difficulties, were quick to associate economic depression with Gordon's Christianity. And then suddenly, in the middle of rising discontent in the Sudan, Ismāʿīl's financial position collapsed. In difficulties for years, he could now no longer pay the interest on the Egyptian debt, and an international commission was appointed by the European powers to oversee Egyptian finances. After sixteen years of glorious spending, Ismāʿīl sailed away into exile. Gordon resigned.

Gordon left a perilous situation in the Sudan. The Sudanese were confused and dissatisfied. Many of the ablest senior officials, both European and Egyptian, had been dismissed by Gordon, departed with him, or died in his service. Castigated and ignored by Gordon, the bureaucracy had lapsed into apathy. Moreover, the office of governor general, on which the administration was so dependent, devolved upon Muḥammad Raʾūf Pasha, a mild man, ill-suited to stem the current of discontent or to shore up the structure of Egyptian rule, particularly when he could no longer count on Egyptian resources. Such then was the Sudan in June of 1881 when Muḥammad Aḥmad declared himself to be the Mahdī ("the divinely guided one").

THE MAHDIYAH

Muḥammad Aḥmad ibn ʿAbd Allāh was the son of a Dunqulah boatbuilder who claimed descent from the Prophet Muḥammad. Deeply religious from his youth, he was educated in one of the Ṣūfī orders, the Sammānīyah, but he became disgusted with the worldly ways of his teacher and secluded himself on Ābā Island in the White Nile to practice religious asceticism. In 1880 he toured Kordofan, where he learned of the discontent of the people and observed those actions of the government that he could not reconcile with his own religious beliefs. Upon his return to Ābā Island he clearly viewed himself as a *mujaddid*, a renewer of the Muslim faith, his mission to reform Islām and return it to the pristine form practiced by the Prophet. To Muḥammad Aḥmad the orthodox *ʿulamāʾ* who supported the administration were no less infidels than Christians, and when he later lashed out against misgovernment, he was referring as much to the theological heresy as to secular maladministration. Once he had proclaimed himself *mahdī* (a title traditionally used by Islāmic religious reformers) Muḥammad Aḥmad was regarded by the Sudanese as an eschatological figure, one who foreshadows the end of an age of darkness (which happened to coincide with the end of the 13th Muslim century) and heralds the beginnings of a new era

of light and righteousness. Thus, as a divinely guided reformer and symbol, Muḥammad Aḥmad fulfilled the requirements of Mahdiship in the eyes of his supporters.

The Mahdī's followers

Surrounding the Mahdī were his followers, the *anṣār*, and foremost among them was ʿAbd Allāh ibn Muḥammad, the *khalīfah* ("deputy"), who came from the Taʿāishah tribe of the Baqqārah Arabs and who assumed the leadership of the Mahdist state upon the death of Muḥammad Aḥmad. The holy men, the *fakis*, who for long had lamented the sorry state of religion in the Sudan brought on by the legalistic and unappealing orthodoxy of the Egyptians, looked to the Mahdī to purge the Sudan of the faithless ones. Also in his following, more numerous and powerful than the holy men, were the merchants formerly connected with the slave trade. All had suffered from Gordon's campaign against the trade, and all now hoped to reassert their economic position under the banner of religious war. Neither of these groups, however, could have carried out a revolution by themselves. The third and vital participants were the Baqqārah Arabs, the cattle nomads of Kordofan and Darfur who hated taxes and despised government. They formed the shock troops of the Mahdist revolutionary army, whose enthusiasm and numbers more than made up for its primitive technology. Moreover, the government itself only managed to enhance the prestige of the Mahdī by its fumbling attempts to arrest him and proscribe his movement. By September 1882, the Mahdists controlled all of Kordofan and at Shaykān on November 5, 1883, destroyed an Egyptian army of 10,000 men under the command of a British colonel. After Shaykān, the Sudan was lost, and not even the heroic leadership of Gordon, who was hastily sent to Khartoum, could save the Sudan for Egypt. On January 26, 1885, the Mahdists captured Khartoum and massacred Gordon and the defenders.

The reign of the Khalīfah. Five months after the fall of Khartoum, the Mahdī suddenly died on June 22, 1885. He was succeeded by the Khalīfah ʿAbd Allāh. The Khalīfah's first task was to secure his own precarious position among the competing factions in the Mahdist state. He frustrated a conspiracy by the Mahdī's relatives, and disarmed the personal retinues of his leading rivals in Omdurman, the Mahdist capital of the Sudan. Having curtailed the threats to his rule, the Khalīfah sought to accomplish the Mahdī's dream of a universal *jihād* (holy war) to reform Islām throughout the Muslim world. With a zeal compounded from a genuine wish to carry out religious reform, a desire for military victory and personal power, and an appalling ignorance of the world beyond the Sudan, the forces of the Khalīfah marched to the four points of the compass to spread Mahdism and extend the domains of the Mahdist state. By 1889 this expansionist drive was spent. In the west the Mahdist armies had achieved only an unstable occupation of Darfur. In the east they had defeated the Ethiopians, but the victory produced no permanent gain. In the Southern Sudan the Mahdists had scored some initial successes but were driven from the Upper Nile in 1897 by the forces of the Congo Free State of Leopold II of Belgium. On the Egyptian frontier in the north the *jihād* met its worst defeat at Tūshkī in August 1889, when an Anglo-Egyptian army under General Francis W. Grenfell destroyed a Mahdist army led by ʿAbd ar-Raḥmān an-Nujūmī. The Mahdist state, having squandered its resources on the *jihād*, a period of consolidation and contraction followed, necessitated by a sequence of bad harvests resulting in famine, epidemic, and death. Between 1889 and 1892 the Sudan suffered its most devastating and terrible years, as the Sudanese sought to survive on their shrivelled crops and emaciated herds. After 1892 the harvests improved and food was no longer in short supply. Moreover, the autocracy of the Khalīfah had become increasingly acceptable to most Sudanese, and having tempered his own despotism and eliminated the gross defects of his administration he too received the widespread acceptance, if not devotion, that the Sudanese had accorded the Mahdī.

In spite of its many defects, the Khalīfah's administration served the Sudan better than its many detractors would admit. Certainly the Khalīfah's government was autocratic, but while autocracy may be repugnant to European democrats, it was not only understandable to the Sudanese but appealed to their deepest feelings and attitudes formed by tribe, religion, and past experience with the centralized authoritarianism of the Turks. For them, the Khalīfah was equal to the task of governing bequeathed him by the Mahdī. Only when confronted by new forces from the outside world, of which he was ignorant, did ʿAbd Allāh's abilities fail him. His belief in Mahdism, his reliance on the superb courage and military skill of the *anṣār*, and his own ability to rally them against an alien invader were simply insufficient to preserve his independent Islāmic state against the overwhelming technological superiority of Great Britain. And as the 19th century drew to a close, the rival imperialisms of the European powers brought the full force of this technological supremacy against the Mahdist state.

The British conquest. British forces invaded and occupied Egypt in 1882 to put down a nationalist revolution hostile to foreign interests, and remained there to prevent any further threat to the khedive's government or the possible intervention of another European power. The consequences of this were far-reaching. A permanent British occupation of Egypt required the inviolability of the Nile waters without which Egypt could not survive, not from any African state, who did not possess the technical resources to interfere with them, but from rival European powers who could. Consequently, the British government, by diplomacy and military manoeuvres, negotiated agreements with the Italians and the Germans to keep them out of the Nile Valley. They were less successful with the French, who wanted them to withdraw from Egypt. Once it became apparent that the British were determined to remain, the French cast about for means to force the British from the Nile Valley; in 1893 an elaborate plan was concocted by which a French expedition would march across Africa from the west coast to Fashoda (Kodok) on the Upper Nile, where it was believed a dam could be constructed to obstruct the flow of the Nile waters. After inordinate delays, the French Nile expedition set out for Africa in June 1896, under the command of Capt. Jean-Baptiste Marchand.

British interest in the Nile Valley

As reports reached London during 1896 and 1897 of Marchand's march to Fashoda, Britain's inability to insulate the Nile Valley became embarrassingly exposed. British officials desperately tried one scheme after another to beat the French to Fashoda. They all failed, and by the autumn of 1897 British authorities had come to the reluctant conclusion that the conquest of the Sudan was necessary to protect the Nile waters from French encroachment. In October an Anglo-Egyptian army under the command of General Sir Horatio Herbert Kitchener was ordered to invade the Sudan. Kitchener pushed steadily but cautiously up the Nile. His Anglo-Egyptian forces defeated a large Mahdist army at the ʿAṭbarah River on April 8, 1898. Then, after spending four months preparing for the final advance to Omdurman, Kitchener's army of about 25,000 troops met the massed 60,000-man army of the Khalīfah outside the city on September 2, 1898. By midday the battle was over. The Mahdists were decisively defeated with heavy losses, and the Khalīfah fled, to be killed nearly a year later. Kitchener did not long remain at Omdurman but pressed up the Nile to Fashoda with a small flotilla. Here on September 18, 1898, he met Captain Marchand, who declined to withdraw—the long expected Fashoda crisis had begun. Both the French and British governments prepared for war. Neither the French army nor the navy was in any condition to fight, however, and the French were forced to give way. An Anglo-French agreement of March 1899 stipulated that French expansion eastward in Africa would stop at the Nile watershed.

THE ANGLO-EGYPTIAN CONDOMINIUM

The early years of British rule. Having conquered the Sudan, the British now had to govern it. But the administration of this vast land was complicated by the legal and diplomatic problems that had accompanied the conquest.

The Sudan campaigns had been undertaken by the British to protect their imperial position as well as the Nile waters, yet the Egyptian treasury had borne the greater part of the expense, and Egyptian troops had far outnumbered those of Britain in the Anglo-Egyptian Army. The British, however, did not simply want to hand the Sudan over to Egyptian rule; most Englishmen were convinced that the Mahdīyah was the result of sixty years of Egyptian oppression. To resolve this dilemma the Anglo-Egyptian condominium was declared in 1899 whereby the Sudan was given separate political status in which sovereignty was jointly shared by the khedive and the British crown, and the Egyptian and the British flags were flown side by side. The military and civil government of the Sudan was invested in a governor general appointed by the khedive of Egypt but nominated by the British government. In reality, there was no equal partnership between Britain and Egypt in the Sudan. From the first the British dominated the condominium and set about pacifying the countryside and suppressing local religious uprisings, which created insecurity among British officials but never posed a major threat to their rule. The north was quickly pacified and modern improvements were introduced under the aegis of civilian administrators, who began to replace the military as early as 1900. In the south, resistance to British rule was more prolonged; administration there was confined to keeping the peace rather than making any serious attempts at modernization.

Wingate's administration

The first governor general was Lord Kitchener himself, but in 1899 his former aide, Sir Reginald Wingate, was appointed to succeed him. Wingate knew the Sudan well, and during his long tenure as governor general (1899–1916) became devoted to its people and their prosperity. His tolerance and trust in the Sudanese resulted in policies that did much to establish confidence in Christian British rule by a devoutly Muslim, Arab-oriented people.

Modernization was slow at first. Taxes were purposely kept light, and the government consequently had few funds available for development. In fact, the Sudan remained dependent on Egyptian subsidies for many years. Nevertheless, railways, telegraph, and steamer services were expanded, particularly in al-Jazīrah, in order to launch the great cotton-growing scheme that remains today the backbone of the Sudan's economy. In addition, technical and primary schools were established, including the Gordon Memorial College, which opened in 1902 and soon began to graduate a Western educated elite that was gradually drawn away from the traditional political and social framework. Scorned by the British officials, who preferred the illiterate but contented fathers to the ill-educated, rebellious sons, and adrift from their own customary tribal and religious affiliations, these Sudanese turned for encouragement to Egyptian nationalists; from that association Sudanese nationalism in this century was born.

Its first manifestations occurred in 1921, when ʿAlī ʿAbd al-Laṭīf founded the United Tribes Society, and was arrested for nationalist agitation. In 1924 he formed the White Flag League, dedicated to driving the British from the Sudan. Demonstrations followed in Khartoum in June and August and were suppressed. When the governor general, Sir Lee Stack, was assassinated in Cairo on November 19, 1924, the British forced the Egyptians to withdraw from the Sudan, and annihilated a Sudanese battalion that mutinied in support of the Egyptians. The Sudanese revolt was ended, and British rule remained unchallenged until after World War II.

The growth of national consciousness. In 1936 Britain and Egypt had reached a partial accord in the Anglo-Egyptian treaty that enabled Egyptian officials to return to the Sudan. Although the traditional Sudanese *shaykhs* and chiefs remained indifferent to the fact that they had not been consulted in the negotiations over this treaty, the educated Sudanese elite were resentful that neither Britain nor Egypt had bothered to solicit their opinions.

The Graduates' General Congress

Thus, they began to express their grievances through the Graduates' General Congress, which had been established as an alumni association of Gordon Memorial College and soon embraced all educated Sudanese. At first the Graduates' General Congress confined its interests to social and educational activities, but with Egyptian support the organization demanded recognition by the British to act as the spokesman for Sudanese nationalism. The Sudan government refused, and the Congress split into two groups: a moderate majority prepared to accept the good faith of the government, and a radical minority, led by Ismāʿīl al-Azharī, turned to Egypt. By 1943 Azharī and his supporters had won control of the Congress and organized the Ashiqqā' ("Brothers"), the first genuine political party in the Sudan. Seeing the initiative pass to the militants, the moderates formed the Ummah ("Nation") Party under the patronage of Sayyid ʿAbd ar-Raḥmān al-Mahdī, the posthumous son of the Mahdī, with the intention of cooperating with the British toward independence. Sayyid ʿAbd ar-Raḥmān had inherited the allegiance of the thousands of Sudanese who had followed his father. He now sought to combine to his own advantage this power and influence with the ideology of the Ummah. His principal rival was Sayyid ʿAlī al-Mirghanī, the leader of the Khatmīyah brotherhood. Although he personally remained aloof from politics, Sayyid ʿAlī threw his support to Azharī. The competition between the Azharī-Khatmīyah faction—remodelled in 1951 as the National Unionist Party (NUP)—and the Ummah-Mahdist group quickly rekindled old suspicions and deep-seated hatreds that soured Sudanese politics for years and eventually strangled parliamentary government.

Although the Sudan government had crushed the initial hopes of the congress, the British officials were well aware of the pervasive power of nationalism among the elite and sought to introduce new institutions to associate the Sudanese more closely with the task of governing. An Advisory Council was established for the Northern Sudan consisting of the governor general and 28 Sudanese, but Sudanese nationalists soon began to agitate to transform the Advisory Council into a legislative one in which the Southern Sudanese would be included. The British had facilitated their control of the Sudan by segregating the pagan or Christian Africans of the south from the Muslim Arab northerners. The decision to establish a legislative council forced them to abandon this policy; in 1947 they permitted southern participation in a legislative council to represent the Sudan as a whole.

The creation of this council produced a strong reaction on the part of the Egyptian government, which in October 1951 unilaterally abrogated the Anglo-Egyptian agreement of 1936 and proclaimed Egyptian rule over the Sudan. These hasty and ill-considered actions only managed to alienate the Sudanese from Egypt until the Nasser-Naguib revolution in July 1952 placed men with more understanding of Sudanese aspirations in power in Cairo. Not only did they seek to regain Egyptian influence in the Sudan but on February 12, 1953, signed an agreement with Great Britain granting self-government for the Sudan and self-determination within three years for the Sudanese. Elections for a representative Parliament to rule the Sudan followed in November and December 1953. The Egyptians threw their support behind Ismāʿīl al-Azharī, the leader of the National Unionist Party, who campaigned on the slogan "Unity of the Nile Valley." This position was opposed by the Ummah Party, which had the less vocal but pervasive support of British officials. To the shock of many British officials and to the chagrin of the Ummah, which had enjoyed power in the Legislative Council for nearly six years, Ismāʿīl al-Azharī's NUP won an overwhelming victory. Although Azharī had campaigned to unite the Sudan with Egypt, the realities of disturbances in the Southern Sudan and the responsibilities of political power and authority ultimately led him to disown his own campaign promises and to declare The Sudan an independent republic with an elected representative Parliament on January 1, 1956.

The end of the condominium

THE REPUBLIC OF THE SUDAN

The triumph of liberal democracy in The Sudan was short-lived. Compared to the strength of tradition, which still shaped the life of the Sudanese, the liberalism imported from the West, disseminated through British edu-

cation and adopted by the Sudanese intelligentsia, was a weak force. At first parliamentary government had been held in high esteem; it was the symbol of nationalism and independence, signifying the nation's coming of age and freedom from alien rule. But at best Parliament was a superficial instrument. It had been introduced into The Sudan at precisely the time when parliamentary forms were rapidly disappearing from other countries in the Middle East. Parties, the machinery by which parliamentary government functions, were not well-organized groups with distinct objectives, but loose alliances attached opportunistically to personal interests and sectarian loyalties. Such groups were difficult to manage, almost impossible to direct. When the tactics of party management were exhausted, Parliament became debased, benefitting only those politicians who reaped the rewards of power and patronage. Disillusioned with their experiment in liberal democracy, the Sudanese turned once again to the authoritarianism to which their traditions had accustomed them.

The Abboud government. On the night of November 16, 1958, the commander in chief of the Sudan army, Gen. Ibrahim Abboud, carried out a bloodless coup d'etat and took charge of the government. The following day Abboud himself broadcast to the nation, blaming the "state of degeneration" on the political strife between rival factions. He dissolved all political parties, prohibited assemblies, and temporarily suspended newspapers. The country was henceforth governed by a Supreme Council of the Armed Forces, consisting of 12 senior officers. Parliament was abolished, the transitional constitution suspended, and a state of emergency proclaimed. Only the Communists opposed the military regime, but they were powerless against the army, and despite several attempts by dissident officers to seize control of the Supreme Council, at the end of its first year in power the Abboud government was securely in control.

Economic improvements

Despite these internal struggles, army rule brought rapid improvements in The Sudan's deteriorating economic position. The Abboud government at once abolished the fixed price on cotton and within six months had sold all the Sudanese cotton. Although disposed of at lower rates, the ultimate effect of the cotton sale was to give The Sudan a surplus revenue and dramatically rebuild the nation's foreign reserves. The other achievement of the military government was the conclusion of a Nile waters agreement with the United Arab Republic on November 8, 1959. Although both the monetary compensation and the share of the Nile waters that The Sudan received have proven insufficient, the Nile waters agreement was more significant than the inadequate technical division implies, for the United Arab Republic not only recognized but appeared to be reconciled to an independent Sudan on its southern frontier astride the water upon which Egypt depends.

The problem of the south. In the Southern Sudan Abboud's policies were less successful. Partly in fear of another rebellion and partly restricted by the moderation imposed upon them under a parliamentary system, the politicians had left the south to follow its own ways. The army officers of the military government were under no such restraints. In the name of national unity they introduced numerous measures designed to facilitate the spread of Islām and of the Arabic language. Important positions in the administration and police were staffed by northerners. The Southern Liberal Party, which represented the views of the core of southern intellectuals, was proscribed. Education was shifted from the English curriculum of the Christian missionaries, who had long been solely responsible for education in the south, to an Arabic, Islāmic orientation. Mosques and Islāmic schools were established under the direction of the department of religious affairs. Foreign Christian missionaries were increasingly restricted and finally expelled in February and March 1964.

In the Southern Sudan itself, the measures of the central government were greeted by ever-increasing resistance. In October 1962 a widespread strike in southern schools resulted in antigovernment demonstrations followed by a general flight of students and others over the border. During the spring and summer of 1963 feelings relaxed and affairs appeared to improve, most of the discontented having fled, and The Sudan government reinforced its army and police units in the southern provinces. Suddenly, however, in September 1963, rebellion again erupted in eastern Equatoria and in the Upper Nile Province led by the Anya Nya, a Southern Sudanese guerrilla organization. Bitterly opposed to the policies of the government and disenchanted with the ineffectual attempts by the more moderate southern leaders to reach a settlement, the Anya Nya declared open hostility against the Northern Sudanese in the belief that only violent resistance would make the government of General Abboud seek a solution acceptable to the southerners. Taking refuge in the illimitable bush, the guerrillas continued their sporadic attacks on isolated posts of the Sudan army, while the Southern Sudanese carried on their way of life much as before, wandering off when the pressure of government became too great, returning when the weight of administration relaxed. Without confidence, however, there was little growth or progress in any field. Thus, while economic and educational developments transformed the north, the south remained a stagnant backwater. Blind to the aspirations of the Southern Sudanese and devoid of imagination in their dealings with them, the generals in Khartoum sought to establish their authority by repression, which increased in proportion to southern discontent.

The Anya Nya revolt

Resistance in the north. Although the Northern Sudanese had little sympathy for their countrymen in the south, the intelligentsia was able to use the government's failure there to assail authoritarian rule in the north and to revive demands for democratic government that the military coup d'etat had brought to an end. By 1962 numerous urban elements, including the intelligentsia, the trade unions, and the civil servants, as well as the powerful religious brotherhoods, had become alienated from the military regime. The intelligentsia resented its exclusion from the councils of government, the trade unions chafed at the restrictions placed upon their activities, and the civil servants sulked at orders from their military ministers. Even the more conservative religious brotherhoods grew restless when they were unable to carry on their former political activities. Moreover, the tribal masses and growing proletariat had become increasingly apathetic toward the government. Military reviews, parades, and heroic pronouncements were no substitute for the enthusiasm of party politics and the passions stirred by political action. Even if they considered the problem, the military rulers never provided an outlet for the political frustrations of the Sudanese, and in the end the regime was overwhelmed by boredom and overthrown by the reaction to its lassitude. The means of its overthrow was the southern problem.

In October 1964 students at the University of Khartoum held a meeting in defiance of a government prohibition in order to condemn publicly government action in the Southern Sudan and to denounce the regime. In the ensuing clash with police one student was killed. Larger demonstrations followed, and with most of its forces committed in the Southern Sudan, the military regime was unable to maintain control. The disorders soon spread and General Abboud, unwilling to crush the disturbances with massive repression, resigned as head of state, and a transitional government was appointed to govern under the provisional constitution of 1956.

The Sudan since 1964. Under the leadership of Sirr al-Khātim al-Khalīfah, the transitional government held elections in April and May 1965 to determine an elected, representative government. A coalition government led by a leading Ummah politician, Muḥammad Aḥmad Maḥjūb, was formed in June 1965. As before, parliamentary government was characterized by factional disputes. On the one hand, Maḥjūb enjoyed the support of the traditionalists within the Ummah Party represented by the imam al-Hādī, the spiritual successor to Sayyid ʿAbd ar-Raḥmān al-Mahdī, while he was challenged by the Sayyid Ṣādiq al-Mahdī, the young great grandson of the Mahdī, who led the more progressive forces within the

Ummah. Unable to find common objectives to unite their differences, the Parliament failed to deal with the economic, social, and constitutional problems in The Sudan. Moreover, the earlier hopes expressed by the transitional government of cooperation with the southerners soon vanished before northern intransigence and southern indecision, disunity, and paucity of leadership. Conflict continued in the south, and with little hope of resolution dragged on tragically from one year to the next. Having no workable constitution, a stagnant economy, a political system torn by secretarian interests, and a continuing civil conflict in the Southern Sudan, a group of young officers led by Col. Gaafar an-Nimeiry seized the government on May 25, 1969. For the second time the army had seized control of The Sudan in an effort to resolve the formidable problems that had defied resolution by the politicians.

Resumption of military control

BIBLIOGRAPHY. The early history of the Sudan is best presented by P.L. SHINNIE in *Meroe: A Civilization of the Sudan* (1967); and YUSUF FADL HASAN, *The Arabs and the Sudan* (1967). The age of the Funj is most comprehensively covered by O.G.S. CRAWFORD in *The Fung Kingdom of Sennar* (1951), supported by ANDREW PAUL, *A History of the Beja Tribes of the Sudan* (1954). The best recent history of the Sudan from the Funj sultanate to the present is P.M. HOLT, *A Modern History of the Sudan* (1961), which is admirably suited to the general reader. Holt's account may be supplemented by more specialized studies, particularly R.L. HILL, *Egypt in the Sudan, 1820–1881* (1959); and Holt's own study of the Mahdist period, *The Mahdist State in the Sudan, 1881–1898* (1958). The reconquest of the Sudan resulted in a great outpouring of historical literature. The campaigns themselves are brilliantly narrated by WINSTON S. CHURCHILL in *The River War: An Account of the Reconquest of the Soudan*, 2 vol. (1899). The formulation of British policy that led to the Anglo-Egyptian reconquest is adequately analyzed by MÉKKI SHIBEIKA in *British Policy in the Sudan, 1882–1902* (1952); G.N. SANDERSON, *England, Europe, and the Upper Nile, 1882–1899* (1965); and in the decade immediately following the Fashoda crisis by ROBERT O. COLLINS, *King Leopold, England, and the Upper Nile, 1899–1909* (1968). There is no adequate single history of the Anglo-Egyptian condominium, although SIR HAROLD A. MACMICHAEL, *The Sudan* (1954); and J.S.R. DUNCAN, *The Sudan* (1952) and *The Sudan's Path to Independence* (1957), present a general survey. Thus the student must seek the story of Anglo-Egyptian rule in a host of biographies, memoirs, and personal reminiscences. The transitional period of self-government and the vicissitudes of independence have been narrated by K.D.D. HENDERSON, *Sudan Republic* (1965). The Southern Sudan has its own historical literature, the most useful being RICHARD GRAY, *A History of the Southern Sudan, 1839–1889* (1961); ROBERT O. COLLINS, *The Southern Sudan, 1883–1898* (1962) and *Land Beyond the Rivers: The Southern Sudan, 1898–1918* (1971); and J. ODUHO and W. DENG, *The Problem of the Southern Sudan* (1963).

(R.O.C.)

Nineveh

Nineveh, the most populous and the oldest city in Assyria, lay on the east bank of the Tigris opposite modern Mosul (in Iraq). From time immemorial, roads from the foothills of Kurdistan debouched there, and a tributary of the Tigris, the Khawṣar River, added to the value of the fertile agricultural and pastoral lands in the district. The first to survey and map Nineveh was the archaeologist Claudius J. Rich in 1820, a work later completed by Felix Jones and published by him in 1854. Excavations have been undertaken intermittently since that period by many persons. A.H. (later Sir Henry) Layard during 1845–51 discovered the palace of Sennacherib and took back to England an unrivalled collection of stone basreliefs together with thousands of tablets inscribed in cuneiform from the great library of Ashurbanipal. Hormuzd Rassam continued the work in 1852. During 1929–32 R. Campbell Thompson excavated the temple of Nabu (Nebo) on behalf of the British Museum and discovered the site of the palace of Ashurnasirpal II. In 1931–32, together with M.E.L. (later Sir Max) Mallowan, Thompson for the first time dug a shaft from the top of the Quyunjik (Acropolis), 90 feet (30 metres) above the level of the plain down through strata of accumulated debris of earlier cultures to virgin soil. It was then proved that over four-fifths of this great accumulation is prehistoric.

Early settlements in Neolithic times

The first settlement, a small Neolithic hamlet, was probably founded not later than the 7th millennium BC. An obsidian-blade industry suggests that there was contact from the beginning with Van in eastern Armenia. Hassuna-Sāmarrā' and Tall Halaf painted pottery of the subsequent Early Chalcolithic phases, characteristic of the north, was succeeded by gray wares such as occur westward in the Jabal Sinjār. Farmers during the 4th millennium used clay sickles of a type found in the Ubaid period, and these imply contact with the south.

One of the most remarkable discoveries that Mallowan and Thompson made within the mound of Quyunjik in the prehistoric strata consisted of many hundreds of roughly made, bevelled bowls, overturned in the soil and filled with vegetable matter. On the analogy of much later medieval deposits they may have been intended as magical offerings to expel evil spirits from houses. Their typology conforms exactly with that of Uruk (Erech) pottery, which was widespread throughout the Tigris-Euphrates Valley in the late 4th millennium. In these levels also large metal vases occur, again characteristic of southern Babylonia, and there can be no doubt that technologically this district of the Tigris had much in common with the cities of the lower Euphrates Valley at this period. This similarity is of particular interest because it indicates that some time before 3000 BC, a period of economic prosperity had united the commercial interests of north and south in an altogether exceptional manner, for later these two civilizations diverged widely.

A little before and after 3000 BC, unpainted Ninevite pottery was similar to that used at Sumerian sites such as Ur and Uruk (biblical Erech); to approximately the same period belongs a series of attractively painted and artistically incised ware known as Ninevite V, which is a home product distinct from that of the south. Hoards of beads found in these strata may be dated *c.* 2900 BC.

Metal sculpture from the 3rd millennium

The most remarkable object of the 3rd millennium BC is a realistic bronze head—life-size, cast, and chased—of a bearded monarch. This, the finest piece of metal sculpture ever recovered from Mesopotamia, may represent the famous King Sargon of Akkad (*c.* 2334–*c.* 2279 BC). This bronze head, however (now in the Iraq Museum, Baghdad), because of its brilliant technique and elaborately modelled features, is thought by some authorities to belong to a rather later stage of the Akkadian Period (*c.* 2334–*c.* 2154 BC); if so, the head might represent King Naram-Sin (*c.* 2254–*c.* 2218 BC). The hypothesis for the earlier period seems preferable, for metal work advanced more rapidly in style in Mesopotamia at that period than did stone sculpture, and it is known from inscriptions that Sargon's second son, Manishtusu, had built the temple of E-Mashmash at Nineveh by virtue of being the "son of Sargon"; thus a model of the founder of the dynasty would have been appropriately placed there.

Surprisingly, there is no large body of evidence to show that Assyrian monarchs built at all extensively in Nineveh during the 2nd millennium BC: certainly not during the four centuries that succeeded Shamshi-Adad I (*c.* 1813–*c.* 1781 BC), when Assyria was of little account because of the power of the superior Hittite, Kassite, and Mittanian dynasties. An interesting historical document, however, describes a victorious campaign of Ashur-uballit I (ruled *c.* 1365–*c.* 1330 BC) against a Kassite usurper, during a period of Assyrian renascence. The fame of Ishtar of Nineveh had indeed reached the ears of the Pharaoh before that, for her statue was sent to Egypt by Tushratta, king of Mitanni, to restore the Pharaoh's health. Later monarchs whose inscriptions have appeared on the Acropolis include Shalmaneser I and Tiglath-pileser I, both of whom were active builders in Ashur; the former had founded Calah (Nimrūd).

But Nineveh had to wait for the neo-Assyrians, particularly from the time of Ashurnasirpal II (ruled 883–859 BC) onward, for a considerable architectural expansion. Thereafter successive monarchs kept in repair and founded new palaces, temples to Sin, Nergal, Nanna, Shamash, Ishtar, and Nabu (Nebo). Unfortunately, severe depredations have left few remains of these edifices.

It was Sennacherib who made Nineveh a truly magnificent city (*c.* 700 BC). He laid out fresh streets and squares and built within it the famous "palace without a rival," the plan of which has been mostly recovered and has overall dimensions of about 600 by 630 feet (180 by 190 metres). It comprised at least 80 rooms, of which many were lined with sculpture. A large part of the famous "K" collection of tablets was found there (see below); some of the principal doorways were flanked by human-headed bulls. At this time the total area of Nineveh comprised about 1,800 acres (700 hectares), and 15 great gates penetrated its walls. An elaborate system of 18 canals brought water from the hills to Nineveh, and several sections of a magnificently constructed aqueduct erected by the same monarch were discovered at Jerwan about 25 miles (40 kilometres) distant.

His successor Esarhaddon built an arsenal in the Nabī Yūnus mound, south of Quyunjik, and either he or his successor set up statues of the pharaoh Taharqa (Tarku) at its entrance as trophies to celebrate the conquest of Egypt. These were discovered by Fuad Safar and Muḥammad ʿAlī Muṣṭafā on behalf of the Iraqi Department of Antiquities in 1954.

Ashurbanipal's library and the "K" collection

Ashurbanipal later in the 7th century BC constructed a new palace at the northwest end of the Acropolis. He also founded the great library and ordered his scribes to collect and copy ancient texts throughout the country. The "K" collection included more than 20,000 tablets or fragments of tablets and incorporated the ancient lore of Mesopotamia. The subjects are literary, religious, and administrative, and a great many tablets are in the form of letters. Branches of learning represented include mathematics, botany, chemistry, and lexicology. The library contains a mass of information about the ancient world and will exercise scholars for generations to come.

Fourteen years after the death of Ashurbanipal, however, Nineveh suffered a defeat from which it never recovered. Extensive traces of ash, representing the sack of the city by Babylonians, Scythians, and Medes in 612 BC, have been found in many parts of the Acropolis. From the ruins it has been established that the perimeter of the great Assyrian city wall was about 7.5 miles (12 kilometres) long and in places up to 148 feet (45 metres) wide; there was also a great unfinished outer rampart, protected by a moat, and the Khawṣar River flowed through the centre of the city to join the Tigris on the western side of it.

The 15 great gates that intersected the Acropolis walls were built partly of mud-brick and partly of stone. The long eastern sector, about three miles (five kilometres), contained six gates; the southern sector, 2,624 feet (800 metres), contained only one, the Ashur Gate; the western sector, about 2.5 miles (four kilometres), had five gates; the northern sector, about 1.2 miles (two kilometres), three gates, Adad, Nergal, and Sin. Several of these entrances are known to have been faced with stone colossi (lamassu). In the Nergal Gate two winged stone bulls, attributable to Sennacherib, have been reinstalled: a site museum has been erected adjacent to it by the Iraqi Department of Antiquities. The Adad Gate contained many inscribed tiles, and what may prove to be the Sin Gate contained a corridor that led through an arched doorway into a ramp or stairwell giving access to the battlements.

Most impressive was the Shamash Gate, which has been thoroughly excavated by Tariq Madhloum on behalf of the Iraqi Department of Antiquities. It was found to have been approached across two moats and a water course by a series of bridges in which the arches were cut out of the natural conglomerate. The wall was faced with limestone and surmounted by a crenellated parapet, behind which ran a defense causeway. The structure was built with mud as well as burnt bricks, which bore the stamp of Sennacherib. There was an entrance 14.8 feet (4.5 metres) wide in the centre of a long, projecting bastion, which was further strengthened by six towers. Crudely incised stone slabs on the inner side of the gateway depicted the burning of a tower; it is possible that these carvings represented the fall of Nineveh and are post-Assyrian. The internal plan of the gate includes six great chambers lined with uncarved orthostats (upright slabs), which were discovered by Layard and Rassam.

Archaeologists also have been active within the Quyunjik (Acropolis). Since 1966 Madhloum has been restoring the throne room of Sennacherib's palace and some of the adjoining chambers. He found that all the entrances to the two main chambers were flanked by winged bull colossi, and he recovered a series of orthostats not recorded by any of the 19th-century excavators. One such slab illustrates a foreign city, heavily defended by towers, surrendering to the Assyrian army. He discovered a stone paved bathroom adjoining the throne room and, in the great antehall, no fewer than 40 carved orthostats. The subjects represented include Sennacherib's campaigns against mountain-dwelling peoples, besieged cities, and units of the Assyrian army.

Decline of Nineveh

After 612 BC the city ceased to be important, although there are some Seleucid and Greek remains. Xenophon in the *Anabasis* recorded the name of the city as Mespila. In the 13th century AD the city seems to have enjoyed some prosperity under the *atabegs* of Mosul. Subsequently, houses continued to be inhabited at least as late as the 16th century AD. In these later levels imitations of Chinese wares have been found.

BIBLIOGRAPHY. A.H. LAYARD, *Nineveh and Its Remains, with an Enquiry into the Manners and Arts of the Ancient Assyrians*, 2nd ed., 2 vol. (1849); *Discoveries in the Ruins of Nineveh and Babylon, with Travels in Armenia, Kurdistan and the Desert* (1853); R. CAMPBELL THOMPSON, R.W. HUTCHINSON, and M.E.L. MALLOWAN, in *Liverpool Annals of Archaeology and Anthropology*, 18:79–112, 19:55–116, 20:71–186 (1931–33), containing accounts of the excavations on the site of the palace of Ashurnasirpal in 1929–30, on the site of the Temple of Ishtar in 1931–32, and in the mound of Quyunjik; R. CAMPBELL THOMPSON and R.W. HUTCHINSON, *A Century of Exploration of Nineveh* (1929); J.P.G. FINCH, "The Winged Bulls at the Nergal Gate of Nineveh," *Iraq*, 10:9–18 (1948); TARIQ MADHLOUM, "Excavations at Nineveh: A Preliminary Report (1965–1967)," *Sumer*, 23:76–82 (1967), a short account of the excavations and restoration work carried out in 1965–67 necessitated by the development and rapid growth of the city of Mosul; and "Nineveh (1967–1968) Campaign," *Sumer*, 24:45–52 (1968), containing the results of further excavation work in 1967–68.

(M.Mn.)

Ningsia Hui

Ningsia Hui (Ning-xia Hui in Pin-yin romanization), an autonomous region of the People's Republic of China, is bounded on the east in part by Shensi and on the east, south, and west by Kansu, and on the north by the Inner Mongolian Autonomous Region. About one-third of its population of 2,200,000 in 1970 was Hui (Chinese Muslim). Most of the region is desert, but the vast plain of the Huang Ho (Yellow River) in the north has been irrigated for agriculture for centuries. The total area of the autonomous region is 65,600 square miles (170,000 square kilometres). Its capital is Yin-ch'uan.

Ningsia Hui is nearly coextensive with the ancient kingdom of the Tangut people, known in China as the Hsi Hsia; after its conquest by Genghis Khan, it was named Ningsia (Peaceful Hsia).

History. Irrigation canals on the Ningsia plains of the Huang Ho dating from the time of the Ch'in, Han, and T'ang dynasties (*i.e.*, from about 220 BC) provide evidence that the area has long been inhabited.

In the 11th century AD the area became part of the kingdom of Tangut (Hsi Hsia) in western China. Yin-ch'uan, its chief city and now the capital, was captured by Genghis Khan early in the 13th century and remained tributary to China. In 1914 the area became a part of the province of Kansu, and in 1928 it was constituted as the province of Ningsia. In 1954, under the People's Republic of China, the province was again merged with Kansu; in 1956 that part of Ningsia largely occupied by Mongolians became a part of the Inner Mongolian Autonomous Region; and in 1958 the rest of the former province became the Ningsia Hui Autonomous Region.

Physical geography. Physiographically, the Ningsia Hui region can be divided into two parts. Southern Ning-

sia Hui is part of the Loess Plateau, with the Liu-p'an Shan (Liu-p'an Mountains) as the main ridge. The region is covered with a thick layer of loess, in some places over 300 feet deep, and the topography is generally fairly flat. Northern Ningsia Hui is made up for the most part of the Ningsia plain of the Huang Ho, which enters Ningsia Hui from the Tsinghai plateau in Kansu and flows east, then north into Inner Mongolia. On an elevation of 3,600–3,900 feet (1,100–1,200 metres), the plain slopes gradually from south to north. West of the plain are the Ho-lan Shan. These mountains serve as a shelter against the sandstorms from the Ala Shan, a desert to the west. In the piedmont belt, a few coal mines have been opened.

With an annual precipitation of only eight inches (200 millimetres), the Ningsia plain is an arid area, but the Huang Ho provides irrigation. Many canals have been built over the centuries. The network of willow-lined canals and paddy fields gives the landscape a look resembling that of southern China. The climate is continental. Temperatures range from an annual average maximum of 80° F (27° C) to an annual average minimum of 7° F (−14° C).

Ethnic composition of the people

Population. The ethnic composition of Ningsia Hui includes the Han (Chinese-speaking Chinese), Hui (Chinese Muslims), Mongolians, and Manchus. Out of the total population of about 2,200,000 (1970), about 700,000, or one-third, were Hui and 1,000,000 were Han. The Mongolians and Manchus, combined, numbered about 500,000. The people speak Mandarin Chinese, Tibetan, and Mongolian, and the predominant religions are Islām or Buddhism. Islām is well developed and has the most believers, including the Hui. In the late 1940s, before the Communists came into power, it was estimated that Muslims comprised 60 percent of the total population.

The region is predominantly rural, with most of the population engaged in pasturing and farming the land. One of China's more sparsely settled areas, Ningsia Hui has a population density of about 34 persons per square mile (13 per square kilometre). In the widely scattered cities, residents have traditionally been devoted to handicrafts. In recent years, however, more workers have begun to be employed in mining and manufacturing.

Administration. Until 1958 the Hui minority of the region was organized into two autonomous districts (*tzu-chih-chou*) and one autonomous county (*tzu-chih-hsien*). Wu-chung Hui Autonomous District had a population of 230,000, about 62 percent of whom were Hui. Its capital was at Wu-chung. Ku-yüan Hui Autonomous District had a population of 220,000. Its capital was at Ku-yüan. Finally, there was the Ching-yüan Hui Autonomous County, on the southern border of the Ku-yüan District, with its seat in the former county town of Hwa-pingling (now known as Ching-yüan).

In 1958 the two autonomous districts and the Ching-yüan County became part of the Ningsia Hui Autonomous Region. In 1969 the region acquired a Mongol banner (district)—A-la-shan East. The banner had its headquarters at A-la-shan-tso-ch'i; its annexation increased the total area from 25,000 to about 65,000 square miles. Local government is organized in accord with the Chinese Communist Party plan.

Social conditions. Ningsia Hui was formerly a backward area in education. In 1935 there was not a single college in the province. There were only two high schools; two normal schools, with about 425 students; and about 200 elementary schools, with 12,600 students.

Advances in education

About 85 percent of the children did not attend schools. Since the Communist government was organized in 1949, much improvement has been reported. It was estimated that by 1958 illiteracy had been reduced by more than 60 percent. In 1965 the region was reported to have 3,000 elementary schools, 185 high schools, and three colleges and special technical colleges, with a total enrollment of 289,000 students.

In the area of public health, there are reported to be about 860 hospitals.

Economy. The Ningsia plain produces abundant wheat and good-quality rice. Cash crops include sesame and sugar beet. In the mixed agricultural and pastoral areas, a good breed of sheep, a domesticated form of the argali of eastern Mongolia, is being raised. Its wool is soft, white, and lustrous. The Manchus especially have long been known for breeding and raising pigs. The principal crops are rice, wheat, millet, cotton, and sugar beets. Melons and apricots are also grown in quantity.

Mineral resources of Ningsia Hui are limited to coking-coal reserves in the P'ing-lo area, near the Inner Mongolian border. Coal was mined on a small scale in the past but has been expanded since the construction of the Pao-t'ou–Lan-chou railroad in 1958. Well-known in ancient times as a border city on the western frontier of China, the capital, Yin-ch'uan, used to be a trading centre for farm and animal products. Now medium-sized and small factories have been built there, including a farm-tool plant and a woolen-textile mill.

Transportation. Yin-ch'uan, the capital of the autonomous region, lies in the centre of the Ningsia plain. The Huang Ho, to the east, provides irrigation and water transportation facilities.

The Pao-t'ou–Lan-chou railway, an extension of the main Peking–Pao-t'ou railway completed in 1958, now links Yin-ch'uan to the two new industrial bases of Pao-t'ou (in Inner Mongolia, to the north) and Lan-chou (in Kansu, to the south) and helps the development of industry and agriculture in Ningsia Hui.

Cultural life. Hui cultural life is intimately interrelated with the Hui religion, Islām. Muslims, for example, never eat pork, and this religious practice has caused much trouble between the Hui and the Han, who, like the Manchus, are fond of pork. The Hui woman traditionally keeps house; her role is domestic, and she cannot have contact with outside work. When they go out, Hui women must wear the veil to conceal their faces, and they are forbidden to talk to males. Under the Communist regime, however, Hui women have had to do farm work in the communes and production work in the factories. The traditional culture is undergoing change.

BIBLIOGRAPHY. GEORGE B. CRESSEY, *Land of the 500 Million: A Geography of China* (1955), a standard geography text; *Asia's Lands and Peoples*, 3rd ed. (1963), a widely used college text; THEODORE SHABAD, *China's Changing Map*, 2nd ed. (1972), an up-to-date book on the geography of China, using the province as the unit; T.R. TREGEAR, *A Geography of China* (1965), a general geography text; CHIAO-MIN HSIEH, *China: Ageless Land and Countless People* (1967), a series on political geography emphasizing historical perspective; KEITH M. BUCHANAN, *The Chinese People and the Chinese Earth* (1966), a geography text using Communist information; OWEN LATTIMORE, *Inner Asian Frontiers of China* (1950), a classic study; SIR AUREL STEIN, *Innermost Asia: Its Geography As a Factor in History, Geogr. J.*, 65:377–403, 473–501 (1925); section on Ningsia Hui in the HUMAN RELATION AREA FILES, *A Regional Handbook for Northwest China* (1956), a useful collection.

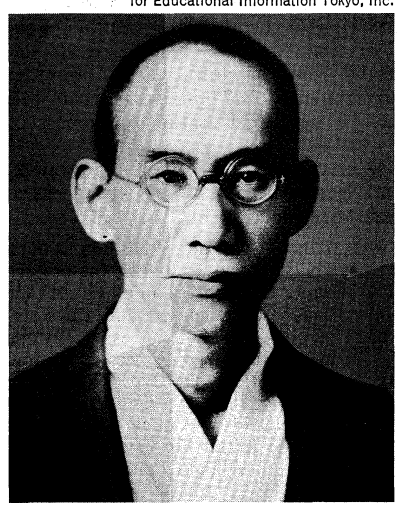
(C.–M.H.)

Nishida Kitarō

Nishida Kitarō was the outstanding Japanese philosopher of the last 100 years, during which period the Japanese endeavoured to assimilate Western philosophy and to produce original works based on the Oriental spiritual tradition. Nishida was both an outstanding example of and influence upon this cultural development.

Nishida's life. He was born on April 19, 1870, in a little village near Kanazawa in Ishikawa Prefecture. His father, Nishida Yasunori, was for a time a teacher of the elementary school in the village, among whose few pupils was Kitarō. His mother, Tosa, was a pious devotee of the Jōdo, or "True Pure Land," school of Buddhism. He respected his mother very highly and cherished her memory. Although Nishida's family was descended from a former village landowner, Yasunori ruined his fortune when Kitarō was young, and all of the Nishidas had to move to Kanazawa in 1883. Kitarō entered primary course at Kanazawa Normal School in that year but had to leave on account of sickness in the following year. He was admitted in 1886 into the second

class of the high school, and in 1888 he became a student of the Fourth Higher School (junior college). In his boyhood, Nishida took traditional lessons in Chinese from an excellent Confucian teacher, and in his higher school days he was taught by another good scholar erudite in Chinese. Another important teacher of Nishida's was Hōjō Takiyoshi, a professor of mathematics of the Fourth Higher School, under whom Nishida had studied mathematics even before he entered high school. Such a training and refinement of his personality through Chinese culture enriched his life with a lasting Confucian quality and worldview. Later, when Western philosophy and Buddhism (especially Zen Buddhism) were merged in his mature mind, there remained deep within him an undercurrent of Confucian conviction with regard to "the ideal man," "the Way" to good and truth, sincerity, self-negation, and detachment. He and his contemporaries belonged to the last Japanese generation whose education in the Chinese Classics molded their personal character. From his boyhood days, he made several good friends in Kanazawa, among whom was D.T. Suzuki, later an eminent Buddhologist and the main interpreter of Zen Buddhism to the West. Nishida and Suzuki became classmates at higher school, and from that time their mutual spiritual influence continued till Nishida's death.



Nishida Kitarō.

In his memoirs, entitled "A Certain Professor's Statement upon Retirement (from Kyōto University, December 1928)," he writes:

> My student days at the Fourth Higher School were the happiest of my life. I was filled with youthful zest. I did anything I wished, heedless of the consequences. As a result I had to leave school before my graduation. At the time I thought it was not necessarily true that one could not achieve anything by studying alone. In fact I thought it would be better to rid myself of the fetters of school and to read freely. But within one year I was prohibited to read any more by my doctor, since I was afflicted with an eye disease. I had to abandon my principle, and went to Tokyo to be a non-regular student of philosophy, in Tokyo University (1891–1894).

At that time in the Faculty of Letters and Law, there were several promising students who later became famous, some as men of letters and others as university professors. Together with them, Nishida appeared at the same lectures but could not form close friendships, as he had done in higher school. After graduation he became a teacher in a middle school near his home (1895). In the following year, he was appointed a lecturer in the Fourth Higher School at Kanazawa; and after two years as a lecturer and later as a professor at the Yamaguchi Higher School in Yamaguchi, he was again appointed as a professor of the Fourth Higher School, teaching psychology, logic, ethics, and German (1899–1909). During his Yamaguchi and Kanazawa teaching periods, he was much engaged in the practice of Zen meditation. Remarks about Zen practice are overwhelmingly conspicuous in his diary of this period. From this effort and through his lectures at the higher school came Nishida's maiden work, *Zen-no-Kenkyū*, (1911; Eng. trans., *A Study of Good*, 1960). At about this time parts of the book were published in Japanese philosophical journals, and his name as an original philosopher attracted attention in the Japanese philosophical world.

After one year as professor at Gakushūin University (Tokyo) in 1909, he was appointed associate professor of ethics at Kyōto University. In 1913 he was appointed professor of philosophy of religion and in 1914 professor of philosophy, a post he held until his retirement in 1928. About the end of his professorship in Kyōto University, Nishida's philosophy attained its maturity, which can be defined as "the philosophy of the *topos* [place] of Nothingness." In his latter years he delved most deeply into philosophical problems and endeavoured to explain more concrete facts by his logic. Thus his idea of the true reality that overcomes the dichotomy of subjectivity and objectivity (the mind and its objects) in the *topos* of Nothingness became significant, he emphasized, for "historical reality in the historical world." Nishida developed this implication of absolute Nothingness in his *Tetsugakulon-Bunshū* ("Philosophical Essays"; 7 vols.), which he wrote after his retirement. In his last days, as World War II was coming to an end, looking at the fires of the burning cities in the darkness of night, Nishida was much inspired by the words of the prophets of Israel in the Old Testament. He said that the result of this war might be something that neither the victors nor the vanquished could at that time foresee. He died on June 7, 1945, at Kamakura.

Philosophy of Nothingness

Nishida's philosophy was attacked by the nationalists and militarists during World War II because of the Western way of his philosophical thinking. Since then he has been criticized for his loyalty to his nation and for his alleged metaphysical obscurantism by Marxist philosophers and antimetaphysical Rationalist philosophers, but such criticism is decreasing for lack of solid grounds. More philosophically important are the criticisms by Takahashi Satomi and Tanabe Hajime. Takahashi was the first scholar to appreciate and evaluate the distinctively Japanese philosophy in Nishida's *Zen-no-Kenkyū*, and later he contributed his critical investigation of Nishida's philosophy in its mature form. Tanabe, Nishida's disciple, who succeeded him in the chair of professor of philosophy at Kyōto University (1927–45), contributed valuable criticism from his own philosophical point of view.

The three stages of Nishida's thought. Nishida says in his memoirs that he thought of his life in terms of a change of position with the blackboard as an axis; in the first half of his life he sat at a desk facing the blackboard, while in the latter half he sat with the blackboard behind him. Continuing this metaphor, it may be said that in the third stage (see below), represented by his philosophy of the *topos* of Nothingness, he wished to relinquish both positions, whether in front of or behind the blackboard, so that he and his logic became chalk on the blackboard of the historical world. In Nishida's philosophy, each stage has its independent value. Like a series of vortexes floating on the stream when two (*i.e.*, Western and Eastern) rivers converge, each closes its own circle. The preceding system should not be replaced by the later, even if they flow successively.

In the first stage of his philosophy, Nishida derived his basic insights from his long, concentrated practice of Zen. He was also much inspired by William James's philosophy and psychology, and tried to interpret his own basic insights philosophically with the use of psychological concepts borrowed from James. The opening page of Nishida's *Zen-no-Kenkyū* indicates the general direction of his thought:

> To experience means to know events precisely as they are. It means to cast away completely one's attitude of discriminative reflection, and to know in accordance with the events. Since people include some reflection even when speaking of experience, the word "pure" is here used to

His notion of pure experience

signify a condition of true experience itself without the addition of the least thought or reflection. For example, it refers to that moment of seeing a color or hearing a sound which occurs not only before one has added the judgment that this seeing or hearing relates to something external or that one is feeling this sensation, but even before one has judged what color or what sound it is. Thus, *pure experience* is synonymous with *direct experience*. When one experiences directly one's conscious state there is as yet neither subject nor object, and knowledge and its object are completely united. This is the purest form of experience.

The concept of pure experience expounded here is the Western philosophical mold into which Nishida poured his own religious experience cultivated by his Zen training. As it is beyond the dichotomy of subject and object, so it is far removed from the difference of whole and part. The whole universe is, as it were, crystallized into one's own being. In the total activity of one's own pure and alert life, one's entire being becomes transparent, so that it reflects, as in a mirror, all things as they become and also participates in them. This is "to know in accordance with the events." The profoundness of reality, the directness of one's experience of reality, a dynamic system developing itself in the creative stream of consciousness —these are the characteristic motifs of Nishida's philosophy, all suggesting where his thinking was ultimately rooted. According to Nishida, judgment is formed by analysis of the intuitive whole. For instance, the judgment that a horse runs is derived from the direct experience of a running horse. The truth of a judgment is grounded on the truth of the original intuitive whole from which the judgment is formed through the dichotomy of subject and predicate or that of subject and object. For the establishment of its truth a judgment is, through its dichotomy itself, referred back to intuition as its source, because intuition is here considered a self-developing whole, similar to Hegel's Notion (*Begriff*). As Hegel says, "All is Notion," or "All is judgment," so could Nishida say, "All (reality) is intuition," or "All reality is immediate consciousness." For this is practically the import of his dictum, "Consciousness is the Unique Reality."

In the second stage of his philosophy, Nishida was under the influence of the philosophy of Henri Bergson, a French philosopher, which he tried to synthesize with a Neo-Kantian type of German thought that was then prevalent in Japanese philosophical circles. He thus entered the second stage of his thinking, the result of which was incorporated into a book entitled *Jikaku-ni okeru chokkan to hansei* (1917, "Intuition and Reflection in Self-consciousness"). His basic notion did not undergo any change, but he tried to express what he once called pure experience in a different way. Neo-Kantian influence led him to eliminate from his thought all psychological terms and to follow strictly the path of logical thinking to the end. Actually, however, he found himself standing at the end of a blind alley, where he came up against something that was impenetrable to his logic. "After a long struggle with the Unknowable my logic itself bade me surrender to the camp of mysticism," so he himself says in the preface. Thus the self as the unity of thought and intuition acquires a mystical background. It is pure activity but ultimately finds itself in the abyss of darkness, enveloping every light of self-consciousness. This darkness, however, is "dazzling obscurity" giving the self the unfathomable depth of meaning and being. The self is thus haloed with a luminous darkness.

The "Non-self" and individuality

The third stage of Nishida's philosophy was marked by a reversal of his whole procedure, as is shown in his *Hataraku-mono kara miru-mono e* (1927, "From the Acting to the Seeing Self"). Whereas he had always made the self the starting point for his philosophical thinking, he now parted definitely with Transcendental Idealism (see IDEALISM) or, rather, broke through it to find behind it a realm of reality corresponding to his own mystical experience. This may be called the realm of Non-self, or Nothingness, which should not be confused with the non-self of Idealism as the realm of the objective over against that of the subjective, or with annihilating nothingness of Sartre's Existentialism. The "Non-self" of Nishida is the ultimate reality where all subject–object cleavage is overcome. In accordance with Buddhist tradition he called it "Nothingness" and sought to derive the individual reality of everything in the world, whether it be a thing or a self, from the supreme identity of Nothingness. The Idealist "pure self," as the universal consciousness or consciousness in general, is still abstract, while the "Non-self" of Nishida establishes itself as true individuality in the absolute Nothingness, which includes, not excludes, the individual reality of the thing-in-itself (the ultimate reality of things). Indeed, the problem of the individual now became Nishida's chief concern. In his quest for its solution he made an intensive study of Greek philosophy, especially Plato and Aristotle. He found the thinking of these philosophers to be relatively free from the cleavage of subject and object, in comparison with modern Western philosophy, which always presupposes, consciously or unconsciously, the *cogito* (the thinking subject) as the starting point of thought. The ontology of Plato and Aristotle rather makes a logic of reality reveal itself, a logic that explains the world of reality as seen from within. Whether "explaining" or "seeing," such a logic is to be understood as an act taking place in the world of reality itself. Nishida seeks thus to clarify the significance of the individual and the universal from the viewpoint of the Absolute Nothingness. Thus he propounds that Nothingness or *mu* is the universal that is to be sought behind the predicate as the universal concept and, at the same time, is the abyss of Nothingness in which the self as the individual is crystalized. He developed the idea of the "*topos* of Nothingness," adopting the concept of *topos* from Plato's *Timaeus* and from this time on Nothingness is explained as the uniqueness of the *topos*.

In the fourth stage of the development of his thought, Nishida applied the idea of the *topos* of Nothingness to the explanation of his "historical world."

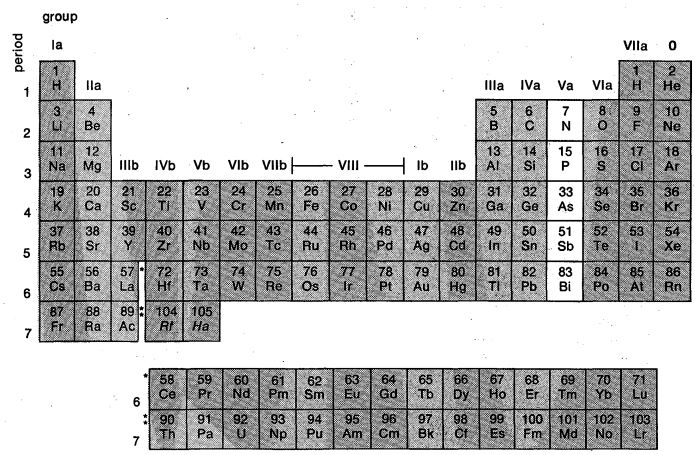
BIBLIOGRAPHY. SHIMOMURA TORATARO, "Nishida Kitaro and Some Aspects of His Philosophical Thought," the preface to Nishida's *A Study of Good* (1960), gives an excellent view of the history of modern Japanese philosophy and a description of the development of Nishida's thought, especially the early period, when Japan was undergoing westernization. VALDO H. VIGLIELMO, "Nishida Kitaro, The Early Years," in *Tradition and Modernization in Japanese Culture*, ch. 13 (1971), goes into details of Nishida's younger days (until 1903). DAVID A. DILWORTH, "The Initial Formations of 'Pure Experience' in Nishida Kitaro and William James," in *Monumenta Nipponica*, 24:93–111 (1969), treats the influence of William James on Nishida's thought, and compares the two philosophies. See also the same author's "The Range of Nishida's Early Religious Thought: Zen no Kenkyū," *Philosophy East and West*, 19:409–421 (1969); and "Nishida's Final Essay: The Logic of Place and Religious World-View," *ibid.*, 20:355–367 (1970); and NODA MATAO, "East-West Syntheses in Kitarō Nishida," *ibid.*, 4:345–359 (1955), which stresses the significance of Nishida's later thought as compared with the philosophical ideas of Whitehead. The following works of Nishida are available in English translation. *A Study of Good*, trans. by VALDO H. VIGLIELMO (1960), is Nishida's maiden work which gives a well-balanced treatment of philosophical problems. *Intelligibility and the Philosophy of Nothingness*, trans. by ROBERT SCHINZINGER (1958), and *Fundamental Problems of Philosophy*, trans. by DAVID A. DILWORTH (1970), include introductory remarks on Nishida's life and work. See also "Nishida Kitarō: The Problem of Japanese Culture," trans. by ABE MASAO in *Sources of the Japanese Tradition*, pp. 857–872 (1958); and GINO K. PIOVESANA, *Recent Japanese Philosophical Thought, 1862–1962* (1963).

(Y.T.)

Nitrogen Group Elements and Their Compounds

In the formal classification of the chemical elements known as the periodic table, the nitrogen group elements, consisting of nitrogen (N), phosphorus (P), arsenic (As), antimony (Sb), and bismuth (Bi), form a vertical file commonly designated group Va (see figure). These elements share certain general similarities in chemical behaviour, though they are clearly differentiated from one another

chemically, and these similarities reflect common features of the electronic structures of their atoms. For this reason, they are conveniently discussed together.



Nitrogen group elements in the periodic table of elements.

Probably no other group of the elements is more familiar to the layman than this group. Although the five elements together make up less than 0.2 percent by weight of the earth's crust, they assume an importance far out of proportion to their abundance. This is especially true of the elements nitrogen and phosphorus, which comprise 2.4 and 0.9 percent, respectively, of the total weight of the human body.

The nitrogen elements have, perhaps, the widest range in physical state of any group in the periodic table. Nitrogen, for example, is a gas that liquefies at about −200° C, whereas bismuth is a solid melting at 271° C and boiling at about 1,560° C. Chemically, too, the range in properties is wide, nitrogen and phosphorus being typical nonmetals; arsenic and antimony, metalloids; and bismuth, a metal. Even in appearance these elements exhibit great variety. Nitrogen is colourless both as a gas and as a liquid. Phosphorus exists in a variety of physical modifications, or allotropic forms, including the familiar white, highly reactive form that must be stored under water to prevent it from igniting in the air; a much less reactive red or violet form; and a black modification that, although least known, appears to be the most stable of all. Arsenic exists mainly as a dull gray metallic solid, but a more reactive yellow, solid form is also known, and there are indications that other forms exist under certain conditions. Antimony is a silver, metallic appearing, but somewhat brittle solid; and bismuth is a silver-white metal with a trace of pink in its lustre.

Physical appearance

Together with carbon, hydrogen, oxygen, and sulfur, the first two members of this group, nitrogen and phosphorus, are the principal chemical elements incorporated into living systems. Nitrogen and phosphorus are readily removed from the soil by plant growth, and therefore they are immensely important components of plant foods. Such designations as "5–10–5" on commercial fertilizers represent the respective weight percentage composition of the material in terms of nitrogen, phosphoric oxide, and potassium oxide (potassium being the third principal element needed for healthy plant growth). Nitrogen in fertilizers may be in the form of sodium or potassium nitrates, ammonia, ammonium salts, or various organic combinations. Phosphorus is supplied chiefly as inorganic phosphate.

Ironically enough, these same elements, nitrogen and phosphorus, can also be used in ways less helpful to man. The explosives in conventional warfare are heavily dependent on their content of nitrogen compounds, and the deadly nerve gases are composed of organic compounds of phosphorus.

On the other hand, arsenic, which is notorious for its toxicity, is most useful in agriculture, where its compounds are an aid in controlling harmful insect pests. Antimony and bismuth are used chiefly in metal alloys, because they impart unique and desirable properties to these alloys.

Beginning with a brief historical review, this article first studies carefully the fundamental nature of this group to see how its members are related to one another, and to reveal how this group fits into the periodic system as a whole. This study is then followed by a detailed survey of the individual elements and some of their most important compounds.

DISCOVERIES OF THE NITROGEN GROUP ELEMENTS

Nitrogen. Since four-fifths of the earth's atmosphere is nitrogen, the recognition of nitrogen as a specific substance came about during early investigations of oxygen. Carl Wilhelm Scheele, a Swedish druggist, showed in 1772 that air is a mixture of two gases, one of which he called "fire air," because it supported combustion, and the other "foul air," because it was left after the "fire air" had been used up. The "fire air" was, of course, oxygen, and the "foul air" nitrogen. At about the same time nitrogen also was recognized by a Scottish botanist, Daniel Rutherford, and by the controversial British clergyman, Joseph Priestley, who, with Scheele, is given credit for the discovery of oxygen. Later work showed the new gas to be a constituent of nitre, a common name for potassium nitrate (KNO_3); and, accordingly, it was named nitrogen by the French chemist Jean-Antoine-Claude Chaptal in 1790. Nitrogen also was considered a chemical element by Antoine-Laurent Lavoisier, whose explanation of the role of oxygen in combustion eventually overthrew the phlogiston theory, an erroneous view of combustion that became popular in the early 18th century. The inability of nitrogen to support life led Lavoisier to name it *azote*, still the French equivalent of nitrogen.

Early theories of air

Phosphorus. Arabian alchemists of the 12th century may have isolated elemental phosphorus by accident, but the records are unclear. Phosphorus appears to have been discovered in 1669 by Hennig Brand, a German merchant whose hobby was alchemy. Brand allowed 50 buckets of urine to stand until they putrified and "bred worms." He then boiled the urine down to a paste and heated it with sand, thereby distilling elemental phosphorus from the mixture. Brand reported his discovery in a letter to Gottfried Wilhelm Leibniz, and, thereafter, demonstrations of this element and its ability to glow in the dark, or "phosphoresce," excited public interest. Phosphorus, however, remained a chemical curiosity until about a century later when it proved to be a component of bones. Digestion of bones with nitric or sulfuric acid formed phosphoric acid, from which phosphorus could be distilled by heating with charcoal. In the late 1800's, James Burgess Readman of Edinburgh developed an electric furnace method for producing the element from phosphate rock, which is essentially the method employed today.

Arsenic. Arsenic was known in the form of certain of its compounds long before it was clearly recognized as a chemical element. In the 4th century BC Aristotle wrote of a substance called *sandarachē*, now believed to have been the mineral realgar, a sulfide of arsenic. Then, in the 1st century AD, the writers Pliny the Elder and Pedanius Dioscorides both described *auripigmentum*, a substance now thought to have been the dyestuff orpiment, As_2S_3. By the 11th century AD three species of "arsenic" were recognized: white (As_4O_6), yellow (As_2S_3), and red (As_4S_4). The element itself possibly was first observed in the 13th century by Albertus Magnus, who noted the appearance of a metallike substance when *arsenicum*, another name for As_2S_3, was heated with soap. It is not certain, however, that this natural scientist and scholar actually observed the free element. The first clearly authentic report of the free substance was made in 1649 by Johann Schroeder, a German pharmacist, who prepared arsenic by heating its oxide with charcoal. Later, Nicolas Lémery, a French physician and chemist, observed the formation of arsenic when heating a mixture of the oxide, soap, and potash. By the 18th century, arsenic was well known as a unique semimetal.

Antimony. The ancients were familiar with antimony both as a metal and in its sulfide form. Fragments of a

Some Properties of Nitrogen Group Elements

	nitrogen	phosphorus (white)	arsenic	antimony	bismuth
Atomic number	7	15	33	51	83
Atomic weight	14.0067	30.9738	74.9216	121.75	208.980
Colour of element	colourless	white	steel gray	silver	pinkish silver
Melting point (° C)	—209.86	44.1	817 (28 atm)	630.5	271.3
Boiling point (° C)	—195.8	280	613 (sublimes)	1,380	1,560
Density at 25° C (g/cm³)	1.25 (g/l)	1.82	5.73	6.684	9.8
Solubility in water (vol/vol water)	0.0231	none	none	none	none
Valence	3, (5)	3, 5	3, 5	3, 5	3, 5
Electronic configuration	$1s^2 2s^2 2p^3$	(Ne)$3s^2 3p^3$	(Ar)$3d^{10} 4s^2 4p^3$	(Kr)$4d^{10} 5s^2 5p^3$	(Xe)$4f^{14} 5d^{10} 6s^2 6p^3$
Isotopic abundance (terrestrial, percent)	^{14}N (99.63), ^{15}N (0.37)	^{31}P (100)	^{75}As (100)	^{121}Sb (57.25), ^{123}Sb (42.75)	^{209}Bi (100)
Radioactive isotopes (mass numbers)	12, 13, 16, 17	28–30, 32–34	68–74, 76–81, 85	112–120, 122, 124–135	196, 199–208, 210–215
Heat of fusion (cal/g)	6.1	5.03	88.5	38.3	12.5
Heat of vaporization (cal/g)	47.7	130	102	161	204.3
Specific heat (cal/g/° C)	0.249 (N_2)	0.189	0.082	0.0494	0.0294
Critical temperature (° C)	—147.1	675	—	—	—
Critical pressure (atm)	33.5	80	—	—	—
Electrical resistivity (microhm-cm)	—	1×10^{17}	33.3 (20° C)	39 (0° C)	106.8 (0° C)
Hardness (Mohs scale)	—	0.5	3.5	3.0–3.5	2.5
Crystal structure at 20° C	—	cubic	hexagonal	hexagonal	hexagonal
Radius					
Covalent (Å)	0.74	1.10	1.19	1.38	1.46
Ionic (M^{3+}, Å)	0.16	0.44	0.58	0.76	0.96
Ionization energy (first, kcal/mole)	337	255	228	201	170
Electronegativity					
(Pauling)	3.05	2.15	2.10	2.05	1.8
(Sanderson)	4.49	3.34	3.91	3.37	3.16

Chaldean vase made of antimony have been estimated to date from about 4000 BC. The Old Testament tells of Queen Jezebel using the naturally occurring sulfide of antimony to beautify her eyes. Pliny, during the 1st century AD, wrote of seven different medicinal remedies using *stibium* or antimony sulfide. Early writings of Dioscorides, dating from about the same time, mention metallic antimony. Records of the 15th century show the use of the substance in alloys for type, bells, and mirrors. In 1615 Andreas Libavius, a German physician, described the preparation of metallic antimony by the direct reduction of the sulfide with iron; and a later chemistry textbook by Lémery, published in 1675, also describes methods of preparation of the element. In the same century, a book summarizing available knowledge of antimony and its compounds was purportedly written by a Basil Valentine, allegedly a Benedictine monk of the 15th century, whose name appears on chemical writings over a span of two centuries. The name antimony appears to be derived from the Latin *antimonium*, in a translation of a work by the alchemist Geber, but its real origin is uncertain.

Bismuth. Bismuth evidently was known in very early times, since it occurs in the native state as well as in compounds. For a long period, however, it was not clearly recognized as a separate metal, being confused with such metals as lead, antimony, and tin. Miners during the Middle Ages apparently believed bismuth to be a stage in the development of silver from baser metals and were dismayed when they uncovered a vein of the metal thinking they had interrupted the process. In the 15th-century writings of Basil Valentine this element is referred to as *wismut*. The name may have been derived from the German words *wis mat*, meaning white mass. In any case it was latinized to *bisemutum* by the mineralogist Georgius Agricola, who recognized its distinctive qualities and described how to obtain it from its ores. Bismuth was accepted as a specific metal by the middle of the 18th century, and works on its chemistry were published in 1739 by the German chemist Johann Heinrich Pott and in 1753 by the Frenchman Claude-François Geoffroy.

Comparative chemistry of the nitrogen group elements

ELECTRONIC CONFIGURATIONS

Similarities in orbital arrangement. In the periodic table, each of the nitrogen group elements occupies the fifth position among the main group elements of its period, a position designated Va. In terms of the electronic configuration of its atoms, each nitrogen group element possesses an outermost shell of five electrons. In each case, these occupy an outer *s* orbital completely (with two electrons) and contribute one electron to each of the three outer *p* orbitals (the orbitals being electron regions within the atom and the letter designations, *s*, *p*, *d*, and *f*, being used to designate different classes of orbital). The arrangement of outer electrons in the atoms of the nitrogen elements thus provides three half-filled outer orbitals that, by interaction with half-filled orbitals of the atoms of other elements, can form three covalent bonds. The other atoms may attract the shared electrons either more or less strongly than do the nitrogen group atoms; therefore the latter may acquire either positive or negative charges and exist in oxidation states of +3 or −3 in their compounds. In this respect, the nitrogen elements are alike.

Another similarity among the nitrogen elements is the existence of an unshared, or lone, pair of electrons, which remains after the three covalent bonds, or their equivalent, have been formed. This lone pair permits the molecule to act as an electron pair donor in the formation of molecular addition compounds and complexes. The availability of the lone pair depends upon various factors, such as the relative size of the atom, its partial charge in the molecule, the spatial characteristics of other groups in the molecule, and the as-yet poorly understood phenomenon called the "inert pair effect." This effect consists of a tendency for the paired *s* electrons in the outermost shell of the heavier atoms of a major group to remain chemically unreactive. Because of it, the electron pair-donating ability of the nitrogen group elements is not uniform throughout the group; it is probably greatest with nitrogen, less with the intermediate elements, and nonexistent with bismuth.

Lone pairs

Variations in bonding capacity. Significant differences in electronic configurations also occur among the elements of the nitrogen group with respect both to the underlying shell and to the outer *d* orbitals. Since the latter first appear with the third period of the table, they are present in all elements of the group but nitrogen. The possibility of utilizing these outer *d* orbitals for bonding thus exists for phosphorus, arsenic, antimony, and bismuth, but not for nitrogen.

There are three principal ways in which the outer *d* orbitals can be used to increase the number of bonds or expand the valence octet. One is by providing a space to which one of the *s* electrons can be promoted. This creates two additional half-filled orbitals (one *d* and one *s* orbit-

Utilization of *d* orbitals

al), and it therefore generates the capacity to form two additional covalent bonds. This is exemplified by the production of phosphorus pentafluoride, PF_5, by further fluorination of the trifluoride, PF_3. Such promotion appears to be greatly assisted by the increase in outer *d*-orbital stability that results from the withdrawal of part of the screening electron and the attendant increase of the effective nuclear charge of the central atom. In PF_5, for example, the fluorine atoms, being much more electronegative than the phosphorus atom, draw away a portion of the phosphorus electrons, leaving the outer *d* orbitals more exposed to the phosphorus nucleus and therefore more stable.

A second way in which the outer *d* orbitals can become involved in the bonding is by their becoming sufficiently stable to attract a lone pair of electrons from a donor. For example, PF_5 can serve as an electron pair acceptor through an outer *d* orbital to coordinate a fluoride ion donor and form the complex ion PF_6^-.

A third way of involving *d* orbitals in bonding is for them to become partially occupied in accommodating lone-pair electrons from another atom, which is already attached by a single bond, thereby strengthening the bond. The phosphorus oxyhalides, of general formula POX_3, appear to be examples of this; their phosphorus–oxygen bonds are observed to be shorter and stronger than expected for ordinary single bonds.

The +5 oxidation state. It thus is possible for an atom of phosphorus, arsenic, antimony, or bismuth to expand its valence octet to form five covalent bonds and one additional coordinate covalent bond. This is not possible for nitrogen, which exhibits a maximum coordination number of four: three single covalent bonds and a coordinate covalent bond with nitrogen acting as donor (through its lone pair). Nevertheless, the +5 oxidation state is formally applicable to nitrogen, so that all five elements can be found in this state. When compounds in the +5 oxidation state are studied, however, it is observed that their properties do not exhibit a uniform trend within the group. Rather, a certain degree of alternation is observed, the +5 states of nitrogen, arsenic, and bismuth appearing less stable and more strongly oxidizing than the corresponding states of phosphorus and antimony. In part this alternation may find explanation in the electronic differences among the atoms with respect to their underlying shells. The number of electrons in the shell just below the outermost level, is two for nitrogen, eight for phosphorus, and 18 for arsenic, antimony, and bismuth.

Alternation of properties

Increasing the nuclear charge by 18 from phosphorus to arsenic may be accompanied by incomplete shielding of this extra charge by the ten 3*d* electrons also added. This would imply smaller size and a greater electronegativity for arsenic than for phosphorus and thus a greater similarity between the phosphorus and antimony atoms. This subject, however, is still controversial, and the widely used scale of electronegativities devised by Linus Pauling fails to make this distinction.

An interesting anomaly is presented by the fact that nitrogen as a free element is in the form of gaseous diatomic molecules, while the elements immediately preceding it in its period of the table are solids, as are the other elements in its group. In surveying the elements of the second period, the most obvious difference in atomic structure found on reaching nitrogen is the appearance for the first time in compounds of the element of a lone pair of electrons not used in bonding with other atoms. Calculations suggest that the presence of this lone pair of electrons is associated with a considerable weakening of nitrogen to nitrogen single bonds in compounds where these bonds occur. In the diatomic nitrogen molecule, however, the bonding is of a different variety—triple bonds being found between the atoms. It is thought that the triple bond is unaffected (unweakened) by the lone pairs of electrons on the nitrogen atoms, and this is assumed to be the reason why nitrogen "prefers" to exist as triply bonded gaseous diatomic molecules rather than as a condensed singly bonded solid polymer.

Bond weakening by lone pairs

The same effect might be expected to be operable with the other elements of the nitrogen group, all of which also contain lone electron pairs in their outermost shells. Further calculations disclose, however, that the bond-weakening effect of the lone pair is far less pronounced with these elements than it is with nitrogen. As a result, with these elements, single bonds are favoured over multiple bonds, and the diatomic state of the molecules is not the preferred form.

Relative electronegativities. It might also be expected that the weakening effect of the lone pair would be observed in compounds of the nitrogen group elements. The picture is more complicated here because the bonds under discussion are formed between different types of atoms. Since different elements differ in electronegativity, bonds between the atoms of different elements are inevitably polar. For purposes of discussion it can be assumed that polar bonds consist of blends of nonpolar covalent bonds and completely polar, ionic bonds. It can then be shown that a relatively small amount of ionic character will contribute a disproportionate share to the overall bond strength. Since the weakening effect of the lone pair is felt only on the covalent portion of the polar bond, rather than on the ionic portion, the less polar bonds will exhibit the greater lone-pair weakening effects.

Comparison of compounds of nitrogen group. These considerations become important in comparing the chemical behaviour of the nitrogen group elements. The electronegativity of nitrogen itself, although lower than that of oxygen, is substantially higher than that of any of the other elements of this group. Bonds between nitrogen and oxygen, therefore, will be considerably less polar than those between oxygen and phosphorus, or oxygen and arsenic, antimony, or bismuth. Consequently, for this reason alone, the covalent contribution to the nitrogen–oxygen bond energy will be relatively more important than is the case with the bonds between oxygen and the heavier elements of the group. Thus, single-bond weakening by the lone pair—and a corresponding tendency toward bond multiplicity—is likely to be much greater with oxides of nitrogen than with oxides of the heavier nitrogen group elements.

FORMATION OF CHEMICAL COMPOUNDS

Oxides. The nitrogen oxides are all very volatile compounds, and all but one are gaseous and show no polymerization; those of the heavier elements, on the other hand, are all solids and demonstrate considerable polymerization. Only two of the six oxides of nitrogen have molecular formulas formally similar to those of the oxides of the other elements in the group. Specifically, the nitrogen oxides are formulated as follows: N_2O, NO, N_2O_3, NO_2, N_2O_4, and N_2O_5, all but the last being gaseous at ordinary temperatures and the last very volatile. The principal oxides of the other elements, however, are relatively nonvolatile solids having the formulas P_4O_6, P_4O_{10}, As_4O_6, As_2O_5, Sb_2O_3, Sb_2O_5, Bi_2O_3, Bi_2O_5. The most important contributing factor to the explanation of these differences is the fact that multiplicity is energetically favoured over single bonds between nitrogen and oxygen, while single bonds are favoured over multiple bonds between oxygen and the other elements of the group.

Multiple and single bonding

Acids. Similar differences, attributable to the same basic cause, are observed among the oxygen acids derived from these oxides. For example, two of the nitrogen oxides are anhydrides of acids: dinitrogen trioxide, N_2O_3, which is the anhydride of the unstable nitrous acid, HNO_2, and dinitrogen pentoxide, N_2O_5, which is the anhydride of the strong acid, nitric acid, HNO_3. The analogous phosphorus acid, HPO_3, is a highly polymeric solid, and the fully hydrated form, phosphoric acid, H_3PO_4, has no parallel in nitrogen chemistry. In part, the existence of the multiply bonded HNO_3, rather than H_3NO_4, with its single bonds, represents the great advantage of multiplicity in nitrogen to oxygen bonding. Such multiplicity of bonding is less advantageous in phophorus–oxygen bonding, as evidenced by the fact that HPO_3 exists as the singly bonded polymer.

Because the ionic species in solution are not necessarily

analogous, it is difficult to make a meaningful comparison among acid strengths in this group. The strongest of the acids is, of course, nitric acid. Phosphoric and arsenic acids are roughly comparable to one another in strength, but the latter is a fairly strong oxidizing agent, whereas the former is relatively nonoxidizing. In the +3 state, phosphorous acid is a moderately strong acid, but arsenious acid is very weak and shows some properties similar to those of bases. Antimonous acid also has both acidic and basic properties, but it is somewhat weaker as an acid and stronger as a base. Finally, bismuth(V) oxide is very unstable; it is a strong oxidizing agent but of uncertain acidity. Bismuth(III) oxide is almost exclusively basic. The trend, therefore, although it is confused in interpretation by dissimilarities from element to element, is from strongly acidic to weakly basic oxides as one proceeds down the group.

Hydrides. Somewhat greater similarities are shown by the hydrogen compounds of the nitrogen group elements, ammonia, NH_3; phosphine, PH_3; arsine, AsH_3; stibine, SbH_3; and bismuthine, BiH_3. All are gaseous and, with the exception of ammonia, extremely toxic. All can be formed by hydrolysis of a binary compound with a metal, and all can be oxidized easily. The smaller size and higher electronegativity of nitrogen make ammonia atypical, in that its bonds are more polar and more stable; and, unlike the other hydrides of the group, ammonia tends to associate through protonic bridging. A lone pair of electrons resides on the central atom of each of these hydrogen compounds, but it is much more available for electron-pair donation on the more negatively charged nitrogen atom of ammonia. Ammonium salts, for example, are much more stable and better known than phosphonium salts, which exist, but dissociate readily.

Halides. Numerous halides of the nitrogen group elements are known. These include all the possible trihalides except NBr_3 (NI_3 occurs as $NI_3 \cdot NH_3$), and some pentahalides. The known pentahalides are PF_5, PCl_5, PBr_5, AsF_5, SbF_5, $SbCl_5$, and BiF_5. The possibility of preparing $AsCl_5$ has been explored thoroughly, without success. This failure is probably related to a higher electronegativity for arsenic than for either phosphorus or antimony, although this is controversial. The failure is, however, consistent with the relative instability of the higher oxidation states of arsenic, selenium, and bromine, a fact that must somehow be associated with these substances being the first of the 18-shell elements in their respective groups. In general, the Va group electronegativities are too high to permit highly polar bonds to halogen, and, as a result, the halogen remains potentially a halogenating agent in these compounds. They are all molecular substances of relatively high volatility, with little opportunity for association, except as $PCl_4^+PCl_6^-$, in the case of PCl_5, and as $PBr_4^+Br^-$, in the case of PBr_5. Since the hydroxides of nitrogen and phosphorus are exclusively acidic, their halides are, as expected, susceptible to complete hydrolysis; the other halides of the group also hydrolyze extensively. Nitrogen halides are again exceptional in that they are extremely unstable, tending to explode violently (an exception to this generalization is the fairly stable fluoride, which possesses moderately polar bonds between atoms of a size favourable for strong bonding). This explosive tendency is aided by the low polarity, and therefore low bond energy, of bonds between nitrogen and the halogens, plus the large amount of energy released when two nitrogen atoms join to form the very stable N_2 molecule.

Instability and electronegative differences

Sulfides. The nitrogen elements form a variety of compounds with sulfur. Although these compounds are not always comparable, there is a general similarity between the nitrogen and arsenic compounds and thus an alternation in properties observed upon proceeding down the group. This is shown by the existence of the compounds S_4N_4 and As_4S_4, while no analogous compounds of phosphorus or antimony have been prepared. Major sulfides known are P_4S_7, P_4S_{10}, P_4S_3, As_2S_5, As_4S_6, Sb_2S_5, Sb_2S_3, and Bi_2S_3.

Organic derivatives. Organic derivatives of all the nitrogen elements exist. Some of these also show alternations in properties. Trimethylamine, trimethylphosphine, trimethylarsine, trimethylstibine, and trimethylbismuthine are all volatile compounds; and all but the bismuth compound show an ability to utilize the lone pair on the central atom as a donor to a positive alkyl group, forming salts of the general formula, $R_4E^+X^-$, in which R is an alkyl group, E the group Va element, and X a halogen atom. The alternation in properties of the trimethyl compounds is most conspicuous with respect to oxidizability, for the phosphorus, antimony, and bismuth compounds are spontaneously inflammable in air, whereas those of nitrogen and arsenic are not. Another evidence of alternation in properties is given by the existence of stable tetramethylhydrazine $(CH_3)_2N–N(CH_3)_2$ and cacodyl $(CH_3)_2As–As(CH_3)_2$, whereas the phosphorus analogue is unknown and the antimony analogue very unstable. Surprisingly, pentamethylarsenic and pentamethylantimony have been made, as have pentaphenyl compounds of phosphorus, arsenic, antimony, and bismuth.

That each element is a chemical individual, regardless of how it might be placed within the periodic table, is well substantiated by a close examination of the individuals of the nitrogen group. Further evidences of the uniqueness of the individual elements will be recognized in the more detailed discussion of their chemistry in the following section. General atomic and physical properties are presented in the table.

Individual nitrogen group elements and their compounds

NITROGEN AND ITS COMPOUNDS

Occurrence and distribution. The atmosphere consists of 78.03 weight percent of nitrogen; this is the principal source of nitrogen for commerce and industry. The atmosphere also contains varying small amounts of ammonia and ammonium salts, as well as nitrogen oxides and nitric acid (the latter substances being formed in electrical storms and in the internal combustion engine). Nitrogen occurs also in mineral deposits of nitre or saltpetre (potassium nitrate, KNO_3) and Chile saltpetre (sodium nitrate, $NaNO_3$), but in quantities far inadequate for man's needs. Another material rich in nitrogen is guano, found in bat caves and in dry places frequented by birds. Nitrogen constitutes on the average about 16 percent by weight of the complex organic compounds known as proteins, present in all living organisms. The natural abundance of nitrogen in the earth's crust is 0.3 parts per 1,000. The cosmic abundance—the estimated total abundance in the universe—is between three and seven atoms per atom of silicon, which is taken as the standard.

Commercial production and uses. *Nitrogen.* Commercial production of nitrogen is largely by fractional distillation of liquefied air. The boiling temperature of nitrogen is −195.8° C, about 13° below that of oxygen, which is therefore left behind. Nitrogen can also be produced on a large scale by burning carbon or hydrocarbons in air and separating the resulting carbon dioxide and water from the residual nitrogen. Various laboratory reactions that yield nitrogen include heating ammonium nitrite (NH_4NO_2) solutions, oxidation of ammonia by bromine water, and oxidation of ammonia by hot cupric oxide.

$$NH_4^+ + NO_2^- \rightarrow N_2 + 2\ H_2O$$
$$8\ NH_3 + 3\ Br_2 \rightarrow N_2 + 6\ NH_4^+ + 6\ Br^-$$
$$2\ NH_3 + 3\ CuO \rightarrow N_2 + 3\ H_2O + 3\ Cu$$

Uses of molecular nitrogen

Elemental nitrogen can be used as an inert atmosphere for reactions requiring the exclusion of oxygen. In the liquid state, nitrogen has valuable cryogenic applications; except for the gases hydrogen, methane, carbon monoxide, fluorine, and oxygen, practically all chemical substances have negligible vapour tensions at the boiling point of nitrogen and exist, therefore, as crystalline solids at that temperature.

In the chemical industry, nitrogen is used as a preventive of oxidation or other deterioration of a product, as an inert diluent of a reactive gas, as a carrier to remove heat or chemicals and as an inhibitor of fire or explosions. In the food industry nitrogen gas is employed to prevent

spoilage through oxidation, mold, or insects, and liquid nitrogen is used for freeze drying and for refrigeration systems. In the electrical industry nitrogen is used to prevent oxidation and other chemical reactions, to pressurize cable jackets, and to shield motors. Nitrogen finds application in the metals industry in welding, soldering, and brazing, where it helps prevent oxidation, carburization, and decarburization. As a nonreactive gas, nitrogen is employed to make foamed—or expanded—rubber, plastics, and elastomers, to serve as a propellant gas for aerosol cans, and to pressurize liquid propellants for reaction jets. In medicine rapid freezing with liquid nitrogen may be used to preserve blood, bone marrow, tissue, bacteria, and semen.

Nitrogen fixation

Although the other applications are important, by far the greatest bulk of elemental nitrogen is consumed in the manufacture of nitrogen compounds. The bond between atoms in the nitrogen molecules is so strong (226 kilocalories per mole; more than twice that of molecular hydrogen) that it is difficult to cause molecular nitrogen to enter into other combinations. Most living organisms cannot utilize nitrogen directly and must have access to its compounds. Therefore the fixation of nitrogen, the incorporation of elemental nitrogen into compounds, is vitally important. In nature, two principal processes of nitrogen fixation are known. One is the action of electrical energy on the atmosphere, which dissociates nitrogen and oxygen molecules, allowing the free atoms to form nitric oxide, NO, and nitrogen dioxide, NO_2. Nitrogen dioxide then reacts with water as follows:

$$3\ NO_2 + H_2O \rightarrow 2\ HNO_3 + NO.$$

The nitric acid, HNO_3, dissolves and comes to earth with rain as a very dilute solution. In time it becomes part of the combined nitrogen of the soil. The other principal process of natural nitrogen fixation is that of certain plants and vegetables called legumes. Through a cooperative action with bacteria, legumes are able to convert atmospheric nitrogen directly into nitrogen compounds. Certain bacteria alone, such as *Azotobacter chroococcum* and *Clostridium pasteurianum,* are also capable of fixing nitrogen.

The Haber process

Ammonia. The chief commercial method of fixing nitrogen is the Haber process for synthesizing ammonia. This process was developed during World War I to lessen the dependence of Germany on Chilean nitrate. It involves the direct synthesis of ammonia from its elements. The synthesis is favoured thermodynamically (the standard free energy of formation of ammonia at 25° C being –3.94 kilocalories per mole). The equation of the reaction however,

$$N_2 + 3\ H_2 \rightarrow 2\ NH_3,$$

shows that four moles of gases (one of nitrogen and three of hydrogen) are used to produce two moles of ammonia. As a result there is a substantial decrease in entropy, or unavailable energy in the system during the reaction, and poorer yields of ammonia would be expected at higher temperatures. The ideal of carrying out the process at low temperature has not yet been realized, however, because of the high energy of activation required to cause the very stable nitrogen and hydrogen molecules to react. The Haber process succeeds in satisfying these conflicting requirements by means of a catalyst. The catalyst, based on iron, enables reaction to occur at a useful rate at temperatures around 500°C if the pressure is about 1,000 atmospheres.

Uses of ammonia

In the early 1970s, tens of millions of tons of ammonia were synthesized by the Haber process each year. Large amounts of this ammonia are used directly in agriculture, being applied directly to the soil from tanks containing the liquefied gas. Further amounts are converted to ammonium nitrate, ammonium phosphates, and other salts that also are used principally in commercial fertilizers. Ammonia from the Haber process is supplemented by ammonia obtained as a by-product of coke ovens, about six pounds being released per ton of coal.

In the textile industry, ammonia is used in the production of synthetic fibres, such as nylon and rayon. It is also employed in the dyeing and scouring of cotton, wool, rayon, and silk. Ammonia serves as a catalyst in the production of Bakelite and certain other synthetic resins. It is used in the synthesis of sulfa and many other drugs and vitamins. It neutralizes acidic by-products of petroleum refining, and, in the rubber industry, it prevents a coagulation of raw latex during transportation from plantation to factory. Because of its high heat of vaporization, low density, high stability, and low corrosiveness, ammonia is a valuable refrigerant for air conditioning, ice making, and cold storage.

In the paper industry, ammonia can be used in place of calcium salts in the bisulfite process for making wood pulp, and it is also used to dissolve casein for coating paper. Large quantities of ammonia find application in a widely used method for producing soda ash (Na_2CO_3), called, after its inventor, the Solvay process. Ammonia is used in various metallurgical processes, including the nitriding of alloy sheets to harden their surfaces. Since ammonia can be decomposed easily to give hydrogen, it is a convenient portable source of atomic hydrogen for welding. Together with chlorine, ammonia is used to purify water supplies, the active intermediates being chloramines. Finally, among its minor uses is inclusion in certain household cleansing agents.

Nitric acid. Perhaps the largest part of commercially synthesized ammonia is converted to nitric acid and nitrates. The process depends on the catalyzed oxidation of ammonia to nitric oxide:

$$5\ O_2 + 4\ NH_3 \rightarrow 4\ NO + 6\ H_2O.$$

At lower temperatures, the nitric oxide reacts readily with atmospheric oxygen to form nitrogen dioxide, NO_2. As indicated above, this substance reacts with water to form nitric acid and nitric oxide. The latter is reoxidized and recycled. It is possible also to distill nitric acid from a mixture of a nitrate and concentrated sulfuric acid.

Uses of nitric acid

Nitric acid is used in very large quantities to make explosives, such as TNT (trinitrotoluene), nitroglycerine, and gunpowder. It is also used to carry out other nitrations, which produce valuable synthetic organic chemicals. Commonly it is mixed with concentrated sulfuric acid for this purpose; the principal active component is NO_2^+, and the by-product water is effectively absorbed by the sulfuric acid. Nitrocellulose, prepared by nitration of cellulose, forms the base for pyroxylin lacquers and plastics, as well as certain synthetic fibres.

Other compounds. Ammonia and nitric acid are the principal nitrogen compounds of commerce, from which most other nitrogen compounds are derived. Also of some importance are certain nitrides, solids formed by direct combination of metals with nitrogen, usually at elevated temperatures. Some of these are extremely hard and refractory, as well as chemically unreactive. The nitrides of boron, titanium, zirconium, and tantalum have useful special applications. For example, a diamond-structured modification of boron nitride (borazon) is nearly as hard as diamond and less easily oxidized, thus having advantages for abrasive, grinding, and polishing uses.

Another nitrogen compound of industrial importance is hydrazine, N_2H_4. Hydrazine is a colourless liquid made by careful oxidation of ammonia with sodium hypochlorite. Hydrazine is used chiefly as a rocket fuel, but is also applied in pharmaceuticals, manufacturing, agriculture, polymer preparation, petroleum refining, chemical processing, detergent synthesis, and the metals industry.

Cyanides

Hydrogen cyanide, HCN, is a gaseous compound that condenses easily to a colourless liquid. It can be prepared by the action of strong acids on its salts, but it is usually produced by a catalytic high-temperature reaction of methane, ammonia, and oxygen. It is employed industrially chiefly to form acrylonitrile by reaction with acetylene. (Acrylonitrile is a starting material for the production of synthetic fibres.) Sodium cyanide can be made by heating sodium metal with ammonia to form sodium amide (or sodamide, $NaNH_2$), which is then heated with carbon:

$$2\ Na + 2\ NH_3 \rightarrow 2\ NaNH_2 + H_2$$
$$NaNH_2 + C \rightarrow NaCN + H_2.$$

Sodium cyanide has uses in electroplating and various chemical processes.

When lime, CaO, is heated with coke, calcium carbide, CaC_2, is formed. Heating this substance and nitrogen together in an electric furnace at a high temperature causes release of carbon and formation of calcium cyanamide, $CaCN_2$. This process is a large-scale industrial method for fixing nitrogen. The crude product can be used directly in fertilizers, since slow hydrolysis liberates ammonia in the soil. It also can be the basis of synthesis of other nitrogen compounds.

Properties and reaction. *Nitrogen.* Nitrogen is a colourless, odourless gas, which condenses at −195.8° C to a colourless, mobile liquid. The element exists as N_2 molecules, represented as :N:::N:, for which the bond energy of 226 kilocalories per mole is exceeded only by that of carbon monoxide, 256 kilocalories per mole. Because of this high bond energy the activation energy for reaction of molecular nitrogen is usually very high, causing nitrogen to be relatively inert to most reagents under ordinary conditions. Furthermore, the high stability of the nitrogen molecule contributes significantly to the thermodynamic instability of many nitrogen compounds, in which the bonds, although reasonably strong, are far less so than those in molecular nitrogen. For these reasons, elemental nitrogen appears to conceal quite effectively the truly reactive nature of its individual atoms.

A relatively recent and unexpected discovery is that nitrogen molecules are able to serve as ligands in complex coordination compounds. The observation that certain solutions of ruthenium complexes can absorb atmospheric nitrogen has led to hope that one day a simpler and better method of nitrogen fixation may be found.

Atomic nitrogen

An active form of nitrogen, presumably containing free nitrogen atoms, can be created by passage of nitrogen gas at low pressure through a high-tension electrical discharge. The product glows with a yellow light and is much more reactive than ordinary molecular nitrogen, combining with atomic hydrogen and with sulfur, phosphorus, and various metals, and capable of decomposing nitric oxide, NO, to N_2 and O_2.

A nitrogen atom has the electronic structure represented by $1s^2$, $2s^2$, $2p^3$, or alternatively, 2–5. The five outer shell electrons screen the nuclear charge quite poorly, with the result that the effective nuclear charge felt at the covalent radius distance is relatively high. Thus nitrogen atoms are relatively small in size and high in electronegativity, being intermediate between carbon and oxygen in both of these properties. The electronic configuration includes three half-filled outer orbitals, which give the atom the capacity to form three covalent bonds. The nitrogen atom should therefore be a very reactive species, combining with most other elements to form stable binary compounds, especially when the other element is sufficiently different in electronegativity to impart substantial polarity to the bonds. When the other element is lower in electronegativity than nitrogen, the polarity gives partial negative charge to the nitrogen atom, making its lone pair electrons available for coordination. When the other element is more electronegative, however, the resulting partial positive charge on nitrogen greatly limits the donor properties of the molecule. When the bond polarity is low (due to the electronegativity of the other element being similar to that of nitrogen), multiple bonding is greatly favoured over single bonding. If disparity of atomic size prevents such multiple bonding, then the single bond that forms is likely to be relatively weak, and the compound is likely to be unstable with respect to the free elements. All of these bonding characteristics of nitrogen are observable in its general chemistry, but with kinetic restrictions imposed by the high stability of the N_2 molecule.

Nitrides. The simplest nitride is that of hydrogen, H_3N, which is customarily called ammonia (formulated

Ammonia

NH_3) and not usually thought of as a nitride at all. The synthesis of ammonia by the Haber process represents a successful technological overcoming of the problems caused by the stability of molecular nitrogen. The ammonia molecule has a pyramidal structure, as would be predicted from a knowledge of the presence of the lone-pair electrons on nitrogen, which force the three nitrogen–hydrogen bonds out of planarity to create bond angles of about 107°. Because hydrogen has a lower electronegativity than nitrogen, the N–H bonds are somewhat polar, with calculated partial charges of 0.06 on hydrogen and −0.17 on nitrogen. The combination of polarity of bonds with geometrical dissymmetry gives the ammonia molecule an appreciable dipole moment 1.49 Debye units. The partial positive charge on the hydrogen atoms, together with the lone pair of electrons on the relatively small, negatively charged nitrogen atom, permits the formation of protonic bridges (hydrogen bonds). Although ammonia is unassociated in the gaseous state, it is much more readily condensed than any of its congeners within the group, none of which appear to associate through protonic bridging. Ammonia can be liquefied under a pressure of about ten atmospheres at ordinary temperatures, and the boiling point of the liquid at one atmosphere pressure is −33.4° C.

Liquid ammonia as solvent

As a liquid, ammonia is an interesting and useful solvent, analogous to water but uniquely different. One of the differences is that ammonia is fundamentally a weaker oxidizing agent than water; another is that it is more basic. A third is that the extent of auto-ionization of liquid ammonia is much less than that of water. A very interesting consequence of these differences is that solvated electrons can exist for a considerable length of time in liquid ammonia, whereas they reduce water and lose their identity in that solvent very quickly. Thus, if sodium metal comes in contact with water, the evolution of hydrogen is immediate, leaving behind sodium ions and hydroxide ions. In pure liquid ammonia, however, sodium dissolves as sodium ions and "free" solvated electrons. Other reactive metals behave similarly, all giving highly coloured solutions that demonstrate conduction of electricity comparable to that of metals. Such solutions are blue when diluted and bronze when highly concentrated. Both colour and electrical conductivity are ascribed to the solvated electrons, which are highly mobile but which are believed to lie in cavities surrounded by the hydrogen portions of four ammonia molecules. Since the action of a reducing agent is to supply electrons, solutions of metals in liquid ammonia—with their large solvated-electron content—are extremely effective in that regard.

The relative solubilities of salts in liquid ammonia are different from the corresponding solubilities in water. Liquid ammonia can be used as the solvent medium for a variety of useful inorganic and organic syntheses and reactions. Acid-base reactions in liquid ammonia are related to the autoionization product species:

$$2\ NH_3 \rightarrow NH_4^+ + NH_2^-;\ \text{ionization constant, } K \sim 10^{-33}.$$

In this ionization, ammonium ion, NH_4^+, is analogous to the hydronium ion, H_3O^+ in water, and amide ion, NH_2^-, is analogous to the hydroxide ion, OH^-. According to the solvent theory of acids and bases, any substance, such as an ammonium salt, that increases the concentration of the anion associated with the solvent (in this case, the amide ion) is thereby basic. A typical neutralization reaction in liquid ammonia thus would be:

$$NH_4Br + KNH_2 \rightarrow KBr + 2\ NH_3.$$

Largely because of the formation of protonic bridges, ammonia is very soluble in water; a small amount of ionization occurs as follows:

$$NH_3 + H_2O \rightarrow NH_4^+ + OH^-.$$

Aqueous ammonia, often called ammonium hydroxide although no molecular species NH_4OH has ever been isolated or identified, is therefore a weak base ($K = 1.8 \times 10^{-5}$), comparable in strength to acetic acid as an acid. By virtue of the lone pair of electrons and the partial negative charge on nitrogen, ammonia is also a good Lewis base (electron-pair donor), donating its electron pair—both in and out of water solution—to thousands of Lewis acids (electron-pair acceptors) with the formation of complex compounds called ammines and ammino-complexes. The simplest Lewis acid is the proton, or H^+,

to which ammonia readily coordinates to form the moderately stable ammonium ion, NH_4^+. Ammonium salts, of course, are well-known. They are readily formed by the neutralization of acids by ammonia, both in aqueous solution and in the gaseous phase. These salts appear to sublime when heated, but evidently they are dissociated completely in the gas phase; cooling reforms the ammonium salt as a white smoke, ammonium chloride being a familiar contaminant of any laboratory, wherein hydrochloric acid and aqueous ammonia are commonly used reagents.

Although basicity resulting from the lone-pair electrons is the most familiar property of ammonia, the compound also can exhibit acidity as the result of its partially positive hydrogen atoms. Thus metal–ammonia solutions slowly evolve hydrogen, and the same metals react with ammonia gas on heating to displace part of the hydrogen, forming amides and imides.

Other nitrides

Molecular nitrogen, despite its relative inertness, reacts directly with certain metals at ordinary temperatures, and even more rapidly when heated. Such metals include lithium (but none of the other alkali metals), magnesium, calcium, strontium, barium, and several others. Beryllium unites with nitrogen above 900° C, but the product is unstable above its melting point of about 220° C in the absence of nitrogen. Boron and aluminum form stable polymeric nitrides when heated with nitrogen gas. Other nitrides can be prepared by heating the metal with ammonia or by other indirect methods.

Nitrides of the major group elements have the expected composition and are fairly typical binary compounds. Nitrides of the transitional elements, on the other hand, tend to have indefinite compositions, being of an "interstitial" or "alloy" nature. These transitional metal nitrides are commonly metallic in appearance, highly conducting, hard, high melting, and resistant to chemical reaction. In contrast, the major group nitrides hydrolyze readily to form ammonia and the corresponding hydroxides. The question of whether a metal nitride actually contains a nitride ion, N^{3-}, probably must be answered in the negative, for it is extremely doubtful whether the often asserted existence of anions of higher than unit charge in a crystalline solid is reasonable. Calculations of partial charge according to the principle of electronegativity equalization show that nitrogen retains only a small fraction of the postulated −3 charge in such compounds. Even for Cs_3N the charge on nitrogen should be only about −0.9.

Lithium nitride, Li_3N, is a red solid, in which each nitrogen is equidistant from two lithium atoms, situated at distances of 1.94 Å from the nitrogen, and six more, situated at 2.11 Å. Because oxygen is more electronegative than nitrogen, it can displace nitrogen from its binary compounds; and lithium nitride, therefore, is oxidized readily on warming in air. Sodium nitride, Na_3N, can be made only by indirect methods, since it is much less stable and more reactive than Li_3N. Presumably the crystal energy of the nitrides of these larger atoms (including also the heavier alkali metals) is insufficient to compensate for stability to be gained by conversion to molecular nitrogen.

Boron nitride

Of especial interest is boron nitride, BN, which forms normally as a graphite-like layer structure when boron is heated with nitrogen. This substance is extremely unreactive, being almost inert to air and alkalies; it is hydrolyzed only slowly, but it can be decomposed by acids. The similarity of boron nitride to graphite is an interesting example of how two elements horizontally separated in the periodic table by one element can form a compound whose properties resemble that intermediate element. The resemblance of boron nitride to the graphite form of carbon led to successful attempts to convert graphitic boron nitride to a diamond structure, following the technology of converting real graphite to diamond. This diamond form, called borazon, is second only to diamond itself in hardness. Borazon has greater resistance than diamond to oxidation at the high temperatures caused by friction, and this gives it significant advantages over diamond as an industrial abrasive and polishing agent.

Oxides and acids. The direct combination of nitrogen with oxygen occurs only under the influence of electrical energy. Several oxides are known, of which the most familiar are nitrous oxide, or nitrogen(I) oxide, N_2O; nitric oxide, or nitrogen(II) oxide, NO; nitrogen sesquioxide, or nitrogen(III) oxide, N_2O_3; nitrogen dioxide, or nitrogen(IV) oxide, NO_2; and dinitrogen pentoxide or nitrogen(V) oxide, N_2O_5. The last is a volatile solid; the others are gases under ordinary conditions. The electronegativity of nitrogen is sufficiently close to that of oxygen to prevent oxygen from becoming very negative in these compounds. Thus they are all at least potentially acidic and oxidizing, and although their multiple bonds are quite strong, they tend to be thermodynamically unstable with respect to the free elements.

Cautious heating of ammonium nitrate, NH_4NO_3, produces smooth decomposition:

$$NH_4NO_3 \rightarrow N_2O + 2\,H_2O.$$

Incautious heating may be disastrously explosive. In 1947 a freighter load of ammonium nitrate inexplicably detonated in Texas City, Texas.

Laughing gas

Nitrous oxide, also known as "laughing gas" because it induces mild hysteria when used as an anesthetic, is one of the few gases other than oxygen capable of supporting combustion. This is not surprising, since the structure favours loss of oxygen and the formation of molecular nitrogen; the two nitrogen atoms are already joined together and the oxygen atom is attached to one of them at the end of the linear molecule. The principal characteristic of nitrous oxide is its oxidizing action.

Nitric oxide is the principal product of the catalyzed oxidation of ammonia, nitrous oxide, being unstable at the high temperature of the reaction. Nitric oxide is commonly formed from nitric acid when the latter is used as an oxidant. In addition, nitric oxide is produced in the internal-combustion engine by oxidation of atmospheric nitrogen. After it is emitted in the exhaust from the engine, nitric oxide reacts with oxygen of the air to form the dioxide, NO_2, which in turn reacts with organic air impurities to form highly irritating components of smog.

Nitrogen sesquioxide, N_2O_3, is the least stable of the familiar nitrogen oxides; at ordinary temperatures it is about 90 percent dissociated into NO and NO_2 molecules. An equimolar mixture of these gases, in fact, acts as though it were N_2O_3, which is the anhydride of nitrous acid, HNO_2. For example, an equimolar mixture of NO and NO_2 is absorbed by sodium hydroxide solution, forming a solution of sodium nitrite, $NaNO_2$.

Nitrous acid

Nitrous acid is a weak and unstable acid, known only in solution. Its salts, however, usually are much more stable, those of the alkali metals being resistant toward decomposition at very high temperatures. Ammonium nitrite is an exception, being quite unstable, and its solutions serve as a laboratory source of nitrogen:

$$NH_4NO_2 \rightarrow N_2 + 2\,H_2O.$$

Nitrogen dioxide, NO_2, is not itself an anhydride of a stable nitrogen acid but it reacts with water in the following way:

$$3\,NO_2 + H_2O \rightarrow 2\,HNO_3 + NO.$$

This reaction is used for the commercial production of nitric acid, the nitric oxide formed being oxidized to nitrogen dioxide and recycled. Nitrogen dioxide is a reddish-brown gas, which boils at 22.4° C; it is quite familiar as a reduction product of concentrated nitric acid. The gas is conveniently prepared by thermal decomposition of lead nitrate. Like all molecules having an odd number of electrons, NO_2 is paramagnetic (as is NO), but it loses both its paramagnetism and its colour when cooled, for it dimerizes to dinitrogen tetroxide, N_2O_4. The latter is a colourless liquid, which freezes to a colourless solid at −10.2° C. As would be expected, NO_2 is a powerful oxidant. Dissolved in concentrated nitric acid, it forms a reddish-brown acid called fuming nitric acid, which is also a powerful oxidizing agent, causing easily combustible materials like wood to burst into flame.

The pentoxide, N_2O_5, which in the crystalline form appears to exist as nitronium nitrate, $NO_2^+NO_3^-$, is the anhydride of nitric acid:

$$N_2O_5 + 2\,H_2O \rightarrow 2\,HNO_3.$$

The chief method of preparation of N_2O_5 is by dehydration of nitric acid with phosphoric anhydride. Nitrogen pentoxide is unstable; it loses oxygen easily and is a strong oxidizing agent.

Nitric acid

Nitric acid, HNO_3, which—with hydrochloric and sulfuric acid—is one of the most common laboratory reagent acids, is a colourless liquid boiling at 84° C. It is miscible with water in all proportions and is slowly decomposed by light even at 25° C according to the following equation:

$$4\,HNO_3 \rightarrow 4\,NO_2 + 2\,H_2O + O_2.$$

This decomposition occurs even in the usual 69 percent aqueous solution, commonly called concentrated nitric acid, so that the initially colourless material slowly becomes yellowish brown. Nitric acid is a powerful oxidizing agent and attacks all metals, with the exception of gold and some of the platinum group. It is a strong acid in aqueous solution, and in dilute form it is used as a source of hydrogen ions.

The principal use of nitric acid is as a nitrating agent for the manufacture of nitro compounds employed as explosives and for other purposes. For this use nitric acid is usually mixed with concentrated sulfuric acid, which, being a stronger acid, causes the nitric acid to act essentially as a base:

$$HNO_3 + 2\,H_2SO_4 \rightarrow NO_2^+ + H_3O^+ + 2\,HSO_4^-.$$

In this circumstance, the nitronium ion, NO_2^+, is evidently the active nitrating species, and the overall result is that the sulfuric acid absorbs the by-product water from the nitration reaction.

Nitrate salts

Practically all metal nitrates are soluble in water, and many readily form crystalline hydrates. The alkali metal nitrates have the greatest thermal stability but lose oxygen at high temperatures to form nitrites. Other metal nitrates decompose more easily, forming metal oxides and various oxides of nitrogen.

Halides. The only stable halide of nitrogen is nitrogen trifluoride, NF_3. It is a colourless gas, which boils at −129° C; it can be prepared by electrolysis of fused anhydrous ammonium bifluoride, NH_4HF_2. Explosive by-products appear to be dinitrogen difluoride, N_2F_2, and the tetrafluoride, N_2F_4. Fluorine is sufficiently more electronegative than nitrogen to give appreciable polarity to the N–F bond and leave the nitrogen atom in NF_3 with a partial positive charge. This renders the nitrogen lone-pair electrons almost completely unavailable for the kind of Lewis-base coordination typical of ammonia. The compound is almost insoluble in water and in bases. When mixed with hydrogen and ignited, nitrogen trifluoride reacts explosively to form hydrogen fluoride and liberate nitrogen.

Action of free chlorine or hypochlorite salts on ammonium chloride, or on aqueous ammonia, leads to stepwise substitution of chlorine atoms for hydrogen atoms, forming first chloramine, NH_2Cl; then dichloramine, $NHCl_2$; and finally nitrogen trichloride, NCl_3. The last compound separates as a dark yellow oil, which has a boiling point of 71° C but is so unstable that it is likely to explode violently below 71° C. When dissolved in an inert solvent and stored in the dark, it is more stable. It is hydrolyzed slowly by a reaction that is the reverse of its formation:

$$NCl_3 + 3\,H_2O \rightleftharpoons NH_3 + 3\,HOCl.$$

Attempts to prepare nitrogen bromide have been unsuccessful, free nitrogen being evolved too readily even at lower temperatures. Nitrogen triiodide, on the other hand, appears to form readily from the action of aqueous ammonia on iodine, but the product includes a molecule of ammonia, $NI_3 \cdot NH_3$. This copper-coloured solid is so unstable that it detonates even when touched.

Other compounds. As mentioned earlier, hydrogen cyanide, HCN, is formed, along with acetylene, when a mixture of nitrogen and methane is exposed to an electric arc. Because it is a very weak acid, hydrogen cyanide also is liberated readily by acidification of any of its salts. The molecule is linear, as would be predicted for a molecule in which the outer-shell electrons of the central atom, a carbon atom, are utilized in only two locations, one as a single bond to hydrogen and the other as a triple bond to nitrogen. The nitrogen in the molecule possesses a partial negative charge and a lone pair of electrons, the hydrogen a partial positive charge. These are the two requisites for protonic bridging; hydrogen cyanide is therefore highly associated, melting at −13.4° C and boiling at 25.6° C. Correspondingly, it has a very high dielectric constant, 194.4, more than twice that of water. Nevertheless, its properties as a solvent are disappointingly poor.

Cyanogen and sulfur nitride

Cyanogen, $(CN)_2$, is called a "pseudohalogen" because of its resemblance to the diatomic halogens. It can be prepared by oxidation of a cyanide, much as a halogen can be prepared by oxidation of a halide. Like several of the halogens, it is a gas at normal temperature and pressure; its melting point is −27.8° C, and its boiling point is −21.1° C. The compound is extremely stable toward heat, dissociating only above 1,000° C. It will burn in air, and water slowly hydrolyzes it, principally to ammonium oxalate. (Cyanogen, in fact, may be regarded as the dinitrile of oxalic acid.)

Sulfur tetranitride, S_4N_4, can be formed by the action of ammonia on sulfur:

$$10\,S + 4\,NH_3 \rightarrow S_4N_4 + 6\,H_2S.$$

It is a golden-yellow solid that melts at 178° C; it can sometimes be distilled, but it is apt to explode. Hydrolysis occurs slowly in water; more rapidly in alkaline solution, according to the following equation:

$$S_4N_4 + 6\,NaOH + 3\,H_2O \rightarrow Na_2S_2O_3 + 2\,Na_2SO_3 + 4\,NH_3.$$

Most of the thousands of known organic nitrogen compounds may be regarded as derived from ammonia, hydrogen cyanide, cyanogen, or nitrous or nitric acid.

Analytical chemistry. Often the percentage of nitrogen in gas mixtures can be determined by measuring the volume after all other components have been absorbed by chemical reagents. Decomposition of nitrates by sulfuric acid in the presence of mercury liberates nitric oxide, which can be measured as a gas. Nitrogen is released from organic compounds when they are burned over copper oxide, and the free nitrogen can be measured as a gas after other combustion products have been absorbed. The well-known Kjeldahl method for determining the nitrogen content of organic compounds involves digestion of the compound with concentrated sulfuric acid (optionally containing mercury, or its oxide, and various salts, depending on the nature of the nitrogen compound). In this way, the nitrogen present is converted to ammonium sulfate. Addition of an excess of sodium hydroxide releases free ammonia, which is collected in standard acid; the amount of residual acid, which has not reacted with ammonia, is then determined by titration.

The Kjeldahl method

Biological and physiological significance. As might be expected in view of the importance of the presence of nitrogen in living matter, most—if not all—organic nitrogen compounds are physiologically active. Of the inorganic compounds, ammonia is relatively harmless in very small concentrations; but in higher concentrations it strongly irritates the eyes and upper respiratory tract and, in fact, inhibits respiration. Nitrous oxide finds considerable use as an anesthetic, especially in dentistry; prolonged breathing of higher concentrations, however, can be fatal. The higher oxides of nitrogen are all extremely toxic. They are severe lung irritants that can cause pleural edema and subsequent death. These compounds are especially insidious because they may be formed in fires involving nitrogenous materials and inhaled with smoke without causing awareness of their presence until several hours after exposure. Hydrazine also is an extremely toxic substance, as are hydrogen cyanide (prussic acid) and cyanogen. Nitric acid, especially in concentrated form, inflicts severe burns, which are painful and slow healing. Mild contact produces a typical yellow-brown stain (resulting from the formation of xanthoproteic acids in the skin).

Nitrogen itself, being inert, is innocuous except when breathed under pressure, in which case it dissolves in the blood and other body fluids in higher than normal con-

The bends

centration. This in itself does no harm, but if the pressure is reduced too rapidly, the excess nitrogen evolves as bubbles of gas in various locations in the body. These can cause muscle and joint pain, fainting, partial paralysis, and even death. These symptoms are referred to as "the bends." Divers, and others forced to breathe air under pressure, must therefore be extremely careful that the pressure is reduced to normal very slowly following exposure. This enables the excess nitrogen to be released harmlessly through the lungs without forming bubbles. A better alternative is to substitute mixtures of oxygen and helium for air. Helium is much less soluble in body fluids, and the dangers are thus diminished.

Isotopes of nitrogen. Nitrogen exists as two stable isotopes, ^{14}N and ^{15}N. These can be separated by chemical exchange or by thermal diffusion. Artificial radioactive isotopes have masses of 12, 13, 16, and 17. The most stable has a half-life of only about ten minutes.

PHOSPHORUS AND ITS COMPOUNDS

Occurrence and distribution. Phosphorus is a very widely distributed element—12th most abundant in the earth's crust, to which it contributes about 0.10 weight percent. Its cosmic abundance is estimated to be about one atom per 100 atoms of silicon, the standard. Its high chemical reactivity assures that it does not occur in the free state. The principal combined forms in nature are the phosphate salts. Nearly 190 different minerals have been found to contain phosphorus, but, of these, the principal source of phosphorus is the apatite series in which calcium ions exist along with phosphate ions and variable amounts of fluoride, chloride, or hydroxide ions, according to the formula $[Ca_{10}(PO_4)_6(F, Cl, \text{or } OH)_2]$. Commonly such metal atoms as magnesium, manganese, strontium, and lead substitute for calcium in the mineral; and silicate, sulfate, vanadate, and similar anions substitute for phosphate ions. Very large sedimentary deposits of fluoroapatite are found in many parts of the earth. The phosphate of bone and tooth enamel is hydroxyapatite. (The principle of lessening tooth decay by fluoridation depends upon the conversion of hydroxyapatite to the harder, more decay resistant, fluorapatite.)

Estimates of the total phosphate rock in the earth's crust average about 50,000,000,000 tons, of which north Africa contains two-thirds, and the U.S.S.R. and the United States most of the remaining third. This estimate includes only ore that is sufficiently rich in phosphate for conversion to useful products by present methods. Vast quantities of material lower in phosphorus content also exist.

Acidulation and reduction

Commercial production and uses. Two principal techniques for converting phosphate rock to usable materials are practiced. One involves acidulation of the crushed rock—with either sulfuric or phosphoric acids—to form crude calcium hydrogen phosphates that, being water soluble, are valuable additions to fertilizer. The other method is the reduction of the phosphate with carbon in an electric furnace to give elemental phosphorus. The latter reaction is extremely complex, and its precise details depend upon the composition of the mineral phosphate. A charge of sand, coke, and phosphate rock is melted at about 1,500° C in an electric furnace. The calcium and impurities are left in the form of a complex fluorosilicate slag, and elemental phosphorus vapour, at about 300° C, distills out and is collected, condensed, and stored underwater as the white allotropic form of the element. More than half a million tons of phosphorus are made annually in the United States in this way. Most of the output is burned to phosphoric anhydride and subsequently treated with water to form phosphoric acid, H_3PO_4.

Only about 5 percent of the 2,000,000 tons of phosphorus consumed per year in the United States is used in the elemental form. Pyrotechnic applications of the element include tracers, incendiaries, fireworks, and matches. Some is used as an alloying agent, some to kill rodents, and the rest is employed in chemical synthesis. A large amount is converted to sulfides used in matches and in the manufacture of insecticides and oil additives. Most of the remainder is converted to halides or oxides for subsequent use in synthesizing organic phosphorus compounds.

Most of the elemental phosphorus produced is converted to high purity phosphoric acid. This is used in industrial processes where impurities present in the acidulated phosphate products would be unacceptable. Much of the pure product is used in soft drinks, metal cleaners, and special liquid fertilizers.

Phosphates in fertilizers

Over 70 percent of the total phosphate rock consumed is used to prepare fertilizers for crops and lawns. If the phosphate is treated with sulfuric acid, the mixture of inert calcium sulfate and calcium hydrogen phosphates is used directly as "superphosphate." When the acid used is phosphoric acid, no inert portion exists, and the product is called "triple superphosphate." Nitric acid also may be used as the acidulant, and this produces a mixture of calcium nitrate and calcium hydrogen phosphates, called "nitrophosphates." These three materials make up about two-thirds of the total phosphate fertilizer produced. Most of the remaining fertilizer phosphate is ammonium phosphate made by neutralizing phosphoric acid with ammonia.

The next most important application of phosphate salts is in synthetic detergents. Nearly 60 percent of the elemental phosphorus produced is converted to detergent components, chiefly sodium tripolyphosphate, $Na_5P_3O_{10}$. Other salts used in detergents, water softeners, and metal cleaners are tetrasodium pyrophosphate, $Na_4P_2O_7$, tetrapotassium pyrophosphate, $K_4P_2O_7$; sodium metaphosphate, $(NaPO_3)_x$; trisodium phosphate (TSP), Na_3PO_4; and sodium dihydrogen phosphate, NaH_2PO_4. In the early 1970s, however, increasing awareness of the heavy pollution of natural waters by phosphates from detergents brought about a search for nonpolluting substitutes.

Phosphate salts also are used as leavening agents for baking and as abrasives in toothpastes. About 5 percent of the total phosphate consumption is in the form of feed supplements for poultry and animals.

Properties and reactions. *Phosphorus.* The electronic configuration of the phosphorus atom can be represented by $1s^2$, $2s^2$, $2p^6$, $3s^2$, $3p^3$, or 2–8–5. The outer shell arrangement therefore resembles that of nitrogen, with three half-filled orbitals each capable of forming a single covalent bond and an additional one pair of electrons. Depending on the electronegativity of the elements with which it combines, phosphorus can therefore exhibit oxidation states of +3 or −3, just as does nitrogen. The principal differences between nitrogen and phosphorus are that the latter is of considerably lower electronegativity and has larger atoms, with outer *d* orbitals available. For these reasons, the similarities between nitrogen and phosphorus chemistry are largely formal ones, tending to conceal the actual, wide differences. The outer *d* orbitals in phosphorus permit an expansion of the octet, which leads to the +5 state, with five actual covalent bonds being formed in compounds, a condition impossible for nitrogen to achieve.

Allotropic forms

The first striking difference in chemistry of the two elements is that elemental phosphorus exists under ordinary conditions in any of several modifications, or allotropic forms, all of which are solid. Phosphorus molecules of formula P_2, structurally analogous to N_2 molecules and evidently also triply bonded, exist only at very high temperatures. These P_2 molecules do not persist at lower temperatures—below about 800°—because of the fact that three single bonds in phosphorus, in contrast to the situation with nitrogen, are favoured over one triple bond. On cooling, the triply bonded P_2 molecules condense to form tetrahedral P_4 molecules, in which each atom is joined to three others by single bonds. These molecules further condense to form either hexagonal- or cubic-structured molecular solids, both called "white phosphorus." Because of the relatively weak intermolecular attractions (van der Waals forces) between the separate P_4 molecules, the solid melts easily at 44.1° C and boils at about 280° C. Formation of tetrahedra requires bond angles of 60° instead of the preferred 90°–109° angles, so that white phosphorus is a relatively unstable, or metastable, form. It changes spontaneously, but slowly, at temperatures around 200° or higher, to a polymeric form called "red phosphorus." This substance is amorphous

when formed at lower temperatures, but it can become crystalline, with a melting point of about 590° C. At higher temperatures and pressures, or with the aid of a catalyst, at ordinary pressures and a temperature of about 200° C, phosphorus is converted to a black crystalline form, which somewhat resembles graphite. This may prove to be the most stable form of phosphorus, despite the relative difficulty in its preparation. In both the red and the black forms, each phosphorus atom forms three single bonds, which are spread apart sufficiently to be relatively strain free.

Consistent with the metastable condition of the white modification, and the crowding of its covalent bonds, this form is far more reactive chemically than the others. It is highly toxic, reacts vigorously with most reagents, and inflames in air at only 35° C, so that it must be stored under water or other inert liquid. White phosphorus dissolves readily in solvents such as carbon disulfide, in which it maintains the composition P_4. In contrast, red phosphorus is insoluble and relatively inert, although large quantities of the usual commercial form can ignite spontaneously in air and react with water to form phosphine and phosphorus oxyacids. Black phosphorus is more inert and is capable of conducting electricity. Both these polymeric forms are insoluble and are very much less volatile than white phosphorus.

Phosphine. Elemental phosphorus does not appear to react with hydrogen directly. Action of strong base, or even hot water, on white phosphorus, however, produces the gaseous compound called phosphine, PH_3. This compound is also produced by hydrolysis of metal phosphides. It melts at −133.8° C and boils at −87.7° C, thus showing far less association than ammonia, which boils more than 50° higher. The molecule is structurally similar to that of ammonia, being pyramidal, as expected, with the lone pair of electrons occupying one corner of a tetrahedron formed by it and three P–H bonds. Although hydrogen is less electronegative than nitrogen, it is slightly more so than phosphorus. The bonds in phosphine are, therefore, only very slightly polar, with the hydrogen partially negative and the phosphorus partially positive (thereby reversing the situation found in the ammonia molecule). In addition, the phosphorus atom is substantially larger than the nitrogen atom, so that its lone-pair electrons are more spread out and thereby less available for attracting a positive hydrogen. Thus, phosphine shows no tendency toward protonic bridging and, correspondingly, liquid phosphine is a much poorer solvent than liquid ammonia. Furthermore, phosphine is much less soluble in water than is ammonia. Similarly, although phosphine does act as an electron-pair donor toward certain Lewis acids, it does not come close to rivalling ammonia in the number and variety of the complexes it forms. Phosphonium salts can be produced by interaction of particularly protonic substances with phosphine, but they are easily dissociated and, in general, much less stable than the corresponding ammonium salts. Of the reaction products with the various hydrogen halides, for example, only phosphonium iodide, PH_4I, is moderately stable at ordinary temperatures.

Phosphonium salts

Phosphine is a good reducing agent and thus is easily oxidized. When it is prepared by the usual method—hydrolysis of calcium phosphide, Ca_3P_2—phosphine contains a small amount of diphosphine, P_2H_4, which is spontaneously inflammable and imparts this property to the mixture. Carefully purified phosphine does not ignite spontaneously, but it is very inflammable. The instability of phosphine is further illustrated by the reaction with atomic nitrogen, which produces the polymeric phosphorous nitride, PN, and liberates hydrogen. Diphosphine is even more unstable; it readily decomposes to elemental phosphorus and phosphine. Diphosphine melts at −99° and boils at 51.7° C.

Oxides and acids. The oxidation of white phosphorus in moist air is often accompanied by a glowing in the dark or phosphorescence. Phosphorus combines very readily with oxygen, the best known products being the so-called phosphorus trioxide, or more properly, phosphorus(III) oxide, P_4O_6, and phosphorus pentoxide, or phosphorus(V) oxide, P_4O_{10}. These bear only a slight, formal resemblance to the oxides of nitrogen with the formulas N_2O_3 and N_2O_5. The structure and bonding of the two pairs of compounds are altogether different, because of the lower electronegativity and larger size of the phosphorus atom and its available outer *d* orbitals.

It may be considered that P_4O_6 is formed by the attachment of single oxygen atoms along each of the six edges of the P_4 tetrahedron, spreading apart the phosphorus atoms by substituting oxygen bridges for the original P–P bonds. This molecule can then easily be oxidized further by joining one oxygen atom to each of the phosphorus atoms at the corners of the P_4O_6 tetrahedron. The last four oxygen atoms, in fact, are found to be bound much more closely and tightly to the phosphorus atoms than the bridging oxygens are, thus indicating a degree of multiple bonding, which probably involves interaction of the outer *d* orbitals of the phosphorus atoms with the lone pairs of electrons on the more tightly bound oxygen atoms.

Phosphorus trioxide is a white solid, which melts at 23.8° C and boils at 173° C. Above 210° C it decomposes yielding, among other products, an intermediate oxide of the empirical formula PO_2, which may correspond to the vapour at high temperatures of formula P_7O_{14}. Phosphorus trioxide slowly takes up oxygen in air to become the pentoxide, P_4O_{10}. The former also dissolves in cold water to form phosphorous acid, H_3PO_3, of which it is thus the anhydride. In hot water, however, the oxide disproportionates, forming phosphine and a mixture of phosphorus acids.

Phosphorous acid is a good reducing agent, being readily oxidized to phosphoric acid, H_3PO_4. It is a weak, diprotonic acid, having only two replaceable hydrogens. Investigation shows that this behaviour corresponds to a tetrahedral structure in which one of the hydrogen atoms is directly attached to a phosphorus atom instead of to oxygen. This is evidently the unreplaceable, nonacidic hydrogen.

Phosphorous and phosphoric acids

Phosphorus pentoxide is a very effective dehydrating agent, being itself hydrated in the process. The first step of its hydration occurs very rapidly to form a polymeric metaphosphoric acid, $(HPO_3)_x$. More slowly, additional water is acquired and orthophosphoric acid, H_3PO_4 commonly called simply phosphoric acid, is formed. It is interesting to note that, again, the similarity to nitrogen chemistry is formal only, in that metaphosphoric acid, a polymeric solid, in almost no way resembles nitric acid, HNO_3, which shows no tendency to polymerize, and that there is no orthonitric acid, H_3NO_4. Phosphoric acid usually is available in the laboratory as a syrupy solution of about 85 percent concentration in water. The high viscosity of this material reflects the extensive protonic bridging that occurs within it. Phosphoric acid is a triprotonic acid; the first ionization constant is that of a moderately strong acid, but the successive ionizations are progressively weaker. A striking difference between nitric and phosphoric acids is that the former is a powerful oxidizing agent, whereas the latter has almost no oxidizing properties at all.

The monohydrogen and dihydrogen phosphates, as well as the neutral phosphates, comprise the salts of phosphoric acid. As is more generally true, the presence of unneutralized hydrogen in these salts makes them much more soluble in water. This effect is the basis for using the calcium hydrogen salts in fertilizers.

Phosphorus–oxygen compounds exist in polymeric form in great variety. There are many families of polyphosphoric acids and their corresponding salts. In a sense these resemble the silicates, in that the phosphorus atoms are surrounded by oxygen atoms in tetrahedral array, some of which are shared with other tetrahedra. The "heteropoly acids" consist of phosphate tetrahedra surrounded by varying numbers of metal-oxide tetrahedra, usually tungstate or molybdate.

Polymeric phosphorus–oxygen compounds

The final important phosphorus acid is hypophosphorous, H_3PO_2. This compound can be formed as the barium salt by treatment of phosphorus with barium hydroxide. The barium salt is converted to the free acid by reaction with sulfuric acid. Hypophosphorous acid is in-

teresting in that only one of its hydrogens is acidic, the other two occupying positions of direct attachment to the phosphorus atom. The acid is a strong reducing agent, as are its salts.

Halides. Phosphorus readily forms a series of halogen-containing compounds: PF_3, PF_5, PCl_3, PCl_5, PBr_3, PBr_5, and PI_3, and many mixed halides as well. These are all volatile, molecular substances, easily susceptible to irreversible hydrolysis, liberating the respective hydrogen halides. They can be made by direct combination of the elements.

Pentahalides

The pentahalides may be thought of as resulting from the promotion of one of the phosphorus lone-pair electrons to an outer *d* orbital, which becomes more available through electron withdrawal from phosphorus by the halogen atoms. As free molecules, or in solution, these substances have trigonal, bipyramidal structures, each comprised of a planar triangle of halogen atoms surrounding the phosphorus atom, with a fourth and fifth halogen above and below the triangle. In the pentafluoride, which is a gas boiling at −84.5° C, it is uncertain whether the apical bonds are greatly different from the equatorial bonds; but in the pentachloride, a volatile solid that sublimes at 159° C, the bonds to the apical chlorine atoms are appreciably longer and only about half as strong as those to the equatorial chlorine atoms. It is possible that the two apical bonds in the chloride are different in nature from those in the fluoride, and perhaps they could be described more accurately as half-bonds.

In the solid state, phosphorus pentachloride has a structure that can be represented as $PCl_4^+PCl_6^-$. The solid state of the corresponding bromide, an orange solid that decomposes above its melting point of about 100° C, can be represented as $PBr_4^+Br^-$. This difference suggests that six bromine atoms cannot be fitted around one phosphorus atom. On the other hand, six fluorine atoms fit very well, for the pentafluoride readily coordinates a fluoride ion to form PF_6^-, of which many salts are known, in addition to the free acid. The failure to detect phosphorus pentaiodide suggests that five iodine atoms cannot be positioned around one phosphorus atom. Also, one would not expect iodine—the least electronegative of the common halogen elements—to activate the *d* orbitals of phosphorus as readily as the other halogens do.

Since the halogen atoms in the pentahalides cannot individually acquire large partial negative charges, these compounds are potentially good halogenating agents. Phosphorus pentachloride is commonly used for this purpose, and, in the process, it becomes converted to phosphorus oxychloride, $POCl_3$. Other, similar, oxyhalides are also known. In these compounds, the oxygen atom is attached by a short, strong bond to the phosphorus atom, a fact that suggests interaction between the oxygen lone pairs and the phosphorus *d* orbitals.

Trihalides

The trihalides of phosphorus range in character from the gaseous fluoride, which boils at −101.2° C, through the liquid chloride and bromide, boiling respectively at 76° C and 176° C, to the red, solid iodide, which melts at 61° C. Ordinarily one would expect the lone-pair electrons on the phosphorus atom in these molecules to be relatively unavailable for coordination because of the partial positive charge induced on the phosphorus atom by the halogen atoms. Although this is generally reflected in the behaviour of these compounds, the trifluoride serves as a ligand (that is, group coordinated with a central atom) in a number of complex compounds. Its behaviour in these compounds is much like that of carbon monoxide in the metal carbonyls—that is, the bonding is substantially reinforced by the use of a vacant orbital (in this case, a phosphorus *d* orbital) to accommodate an electron pair from the metal atom. An interesting reaction of phosphorus trifluoride is its bromination to form the mixed halide PF_3Br_2, which then easily disproportionates to form the simple pentahalides, PF_5 and PBr_5. This is an effective synthesis of PF_5.

Other compounds. The chemical behaviour of phosphorus with sulfur is complex and in some respects surprisingly unlike phosphorus–oxygen chemistry, as well as unlike nitrogen–sulfur chemistry. The principal compounds formed by heating together the appropriate proportions of phosphorus and sulfur are P_2S_5, P_4S_3, P_4S_5, and P_4S_7 (but apparently not P_4S_6). These compounds burn very readily but hydrolyze only slowly, liberating hydrogen sulfide and forming phosphorus acids. Numerous metal phosphides are known, which somewhat resemble the corresponding nitrides.

Analytical chemistry. Elemental phosphorus can be detected by its phosphorescence. It can also be converted to phosphine with boiling sodium hydroxide solution or with zinc and sulfuric acid; the phosphine is identified by means of test paper containing either silver nitrate or mercuric chloride, both of which are reduced to the free metal by phosphine, thereby darkening the paper. Phosphorus vapour also readily darkens silver nitrate test paper. Phosphorus is determined quantitatively by oxidation to phosphate, followed by any of several standard procedures. Phosphate, for example, may be precipitated as the magnesium ammonium salt, $MgNH_4PO_4$, converted by ignition to magnesium pyrophosphate, $Mg_2P_2O_7$, and weighed. Alternatively, phosphate may be precipitated as ammonium phosphomolybdate; this can be weighed as such, converted to magnesium pyrophosphate and weighed, or titrated directly with sodium hydroxide solution.

Biological and physiological significance. Phosphorus is an important constituent of bones and teeth, and it is essential to the growth of living organisms. In organisms the element usually appears as phosphate. In its other forms phosphorus is likely to prove very toxic. White phosphorus attacks the skin and, when ingested, causes a necrosis of the jawbone, called "phossy jaw." Certain organic esters of phosphoric acid, used as lubricating-oil additives, have been found to cause permanent paralysis when accidentally ingested. Phosphine is extremely toxic, as are its organic derivatives. Some of the most toxic substances known to man, collectively termed nerve gas, are organic derivatives of phosphorus.

Nerve gas

Isotopes of phosphorus. The only naturally occurring isotope of phosphorus is that of mass 31. The other isotopes from mass 29 to mass 34 have been synthesized by appropriate nuclear reactions. All of these are radioactive with relatively short half-lives. The isotope of mass 32 has a half-life of about 14 days and has proven extremely useful in tracer studies involving the absorption and movement of phosphorus in living organisms.

ARSENIC AND ITS COMPOUNDS

Occurrence and distribution. The abundance of arsenic in the earth's crust is about five grams per ton; the cosmic abundance is estimated as about four atoms per million atoms of silicon. The element is widely distributed. A small amount exists in the native state, in 90–98 percent purity. Most, however, is combined in more than 150 different minerals, as sulfides, arsenides, sulfoarsenides, and arsenites. Mispickel, or arsenopyrite, FeAsS, is among the most common of arsenic-bearing minerals; others are realgar, As_4S_4; orpiment, As_2S_3; loellingite, $FeAs_2$; and enargite, Cu_3AsS_4. Most commercial arsenic is recovered as a by-product of the smelting of copper, lead, cobalt, and gold ores.

Commercial production and uses. Metallic arsenic forms when arsenopyrite is heated at 650°–700° C in the absence of air. The arsenic in arsenopyrite and the arsenic impurities in other metal ores unite readily with oxygen when heated in air, forming the easily sublimed oxide, As_4O_6, also known as "white arsenic." The vapour of the oxide is collected and condensed in a series of brick chambers and later purified by resublimation. Most arsenic is prepared by carbon reduction of the arsenious oxide dust thus collected.

World consumption of metallic arsenic is relatively small, only a few hundred tons per year. Most of what is consumed comes from Sweden. It is used in metallurgical applications because of its metalloid properties. About one percent arsenic content is desirable in the manufacture of lead shot, for example, because it improves the roundness of the molten drops. Bearing alloys based on lead are improved in both thermal and mechanical properties when they contain about 3 percent arsenic. A small

amount of arsenic in lead alloys hardens them for use in batteries and cable sheathing. Small concentrations of arsenic improve the corrosion resistance and thermal properties of copper and brass. Very highly purified arsenic finds applications in semiconductor technology, where it is used with silicon and germanium, as well as in the form of gallium arsenide, GaAs, for diodes, lasers, and transistors.

Uses of arsenic compounds

In contrast to the small use of metallic arsenic, tens of thousands of tons of the element are consumed annually in the form of its compounds. These are used primarily in agriculture. Calcium arsenate, $Ca_3(AsO_4)_2$, enhances cotton production by controlling boll weevils; it is also used to kill crabgrass. Lead arsenate, $PbHAsO_4$, controls fruit pests, especially the codling moth, whose larvae are very destructive of apples. Sodium arsenite, Na_2AsO_3, helps prevent potato leaf rot. It also debarks trees and kills aquatic weeds in ponds and streams. As an ingredient of sheep and cattle dip, it helps control ticks and other pests. The harvesting of cotton is facilitated by use of arsenic acid, H_3AsO_4, and cacodylic acid, $(CH_3)_2As(O)OH$, as desiccants. These compounds also serve to sterilize soils. Other arsenic compounds are used as selective herbicides and as additives in the feed of swine and poultry. The pentasulfide is an ingredient of fireworks, and it is used in the making of infrared lenses. In glass manufacture pure arsenious oxide is an effective decolorizing agent.

Properties and reactions. *Arsenic.* In its most stable free state, arsenic is a steel-gray, brittle solid with low thermal and electrical conductivity. Other forms have been reported but are not well characterized, including especially a yellow, metastable form, which may consist of As_4 molecules analogous to white phosphorus, P_4. Arsenic sublimes at 613° C, and in the vapour it exists as As_4 molecules, which do not begin to dissociate until about 800° C; dissociation to As_2 molecules becomes complete at about 1,700° C.

Electronic structure

The electronic structure of the arsenic atom, 2–8–18–5, resembles those of nitrogen and phosphorus in that there are five electrons in the outermost shell, but it differs from them in having 18 electrons in the penultimate shell instead of two or eight. The addition of ten positive charges to the nucleus during the filling of the five 3*d* orbitals frequently causes a general contraction of the electronic cloud and a concomitant increase in electronegativity of the elements. In other groups of the periodic table this is clearly shown. Thus, it seems generally accepted that zinc is more electronegative than magnesium and, similarly, that gallium is more electronegative than aluminum. The difference diminishes, however, in the next groups, and many authorities do not agree that germanium is more electronegative than silicon, although an abundance of chemical evidence appears to indicate that this is so. The similar transition from penultimate 8-shell to 18-shell element in passing from phosphorus to arsenic might also produce an increase in the electronegativity of arsenic over phosphorus, but this remains controversial.

Bonding properties

The outer-shell similarity of the two elements suggests that arsenic, like phosphorus, can form three covalent bonds per atom, with an additional lone pair of electrons left unbonded. The oxidation state of arsenic should, therefore, be either +3 or −3 depending on the electronegativity of arsenic and that of the elements with which it is combined. The possibility should also exist of utilizing the outer *d* orbitals to expand the octet, thereby allowing arsenic to form five bonds. This possibility is realized only in compounds with fluorine. The availability of the lone pair for complex formation (through electron donation) appears much less in the arsenic atom than in phosphorus and nitrogen, as evidenced by the chemistry of the element.

Arsenic itself is stable in dry air, but in moist air it tends to become coated with a black oxide. Sublimed arsenic vapour readily burns in air to form arsenious oxide. The free element is essentially unaffected by water, bases, or nonoxidizing acids, but it can be oxidized by nitric acid to the +5 state. Halogens attack arsenic, as does sulfur, and the element will combine directly with many metals forming arsenides.

Arsine. The compound arsine, AsH_3, is evidently too unstable to be formed from the elements by direct combination, but it is readily produced by the hydrolysis of metal arsenides and by the reduction by metals of arsenic compounds in acidic solutions. Since arsine is an extremely toxic gas, it presents a potential hazard whenever metal and acid—either of which may contain arsenic impurities—are brought into contact. Arsine is not very soluble in water. It boils at −62.5° C. Its molecules are pyramidal, with bond angles of 97.5°. The instability of the substance is evidenced by the formation of a "mirror" of arsenic metal and the liberation of hydrogen at about 300° C. Arsine is a strong reducing agent and burns readily to water and oxides. In a limited oxygen supply, only the hydrogen burns, and free arsenic remains behind.

Oxides and acids. The principal oxides of arsenic are arsenious oxide, or arsenic(III) oxide, As_4O_6, formed by burning arsenic, and arsenic pentoxide, or arsenic(V) oxide, As_2O_5. The latter compound cannot be formed by direct oxidation with oxygen of either the free element or arsenious oxide. Like nitrogen pentoxide, it must be prepared by dehydration of the corresponding acid, and it is unstable toward heat, being decomposed completely above 400° C to oxygen and arsenious oxide. In this respect, as in many others, the chemistry of arsenic appears to resemble that of nitrogen somewhat more than it does that of phosphorus; like nitrogen and unlike phosphorus, arsenic in the +5 state is strongly oxidizing.

Arsenious oxide, or white arsenic, is very stable thermally. It is easily reduced to free arsenic, in contrast to phosphorus trioxide, which tends rather to be itself a reducing agent. The arsenic compound is moderately soluble in water forming solutions called arsenious acid, which are very weakly acidic and possibly slightly basic as well. The exact species present in solution are not well established, and no free arsenious acid as such has been isolated—again, unlike phosphorus chemistry, where the corresponding phosphorous acid is a crystalline solid.

The arsenic acids

Arsenic acid, whose molecular formula is usually written as H_3AsO_4, is formed by the action of various strong oxidizing agents on arsenious oxide. The compound isolated from solution has the formula $As_2O_5 \cdot 4H_2O$. It is a fairly strong oxidizing agent, capable of oxidizing chloride ions to free chlorine and sulfurous acid to sulfuric. It is a triprotonic acid, the first dissociation constant of which is about half that of phosphoric acid, but the second and third of which are much larger. When heated to 200° C, arsenic acid is converted to the so-called meta acid, $HAsO_3$. This substance gives no evidence of being highly polymeric like the phosphorus analogue, and, in fact, it is very quickly hydrated when water is added. In general, arsenic–oxygen compounds tend to polymerize much less than do phosphorus–oxygen compounds.

Halides. As expected, arsenic readily forms all four of the possible trihalides by direct combination of the elements. The trifluoride, AsF_3, a liquid boiling at 58° C, can also be formed by heating together solid arsenious oxide and calcium fluoride. The trichloride is also a liquid, boiling at 130° C. Arsenic tribromide is a white solid melting at 31.2° C and boiling at 221° C. The triiodide is a red solid, which melts at about 145° C and boils somewhere in the vicinity of 400° C. All four halides hydrolyze, but slowly. This is especially true of the fluoride and the iodide, which is not very soluble in water.

Instability of pentahalides

Arsenic pentafluoride, AsF_5, is a colourless gas that condenses to a yellow liquid boiling at −52.6° C. It is not very stable and loses fluorine at moderate temperatures. It acts, therefore, as a powerful fluorinating agent. The compound is much less stable than either phosphorus or antimony pentafluoride. Consistent with this fact is the complete failure to prepare arsenic pentachloride, although both the phosphorus and antimony compounds are well known. These examples of the relative instability of arsenic in the +5 state, as compared to both phosphorus and antimony, are in line with the supposition that the electronegativity of arsenic is higher than that of either phosphorus or antimony and that its bonds to halogen are, therefore, less polar and less stable. The pentabromide and pentaiodide of arsenic are also nonexistent.

Other compounds. Arsenic combines directly with many metals forming arsenides, which have a general resemblance to the nitrides and phosphides. Arsenic also forms thousands of organic compounds, which have been investigated since early times because of great interest in their possible medicinal properties. Among the better known organic arsenic compounds is tetramethyldiarsine, or cacodyl, $(CH_3)_2As–As(CH_3)_2$, used in preparing the herbicide cacodylic acid. This was first discovered as a component of "Cadet's liquid"—chiefly cacodyl oxide, $(CH_3)_2As–O–As(CH_3)_2$—which is obtained by dry distillation of arsenious acid and potassium acetate. In the absence of air, cacodyl is stable up to 400° C, but it is spontaneously inflammable, giving as its primary oxidation product cacodyl oxide.

Analytical chemistry. Qualitatively, arsenic may be detected by precipitation as the yellow arsenious sulfide from hydrochloric acid of 25 percent or greater concentration. Trace amounts of arsenic are usually determined by conversion to arsine. The later can be detected by the so-called Marsh test, in which arsine is thermally decomposed, forming a black arsenic mirror inside a narrow tube, or by the Gutzeit method, in which a test paper impregnated with mercuric chloride darkens when exposed to arsine because of the formation of free mercury.

Tests for arsine

Biological and physiological significance. The toxicity of arsenic and its compounds varies widely, ranging from the exceedingly poisonous arsine—and its organic derivatives—to elemental arsenic itself, which is relatively inert. Arsenical compounds in general are skin irritants, which easily cause dermatitis. Protection against inhalation of arsenic-containing dusts is recommended, but most poisoning appears to come from ingestion. The maximum tolerable concentration of arsenic in dusts during an eight-hour day is 0.5 milligrams per cubic metre. For arsine, exposure of similar duration requires that the concentration be less than 0.05 parts per million in the air. In addition to the many uses of arsenic compounds as herbicides and pesticides, they have in several instances been employed as pharmacological agents. The first successful antisyphillitic agent, for example, was an arsenic compound, "Salvarsan," or "606," or 3,3′-diamino-4,4′-dihydroxyarsenobenzene dihydrochloride.

Isotopes of arsenic. Only one stable isotope of arsenic, that of mass 75, occurs in nature. Among the artificial radioactive isotopes is one of mass 76, which has a half-life of 26.4 hours.

ANTIMONY

Occurrence and distribution. Antimony is about one-fifth as abundant as arsenic, contributing on the average about one gram to every ton of the earth's crust. Its cosmic abundance is estimated as about one atom to every 5,000,000 atoms of silicon. Small deposits of native metal have been found, but most of the antimony occurs in the form of more than 100 different minerals. The most important of these is stibnite, Sb_2S_3. Small stibnite deposits are found in Algeria, Bolivia, China, Yugoslavia, Mexico, Peru, and South Africa. Some economic value also attaches to kermesite ($2Sb_2S_3 \cdot Sb_2O_3$), argentiferous tetrahedrite [$(Cu,Fe)_{12}Sb_4S_{13}$], livingstonite ($HgSb_4S_7$), and jamesonite ($Pb_4FeSb_6S_{14}$). Small amounts are also recoverable from the production of copper and lead. About half of all the antimony produced is reclaimed from scrap lead alloy from old batteries, to which antimony had been added to provide hardness.

Stibnite

Commercial production and uses. High-grade or enriched stibnite reacts directly with scrap iron in the molten state, liberating antimony metal. The metal can also be obtained by conversion of stibnite to the oxide, followed by reduction with carbon. Sodium sulfide solutions are effective leaching agents for the concentration of stibnite from ores. Electrolysis of these solutions produces antimony. After further purification of the crude antimony, the metal, called regulus, is cast into cakes.

About half of this antimony is used metallurgically, principally in alloys. It improves the hardness and corrosion resistance of lead. Most of the metal is used for lead alloys, largely for storage batteries, but also for chemical equipment such as tanks, pipes, and pumps. With tin, antimony forms such alloys as britannia metal and pewter, used for utensils and Babbitt metal for bearings. Other applications of antimony alloys are in solder, type metal, and other special materials. Highly purified antimony is used in semiconductor technology to prepare the intermetallic compounds indium, aluminum, and gallium antimonide for diodes and infrared detectors.

The principal commercial compound of antimony is its oxide, Sb_2O_3. This substance is used to prepare opaque enamels for ceramics and metalware, and it is employed as a white pigment for paints. It also aids in flameproofing fabrics and plastics and in decolorizing glass. It finds some use in medicine, in dyeing, and in staining iron and copper. Antimony sulfide, Sb_2S_5, is used as a vulcanizing agent and Sb_2S_3 as a vermilion pigment, and as a component of fireworks, ammunition primers, and tracer bullets.

Properties and reactions. *Antimony.* The most stable form of antimony is a brittle, silvery solid of high metallic lustre. Electrolytic deposition of antimony under certain conditions produces an unstable, amorphous form called "explosive antimony," because, when bent or scratched, it will change in a mildly explosive manner to the more stable, metallic form. There is also an amorphous black form of antimony that results from sudden quenching of the vapour, and a yellow form produced by low temperature oxidation of stibine, SbH_3, with air or chlorine. Metallic antimony is not affected by air or moisture under ordinary conditions, but it can be oxidized easily by oxygen, sulfur, and the halogens, especially when heated.

The electronic structure of antimony closely resembles that of arsenic. It is represented as 2–8–18–18–5, with three half-filled orbitals in the outermost shell. Thus it can form three covalent bonds and exhibit +3 and −3 oxidation states. The electronegativity of antimony, like that of arsenic, remains somewhat controversial. It is generally agreed to be lower than that of arsenic, but whether it is lower also than that of phosphorus is undecided. It can act as an oxidizing agent and reacts with many metals to form antimonides that, in general, resemble nitrides, phosphides, and arsenides, but are somewhat more metallic. The promotion of one of the lone-pair electrons to an outer *d* orbital apparently occurs more easily with antimony than with arsenic, since antimony exhibits the +5 oxidation state in forming both the pentafluoride and the pentachloride.

Electronic structure

Stibine. The formation and properties of stibine, SbH_3, closely resemble those of arsine. Direct combination of antimony with hydrogen does not occur, but reduction of antimony compounds, or hydrolysis of metal antimonides, produces the gaseous stibine, which boils at −17° C. This compound is even less stable than arsine, separating into its elements slowly at 25° C and very rapidly at 200° C. It is a good reducing agent but inactive as an electron donor. It is very toxic.

Oxides and acids. Antimony burns in oxygen, forming principally the so-called trioxide and tetroxide, Sb_4O_6 and Sb_4O_8, respectively. The element is converted to the pentoxide, Sb_2O_5, by concentrated nitric acid. Since the pentoxide is somewhat unstable, its hydrated form loses small amounts of oxygen when dehydrated by gentle heating.

The molecules of the trioxide, which are stable even above 1,500° C in the gas phase, do not form definite hydroxides. Even in the presence of liquid water, the hydrous oxide tends to revert to the oxide by losing water slowly. The species in solution are not well identified, but the solution displays both acidic and basic properties, the supposed "antimonous acid" being somewhat more basic and less acidic than arsenious acid. The salts formed appear to be mainly the salts of the meta acid, $HSbO_2$.

Antimony tetroxide appears to be a stable oxide, especially between 300° and 900° C. It is formed by heating either the pentoxide or the trioxide in air. The pentoxide loses oxygen to form the tetroxide above 300° C, and the trioxide combines with additional oxygen to form the tetroxide up to about 900° C, above which temperature the additional oxygen is lost again. The tetroxide, like the other oxides, is a white solid. It is slightly soluble in water,

Antimony tetroxide

giving a weakly acidic solution, and it dissolves readily in alkaline but not in acid solutions. These solutions contain both antimonites and antimonates, but no Sb(IV) compounds, suggesting that the tetroxide is either a double oxide of Sb_4O_6 and Sb_2O_5 or an antimony(III) antimonate, $SbSbO_4$.

Antimonic acid, a poorly characterized colloidal material, is amphoteric but much weaker in acidity than arsenic acid. There is some evidence that antimonic acid should be represented by the formula $H[Sb(OH)_6]$.

Halides. Antimony pentafluoride, SbF_5, resembles the corresponding phosphorus and arsenic compounds except that, instead of being gaseous, it is a liquid boiling at 149° C. It can be formed by reaction of hydrofluoric acid with antimony pentachloride, which in turn is formed by direct chlorination of antimony trichloride. Like the phosphorus compound, but unlike the arsenic compound, antimony pentafluoride can coordinate fluoride ions, and a number of salts of the resulting SbF_6^- ion are known. Antimony pentachloride is a liquid that freezes at 2.8° C and loses chlorine at about 140° C. In the solid state it appears to consist of $SbCl_4^+Cl^-$. Many metal hexachloroantimonates are known. Antimony forms no pentabromide or pentaiodide.

The antimony trihalides are easily formed by direct halogenation. The trifluoride is a white solid melting at 290° C, and boiling more than 300° higher than the corresponding arsenic compound. Antimony trichloride, called "butter of antimony," is a soft, colourless solid, which melts at 73.2° C and boils at 221° C. It is much more saltlike than is arsenic trichloride. In dilute solution antimony trichloride hydrolyzes to give basic chlorides, such as antimony oxychloride, SbOCl. The trichloride forms many complex compounds with the general formula M_3SbCl_6. Antimony tribromide, which melts at 96.6° C, is very similar to the chloride. The triiodide is a yellow or red solid melting at 171° C.

Other compounds. Antimony forms many other types of compounds, including sulfides, which have industrial importance in the manufacture of rubber. Large numbers of organic compounds have been investigated, some for possible medicinal applications, but, in general, these—like arsenic compounds—tend to be too toxic for wide use.

Analytical chemistry. Antimony may be separated and weighed for analysis as the sulfide, Sb_2S_3. Alternatively, the sulfide may be converted to the oxide and, after careful ignition, weighed as Sb_4O_6. Numerous volumetric methods are also available, including several methods of oxidizing Sb(III) with potassium permanganate, potassium bromate, or iodine. In the absence of arsenic, small amounts of antimony may be determined by a modified Gutzeit method.

Biological and physiological significance. Antimony and a number of its compounds are highly toxic. In fact, the use of antimony compounds for medicinal purposes was temporarily outlawed several centuries ago because of the number of fatalities they had caused. A hydrated potassium antimonyl tartrate called "tartar emetic" is currently used in medicine as an expectorant, diaphoretic, and emetic. The maximum tolerable concentration of antimony dust in air is about the same as for arsenic, 0.5 milligrams per cubic metre.

Tartar emetic

Isotopes of antimony. Two stable isotopes, nearly equal in abundance, occur in nature. One has mass 121 and the other mass 123. Radioactive isotopes of masses 120, 122, 124, 125, 126, 127, 129, and 132 have been prepared.

BISMUTH AND ITS COMPOUNDS

Occurrence and distribution. Bismuth is about as abundant as silver, contributing about 2×10^{-5} weight percent of the earth's crust. Its cosmic abundance is estimated as about one atom to every 7,000,000 atoms of silicon. It occurs both native and in compounds. In the native state, it is found in veins associated with lead, zinc, tin, and silver ores in Bolivia, Canada, England, and Germany. Its naturally occurring compounds are chiefly the oxide (bismite or bismuth ochre), the sulfide (bismuthinite or bismuth glance), and two carbonates (bismutite and bismutosphaerite). Commercial bismuth, however, is produced largely as a by-product in the smelting and refining of lead, tin, copper, silver, and gold ores. Thus, it comes—for example—from tungsten ores in South Korea, lead ores in Mexico, copper ores in Bolivia, and both lead and copper ores in Japan.

Commercial production and uses. Bismuth is volatile at high temperature, but it usually remains with the other metals after smelting operations. Electrolytic refining of copper leaves bismuth behind as one component of the anode sludge. Separation of bismuth from lead by the Betterton–Kroll process involves the formation of high-melting calcium or magnesium bismuthide (Ca_3Bi_2 or Mg_3Bi_2), which separates and can be skimmed off as dross. The dross may be chlorinated to remove the magnesium or calcium, and finally the entrained lead. Caustic soda treatment then produces highly pure bismuth. An alternative separation, the Betts process, involves electrolytic refining of lead bullion (containing bismuth and other impurities) in a solution of lead fluosilicate and free fluosilicic acid, bismuth being recovered from the anode sludge. Separation of bismuth from its oxide or carbonate ores can be effected by leaching with concentrated hydrochloric acid. Dilution then precipitates the oxychloride, BiOCl. This—heated with lime and charcoal—produces metallic bismuth.

Alloys of bismuth

Metallic bismuth is used principally in alloys, to many of which it imparts its own special properties of low melting point and expansion on solidification. Bismuth is thus a useful component of type-metal alloys, which make neat, clean castings; and it is an important ingredient of low-melting alloys, called fusible alloys, which have a large variety of applications, especially in fire-detection equipment. A bismuth–manganese alloy has been found effective as a permanent magnet. Small concentrations of bismuth improve the machinability of aluminum, steel, stainless steels, and other alloys and suppress the separation of graphite from malleable cast iron. Thermoelectric devices for refrigeration make use of bismuth telluride, Bi_2Te_3, and bismuth selenide, Bi_2Se_3. Liquid bismuth has been used as a fuel carrier and coolant in the generation of nuclear energy.

The principal chemical application of bismuth is in the form of bismuth phosphomolybdate, which is an effective catalyst for the air oxidation of propylene and ammonia to acrylonitrile. The latter is used to make acrylic fibres, paints, and plastics. Pharmaceutical uses of bismuth have been practiced for centuries. It is effective in indigestion remedies and antisyphilitic drugs. Slightly soluble or insoluble salts are utilized in the treatment of wounds and gastric disorders, and bismuth is sometimes injected in the form of finely divided metal, or as suspensions of its insoluble salts. Substantial quantities of the oxychloride, BiOCl, have been used to impart a pearlescent quality to lipstick, nail polish, and eye shadow.

Properties and reactions. Bismuth is a rather brittle metal with a somewhat pinkish, silvery metallic lustre. It undergoes a 3.3 percent expansion when it solidifies from the molten state. Its electrical conductivity is very poor, but somewhat better in the liquid state than in the solid. With respect to thermal conductivity, it is the poorest of all metals except mercury. Bismuth is quite resistant to corrosion by air and moisture, but it is oxidized rapidly at its boiling point of 1,560° C. It is oxidized and dissolved by concentrated nitric acid.

Bismuth atoms have the same electronic structure in their outermost shell as do the other elements of the nitrogen group. They can, therefore, form three single covalent bonds, exhibiting either a $+3$ or -3 oxidation state. The element has a somewhat lower electronegativity than the others, and its lone pair of electrons is evidently quite inert, causing the $+5$ state of bismuth to be rare and unstable.

Bismuth compounds

Bismuth forms a very unstable compound with hydrogen, bismuthine, BiH_3, which boils at about 22° C and slowly decomposes above that temperature. Ignition of the basic nitrate produces bismuth trioxide, Bi_2O_3, a yellow solid that is stable even above 1,750° C. The oxide is

nonacidic but only very weakly basic; it forms no definite hydroxide. Powerful oxidants in alkaline solution oxidize bismuth to the +5 state, and acidification then precipitates a red-brown solid (approximating Bi_2O_5 in composition), which is quite unstable, losing some oxygen even on drying at 100° C. This product appears to be acidic. Fluorine is the only halogen capable of oxidizing bismuth to the +5 state; it produces a fairly stable, white, solid pentafluoride, BiF_5, which sublimes at 550° C and appears to decompose at still higher temperatures. The other halides of bismuth are all solid trihalides. They hydrolyze readily to form oxyhalides, which are generally insoluble. Bismuth also forms sulfides and various other salts, including the sulfate. It can form bismuthides with active metals.

Analytical and physiological chemistry. Bismuth is usually determined gravimetrically, being precipitated and weighed as the phosphate or the oxychloride, BiOCl. To produce the latter, a suitable amount of hydrochloric acid is added to a nitric acid solution containing the bismuth, and the resulting solution is poured into a large volume of water, causing the oxychloride to precipitate. Volumetric and colorimetric methods of determination are also available.

Bismuth is relatively nontoxic, the least so of the heavy metals. It is generally not an industrial hazard. Although bismuth and certain of its compounds find considerable therapeutic use, some authorities recommend that other remedies be substituted. Soluble inorganic bismuth compounds are toxic.

Isotopes of bismuth. Bismuth forms only one stable isotope, that of mass 209. A large number of radioactive isotopes are known, as shown in the table, most of them being very unstable.

BIBLIOGRAPHY. A short concise summary of nitrogen chemistry is contained in W.L. JOLLY, *The Inorganic Chemistry of Nitrogen* (1964). More detailed is the comprehensive compendium edited by C.A. STREULI and P.R. AVERELL, *Analytical Chemistry of Nitrogen and Its Compounds*, 2 vol. (1970), which includes especially useful tables. For a popularized, readable, introductory account, see ISAAC ASIMOV, *The World of Nitrogen*, rev. ed. (1962). Information about nitrogen and its compounds, as well as about the other elements of the nitrogen group and their compounds, is found in standard textbooks and treatises of inorganic chemistry. Among these are: R.T. SANDERSON, *Inorganic Chemistry* (1967), a modern textbook emphasizing the periodicity of the chemical elements by surveying their binary compounds with each of principal nonmetals; and *Chemical Bonds and Bond Energy* (1970), a brief book giving new insights into the nature of chemical bonds; HEINRICH REMY, *Grundriss der anorganischen Chemie*, 5th ed. (1955; Eng. trans., *Treatise on Inorganic Chemistry*, 2 vol., 1956), a thorough exposition; E.G. ROCHOW, *The Metalloids* (1966), a brief, very readable survey of the elements that border between metallic and nonmetallic; and N.V. SIDGWICK, *The Chemical Elements and Their Compounds* (1950), a classic, thoroughly documented treatise on inorganic chemistry, largely descriptive in nature.

(R.T.S.)

Nizām al-Mulk

Nizām al-Mulk (Regulator of the Kingdom; personal name Abū ʿAlī Ḥasan ibn ʿAlī), the Iranian vizier of the Seljuq sultans Alp-Arslan (1063–72/73) and Malik-Shāh (1072/73–92), was an outstanding statesman and administrator.

Born in 1018 or 1019 near Ṭūs, in the province of Khorāsān, of a father who was a revenue official for the Ghaznavid dynasty that, before the invasion of the Seljuqs, ruled over Khorāsān from their centre at Ghazna (present-day Ghaznī, Afghanistan), he may have studied with one of the outstanding Shāfiʿī teachers of Nīshāpūr, the imam Muwaffaq. Through his father's position and his family's connections with the sayyids of Bayhaq, he was born into the literate, cultured milieu of the Persian administrative class, a background that molded his attitudes and determined his career. In the years of confusion following the initial Seljuq expansion, his father fled, eventually to Ghazna, where Nizām al-Mulk, too, in due course entered Ghaznavid service. He soon returned to Khorāsān, however, and joined the service of Alp-Arslan, then governing Khorāsān. When Alp-Arslan's vizier died, Nizām al-Mulk was appointed to succeed him, and, when Alp-Arslan himself succeeded his father in 1059, Nizām al-Mulk had the entire administration of Khorāsān in his hands. His abilities so pleased his master that, when Alp-Arslan became the supreme overlord of the Seljuq rulers in 1063, Nizām al-Mulk was made vizier.

For the next 30 years, under two remarkable rulers, he held this position in an empire that stretched from the Oxus in the east, to Khwārezm and the southern Caucasus, and westward into central Anatolia. During these decades, the Seljuq Empire was at its zenith; Nizām al-Mulk's influence guided the Sultan's decisions, sometimes even military ones, and his firm control of the central and provincial administration, through his many dependents and relatives, implemented those decisions. His influence was especially felt in the rule of Malik-Shāh, who succeeded to the throne when he was only 18. Indeed, Nizām al-Mulk's boast shortly before his assassination (1092) was substantially true: "Tell the Sultan, 'If you have not already realized that I am your co-equal in the work of ruling, then know that you have only attained to this power through my statesmanship and judgment'." Such was his reputation among contemporaries that he was compared to the Barmakids, viziers to the 8th-century caliph Hārūn al-Rashīd.

Influence in Seljuq policy

His aim, like that of other great Persian viziers, was to impress on his less sophisticated Turkmen rulers, brought up in the rude tradition of the steppe, the superiority of Persian civilization and its political wisdom. Nizām al-Mulk's conception of society was based partly on the ideals that he had inherited from his background, a Persian tradition of order and hierarchy in the state that reached back beyond the Arab conquest of Persia in the 7th century to the traditions of Sāsānian society. The ruler, chosen by God, had as his main task the preservation of stability in the kingdom and the traditional forms of society. His power was absolute, requiring no authorization, and the administration was centralized in his person.

Yet, despite his immense power and prestige, Nizām al-Mulk was only in some measure able to mold his sultans in this ideal of kingship. Shortly before his death, at Malik-Shāh's request, he wrote down his views on government in the *Seyāsat-nāmeh* (Eng. trans., *The Book of Government; or, Rules for Kings*, 1960). In this remarkable work, Nizām al-Mulk barely refers to the organization of the dewan (administration) because he had been able, with the help of his well-chosen servants, to control and model it on traditional lines. But he never had the same power in the *dargāh* (court) and found much to criticize in the Sultan's careless disregard for protocol, the lack of magnificence in his court, the decline in prestige of important officials, and the neglect of the intelligence service.

The *Seyāsat-nāmeh*

The most severe criticisms in the *Seyāsat-nāmeh*, however, are of those with heterodox religious views, the Shīʿī in general and the Ismāʿīlīs in particular, to whom he devotes his last 11 chapters. His support of "right religion," Sunnī Islām, was not only for reasons of state but also a matter of passionate conviction, providing the other major influence that shaped his view of society. His relations with the ʿAbbāsid caliphate show him interested only in extending the Sultan's and his own influence, until 1086 when the Caliph's gracious reception made him his fervent champion in the later disputes between the Caliph and the Sultan. He expressed his religious devotion in other ways, however, that more effectively contributed to the Sunnī revival. He founded Nizāmīyah *madrasah*s (colleges of higher learning) in many major towns throughout the empire to combat Shīʿī propaganda, as well as to provide reliable, competent administrators, schooled in his own branch of Islāmic law. Less orthodox religious communities among the Ṣūfī orders also benefitted from his generosity; hospices, pensions for the poor, and extensive public works related to the pilgrimage to Mecca and Medina were created or sustained by his patronage. Particularly in his last years, when the Ismāʿīlī threat grew stronger and found a

Opposition to the Assassins

refuge in Alamūt, the castle of the Assassins, he set himself the task of combatting their influence by every means possible. In the closing sections of the *Seyāsat-nāmeh*, when he himself was surrounded by court intrigues, had lost the Sultan's confidence, and felt too acutely the dangers threatening the empire and orthodoxy, he reminded Malik-Shāh of the validity of his warnings:

> My words will be remembered when they [the Ismā'īlīs] begin to throw into the pit the distinguished and the great, when the ears echo with the sound of their drums, and when their evil practices and intrigues are laid bare. At the time of this misfortune he [the Sultan] will realize that I was right in all that I said and never refused to offer any possible advice or goodwill.

His words were prophetic. Niẓām al-Mulk was assassinated on October 14, 1092, on the road from Isfahan to Baghdad, near Nehāvand. The murder was probably committed by an Ismā'īlī from Alamūt, possibly with the complicity of Niẓām al-Mulk's court rival, Tāj al-Mulk, and of the queen, Terkhen-Khatun, if not that of Malik-Shāh himself. Within a month, however, the Sultan, too, was dead, and the disintegration of the great empire had begun.

Despite his accumulation of personal wealth, his nepotism and the arrogance of some of his relatives, and his occasional acts of ruthlessness, Niẓām al-Mulk was, for contemporaries, as he has remained for successive generations who read his *Seyāsat-nāmeh*, the quintessential vizier—wise, prudent, resourceful and successful, and a devout Muslim. By his life and work, he brought the Persian and the Islāmic cultures toward a closer integration at a time when medieval Islām reached its zenith.

BIBLIOGRAPHY. The only complete English translation of Niẓām al-Mulk's *Seyāsat-nāmeh* is by HUBERT DARKE, *The Book of Government; or, Rules for Kings* (1960). There is an older French translation using a different text by CHARLES SCHEFER, *Siasset Namèh, traité du gouvernement*, 2 pt. (1891–93); and a more recent German translation by K.E. SCHABINGER VON SCHOWINGEN, *Siyāsatnāma* (1960), with a useful introduction. There is no biography; HAROLD BOWEN'S article in the *Enclyclopaedia of Islam*, 1st ed., vol. 3 (1936), gives a clear outline of his life; and E.G. BROWNE in *A Literary History of Persia*, vol. 2 (1906), assesses the value of his work.

Nkrumah, Kwame

A political organizer of outstanding ability, the African statesman Kwame Nkrumah achieved the liberation of the Gold Coast from British colonial rule. As first prime minister of independent Ghana—as the former colony came to be called—and subsequently president of the republic, Nkrumah's policies made him a figure of acute controversy: inflexible dictator, according to his opponents; redeemer of Africa, according to his followers. As theoretician of the anti-imperialist cause, his writings achieved a major impact and remain among the more cogent expressions of the black liberation movement.

Early years

Kwame Nkrumah was born in September 1909, at Nkroful, a coastal village. His father was a goldsmith and his mother a retail trader. Baptized a Roman Catholic, Nkrumah spent nine years at the Roman Catholic elementary school in nearby Half Assini. After graduation from Achimota College in 1930, he started his career as teacher at Catholic junior schools in Elmina and Axim and at a seminary, where he became seriously attracted to the priesthood. In 1934, however, he came under the influence of Nnamdi Azikwe and other exponents of the new African nationalism. Increasingly drawn to politics, Nkrumah decided to pursue further studies in the United States. He entered Lincoln University, Pennsylvania, in 1935 and, after graduating in 1939, obtained masters degrees from Lincoln and from the University of Pennsylvania. He began to clarify his ideological position after intensive study of the literature of Socialism, notably Marx and Lenin, and of nationalism, especially Marcus Garvey, the American leader who in the 1920s sought to promote unity among the Negroes of the world and to establish a black-governed Negro nation in Africa. Eventually, he came to describe himself as a "nondenominational Christian and a Marxist socialist." He also immersed himself in political work, reorganizing and becoming president of the African Students' Association. He left the United States in May 1945, intending to study law in Britain, but instead he enrolled at the London School of Economics and Political Science. He became vice president of the West African Students Union; an official of other African organizations; and leader of "The Circle," an experiment in the creation of revolutionary activist cells. In 1947, in his first major publication, *Towards Colonial Freedom*, he outlined an ideological blueprint for the anticolonial struggle.

Nkrumah, 1962.
Marc and Evelyne Bernheim—Rapho Guillumette

Meanwhile, in the Gold Coast, J.B. Danquah had formed the United Gold Coast Convention (UGCC) to work for self-government by constitutional means. Invited to serve as its general secretary, Nkrumah returned home in late 1947. Appealing particularly to the youth and former servicemen, he addressed meetings throughout the country and began to create a mass base for the new movement. When extensive riots occurred in February 1948, the British ordered the detention of Nkrumah and other leaders of the UGCC, but he was released in April to testify before a commission of inquiry. As a split developed between the middle class leaders of the UGCC and the more radical supporters of Nkrumah, he launched, in September 1948, the *Accra Evening News* as a vehicle for his views and in June 1949 announced the creation of the new Convention Peoples' Party (CPP), which was committed to a program of immediate self-government. In January 1950, he initiated the campaign of Positive Action: "Non-violence and non-cooperation—the adoption of all legitimate and consitutional means by which we can cripple the forces of imperialism in the country." In the ensuing crisis, services throughout the country were disrupted, and Nkrumah was again arrested and sentenced to one-year's imprisonment with hard labour. But the first general election (February 8, 1951) demonstrated the support the CPP had already won. Elected to Parliament, Nkrumah was released from prison to become leader of government business and, in 1952, prime minister. During a short visit to the U.S. in 1951, he received the LL.D. degree from Lincoln University.

Release from prison and entry into government

When the Gold Coast and the British Togoland trust territory became an independent state within the British Commonwealth—as Ghana—in March 1957, the CPP held 72 of the 104 seats in Parliament. Such opposition as had emerged was regionally based. Maintaining such a situation was inherently dangerous; in December 1957, Nkrumah forced the regional parties to merge into a national opposition. The Preventive Detention Act of the following year legalized imprisonment without trial of those regarded as security risks. It soon became ap-

parent that Nkrumah's style of government was to be authoritarian—in the tradition not only of the preceding colonial administration but also of the powerful 19th-century kingdoms, such as Ashanti, out of which, historically, Ghana had evolved. Nkrumah's popularity in the country rose, however, as new roads, schools, and health facilities were built and as the policy of Africanization created better career opportunities for Ghanaians.

By a plebiscite of 1960 Ghana became a republic and Nkrumah became its president, with wide legislative and executive powers under a new constitution. Nevertheless, whole sectors of the economy remained essentially controlled by foreign interests, including cocoa marketing, insurance, and banking. The Second Development Plan, announced in 1959, had to be abandoned in 1961 when the deficit in the balance of payments rose to more than $125,000,000. Contraction of the economy led to widespread labour unrest and to a general strike in September 1961. From that time on Nkrumah began to evolve a much more rigorous apparatus of political control and to turn increasingly to the Communist countries for support. The attempted assassination of Nkrumah at Kulugungu in August 1962—the first of several—led to his increasing seclusion from public life and to the growth of a personality cult, as well as to a massive build-up of the country's internal security forces. Early in 1964 Ghana was officially designated a one-party state, with Nkrumah as life president of both nation and party. While the administration of the country passed increasingly into the hands of party officials, many of whose standards of public service were unfortunately low, Nkrumah, although introducing into the government a number of talented and dedicated intellectuals, seems to have interested himself more in the ideological education of a new generation of political activists. It was in this period that he published some of his best known works: *Africa Must Unite*, a general analysis of the problems facing the new African nations; *Consciencism*, a highly theoretical analysis of the ideology of decolonization worked out by his followers; and *Neo-Colonialism; The Last Stage of Imperialism*, an indictment of European and American economic activity in Africa.

Life president of Ghana

As the economic crisis worsened and shortages of foodstuffs and other goods became chronic, Nkrumah explored ways of linking the party with the masses—"Democratic Centralism"—and of representing all major interest groups in the nation within Parliament, thus exemplifying to the end the maxim he had long propounded, "Seek ye first the political kingdom." His considerable international reputation as leading spokesman in the cause of African unity seemed enhanced further when he was invited to Hanoi by President Ho Chi Minh to present proposals for ending the Vietnam War. While he was in Peking, however, the army in Ghana seized power on February 24, 1966. Returning to West Africa, Nkrumah found asylum in Guinea, whose president, Sékou Touré, as a gesture of solidarity declared him a co-head of state. There he immersed himself once more in the cause of black nationalism. His *Handbook of Revolutionary Warfare* (1968) was addressed to the oppressed peoples of Africa, and he also involved himself in the problems of "Black Power" in the United States and the Caribbean. In his last work, *Class Struggle in Africa* (1970), Nkrumah reaffirmed his basic Marxist approach and argued that the African revolutionary struggle "forms part of the world socialist revolution, but must be seen in the context of the Black Revolution as a whole." However posterity may judge Nkrumah's stewardship of Ghanaian affairs between 1957 and 1966, his leadership of the independence movement undoubtedly provided a model of nonviolent political action that changed the course of development of a continent. Nkrumah continued to articulate an ideology of contemporary significance until his death, from cancer, in Bucharest, on April 27, 1972.

Last years in exile

BIBLIOGRAPHY. KWAME NKRUMAH, *The Autobiography of Kwame Nkrumah* (1957); BANKOLE TIMOTHY, *Kwame Nkrumah* (1957), an account of Nkrumah's rise to power by a Sierra Leone journalist; HENRY BRETTON, *The Rise and Fall of Kwame Nkrumah* (1966), an unsympathetic account of the nature of Nkrumah's political machine; ROBERT B. FITCH and MARY OPPENHEIMER, *Ghana: End of an Illusion* (1966), a Marxist evaluation, especially of Nkrumah's economic policies; SAMUEL G. IKOKU, *Le Ghana de Nkrumah* (1971), a valuable account of the CPP by a Nigerian who was one of Nkrumah's leading ideologists; GEOFFREY BING, *Reaping the Whirlwind* (1968), a valuable account of Nkrumah's regime by an Englishman who was one of his principal legal and political advisers; A.A. AFRIFA, *The Ghana Coup* (1966), an account of Nkrumah's policies by one of the leading organizers of the coup. (I.G.W.)

Noble Gases and Their Compounds

The so-called noble gases consist of a group of six chemical elements that exist under ordinary conditions as colourless, odourless, tasteless, nonflammable gases: helium (symbol He, atomic number 2), neon (symbol Ne, atomic number 10), argon (symbol Ar, atomic number 18), krypton (symbol Kr, atomic number 36), xenon (symbol Xe, atomic number 54), and radon (symbol Rn, atomic number 86). They traditionally have been labelled Group 0 in the periodic table (see figure) because for decades after their discovery it was believed that they had a valence of zero; that is, that their atoms could not combine with those of other elements to form chemical compounds. Their electronic structures and the finding that some of them do indeed form compounds suggest that a more appropriate designation would be Group VIIIa.

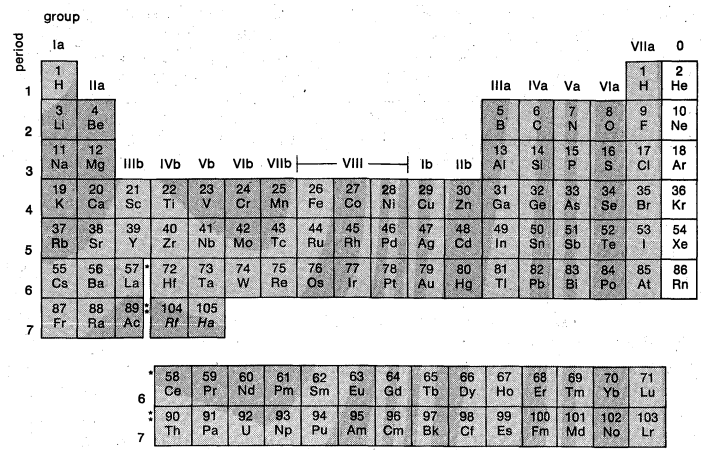
When the members of the group were discovered and identified they were thought to be exceedingly rare, as well as chemically inactive, and therefore were called the rare gases or the inert gases. It is now known, however, that several of these elements are quite abundant on Earth and in the rest of the universe, so the designation rare is misleading. Similarly, use of the term inert has the drawback that it often is applied to gases such as nitrogen and carbon dioxide to connote their nonflammability. In chemistry and alchemy, the word noble long has signified the passivity toward oxygen of a group of metals, such as gold and platinum; it applies in the same sense to the group of gases covered in this article.

THE NOBLE GASES AS A GROUP

Occurrence. The abundances of the noble gases decrease as their atomic numbers increase; helium is, in fact, the most plentiful element in the universe except hydrogen. All the noble gases are present in the Earth's atmosphere and, except for helium and radon, their major commercial source is the air, from which they are obtained by liquefaction and fractional distillation. Most helium is produced commercially from certain natural gas wells. Radon usually is isolated as a product of the radioactive decomposition of dissolved radium compounds (the nuclei of radium atoms spontaneously decay by emitting energy and particles; the particles are the nuclei of helium and radon atoms).

Uses. Several important uses of the noble gases rest on their marked lack of chemical reactivity. Their indifference toward oxygen, for example, confers utter nonflammability upon all six noble gases. Although helium is not quite as buoyant as hydrogen, its incombustibility makes it a safer lifting gas for lighter-than-air aircraft. The noble gases—most often helium and argon, the least expensive—are used to provide chemically unreactive environments for such operations as cutting, welding, and refining of metals (atmospheric oxygen and, in some cases, nitrogen or carbon dioxide would react with the hot metal) and in the handling of other easily attacked materials.

The noble gases absorb and emit electromagnetic radiation in a much less complex way than do other substances. The absorption and emission behaviour is utilized in the employment of these gases in discharge lamps and fluorescent lighting devices: if any of them is confined at low pressure in a glass tube and an electrical discharge is passed through it, the gas glows. Neon produces the familiar orange-red colour of advertising signs; xenon emits a beautiful blue.



Periodic table of elements showing the noble gases.

The very low boiling points and melting points of the noble gases make them useful as refrigerants in the study of matter at extremely low temperatures. The low solubility of helium in fluids leads to its use in admixture with oxygen for breathing by deep-sea divers: because helium does not dissolve in the blood, it does not form bubbles upon decompression (as nitrogen does, leading to the condition known as bends; see COMPRESSION AND DECOMPRESSION INJURIES). Xenon has been used as an anesthetic; although it is costly, it is nonflammable and readily eliminated from the body. Radon is highly radioactive; its only uses have been those that exploit this property, as, for example, in radiotherapy.

The compounds of the noble gases are powerful oxidizing agents (substances that tend to remove electrons from others) and have potential value as reagents in the synthesis of chemical compounds.

History. In 1785 Henry Cavendish, an English chemist and physicist, found that air contains a small proportion (slightly less than 1 percent) of a substance that is chemically less active than nitrogen. A century later Lord Rayleigh, an English physicist, isolated from the air a gas that he thought was pure nitrogen but found that it was denser than nitrogen prepared chemically by liberating it from its compounds. He reasoned that his aerial nitrogen must contain a small amount of a denser gas. Sir William Ramsay, a British chemist, collaborated with Rayleigh in isolating (1894) this gas, which proved to be a new element, argon.

Discovery of argon

After the discovery of argon, and at the instigation of other scientists, Ramsay undertook to investigate the gas evolved upon heating the mineral clevite, which was thought to be a source of argon. The gas proved (1895) instead to be helium, which in 1868 had been detected spectroscopically in the sun but had not been found on Earth. Ramsay and his co-workers searched for related gases and by fractional distillation of liquid air discovered krypton, neon, and xenon, all in 1898. Radon, first identified in 1900, was established as a member of the noble-gas group in 1904.

In 1895 Henri Moissan, a French chemist, failed in an attempt to bring about a reaction between fluorine and argon. In fact, all late-19th- and early-20th-century efforts to prepare chemical compounds of argon failed. The lack of chemical reactivity implied by these failures was of significance in the development of theories of atomic structure. In 1913 a Danish physicist, Niels Bohr, proposed that the electrons in atoms are arranged in successive shells having characteristic energies and capacities and that the capacities of the shells determine the numbers of elements in the periods of the periodic table. On the basis of experimental evidence relating chemical properties to electron distributions, it was suggested that in the atoms of each noble gas beyond helium the electrons are arranged in these shells in such a way that the outermost shell always contains eight electrons, no matter how many others (in the case of radon, 78 others) were arranged inside those of eight. (Further details are discussed in the article ATOMIC STRUCTURE.)

In a theory of chemical bonding advanced by others in 1916, this octet (as it came to be called) of electrons was taken to be the most stable arrangement for the outermost shell of any atom. Although only the noble-gas atoms possessed this arrangement, it was the condition toward which the atoms of all other elements tended in their chemical bonding. Certain elements satisfied this tendency by either gaining or losing electrons outright, becoming electrically charged particles (ions); other elements shared electrons, forming stable combinations linked together by covalent bonds. The valences of elements—that is, the proportions in which their atoms combined to form ionic or covalent compounds—were thus controlled by the behaviour of their outermost electrons, which for this reason were called the valence electrons. This theory rationalized the chemical bonding of the reactive elements as well as the relative inactivity of the noble gases, which is their chief characteristic. (Additional information appears in the article CHEMICAL BONDING.)

Octet theory

The outer electrons of the atoms of the heavier noble gases, screened from the nucleus by intervening electrons, are held less firmly and can be removed more easily from the atoms than can the electrons of the lighter noble gases. This fact had been known for a long time from experiments using electrical and magnetic fields; the energy required for removal of one electron is called the first ionization potential. In 1962 it was discovered that platinum hexafluoride would oxidize molecular oxygen to form a salt. Knowledge that the first ionization potential of xenon is very close to that of oxygen led to the suggestion that a salt of xenon might be formed similarly. In the same year it was established that it is indeed possible to remove electrons from xenon by chemical means—that is, to oxidize xenon—when two teams of chemists independently prepared fluorides of that element. This achievement was followed by the preparation of other xenon compounds and of the fluorides of radon (1962) and krypton (1963).

Properties of the gases. Each noble-gas element is situated in the periodic table between an element of the most electronegative group, the halogens (Group VIIa, the atoms of which add electrons to achieve the octet and thereby become negative ions), and an element of the most electropositive group, the alkali metals (Group Ia, the atoms of which lose electrons to become positive ions). The noble gases thus form a dramatic transition group in the periodic table of the elements and are neither electropositive nor electronegative but, relatively speaking, neutral, neither gaining nor losing electrons easily.

The sizes of the atoms of noble gases increase smoothly with the increase in atomic number, partially because repulsion between the electrons prevents them from occupying the same region of space and forces each successive shell to extend into a larger concentric spherical volume. In the largest atoms, the weaker attraction of the

Some Properties of the Noble Gases

	helium	neon	argon	krypton	xenon	radon
Atomic number	2	10	18	36	54	86
Atomic weight	4.003	20.182	39.948	83.797	131.295	(222)*
Melting point						
(°C)	−272.15†	−248.67	−189.2	−156.6	−111.9	−71
(°K)	1	24.48	83.95	116.55	161.25	202
Boiling point						
(°C)	−268.94	−246.05	−185.88	−152.3	−107.10	−61.8
(°K)	4.21	27.10	87.27	120.8	166.05	211.07
Density at 0° C, 1 atm (g/l)	0.17847	0.89994	1.78403	3.733	5.8811	9.73
Solubility in water at 20° C (cu cm of gas per 1,000 g water)	8.61	10.5	33.6	59.4	108.1	230
Electronic configuration	$1s^2$	$1s^22s^22p^6$	$(Ne)3s^23p^6$	$(Ar)3d^{10}4s^24p^6$	$(Kr)4d^{10}5s^25p^6$	$(Xe)4f^{14}5d^{10}6s^26p^6$
Isotopic abundance (terrestrial, percent)	^{3}He (0.00013) ^{4}He (100)	^{20}Ne (90.92) ^{21}Ne (0.257) ^{22}Ne (8.82)	^{36}Ar (0.337) ^{38}Ar (0.063) ^{40}Ar (99.600)	^{78}Kr (0.354) ^{80}Kr (2.27) ^{82}Kr (11.56) ^{83}Kr (11.55) ^{84}Kr (56.90) ^{86}Kr (17.37)	^{124}Xe (0.096) ^{126}Xe (0.090) ^{128}Xe (1.912) ^{129}Xe (26.44) ^{130}Xe (4.08) ^{131}Xe (21.18) ^{132}Xe (26.89) ^{134}Xe (10.44) ^{136}Xe (8.87)	^{219}Rn (trace) ^{220}Rn (trace) ^{222}Rn (trace)
Radioactive isotopes (mass numbers)	6	17–19, 23, 24	33, 35, 37, 39, 41, 42	74–77, 79, 81, 85, 87–94	118–123, 125, 127, 133, 135, 137–142	204–213 215–224
Colour of light emitted by gaseous discharge tube	yellow	red	red or blue	yellow green	blue to green	—
Heat of fusion (cal/mole)	5 (3.5° K, 100 atm)	80.1	280.8	390.7	548.5	—
Heat of vaporization (cal/mole)	19.4	414	1,557.5	2,158	3,020	4,325
Specific heat (cal/mole/°C)	4.9680	4.9680	4.9680	4.9680	4.9680	4.9680
Critical temperature (°K)	5.25	44.5	150.85	209.35	289.74	378.15
Critical pressure (atm)	2.26	26.9	48.3	54.3	57.64	62
Critical density (g/cm^3)	0.0693	0.484	0.536	0.908	1.100	—
Thermal conductivity at 0° C, 1 atm (cal/cm-sec-°C)	33.90×10^{-5}	11.00×10^{-5}	3.920×10^{-5}	2.09×10^{-5}	1.21×10^{-5}	—
Magnetic susceptibility (cgs units/mole)	−0.0000019	−0.0000072	−0.0000194	−0.000028	−0.000043	—
Crystal structure‡	hcp	fcc	fcc	fcc	fcc	fcc
Radius						
Atomic (Å)	1.3	1.6	1.92	1.98	2.18	—
Covalent (crystal) estimated (Å)	0.4–0.6	0.7	0.95	1.10	1.30	1.5–2.1
Static polarizability (Å^3)	0.204	0.392	1.63	2.465	4.01	—
Ionization potential (first, eV)	24.586	21.563	15.759	13.999	12.129	10.747
Electronegativity (Pauling)	4.5	4.0	2.9	2.6	2.25	2.00

*Stablest isotope. †At 25.05 atm. ‡hcp = hexagonal close-packed, fcc = face centred cubic (cubic close-packed).

nucleus for the electrons in the most distant shell explains the greater ease with which the outer electron clouds of these atoms are distorted by other charged bodies, or polarized. A polarized atom as a whole remains electrically neutral but the distribution of charge within it becomes unsymmetrical, the centre of gravity, so to speak, of the negative charge being displaced from that of the positive charge. The polarizability of the noble gases increases markedly through the group from helium to radon as the ionization potential decreases.

Generally speaking, a polarizable substance is attracted more strongly to surfaces and dissolves in fluids more extensively than a nonpolarizable one. Because the polarizability of helium is minimal, it shows little or no tendency to adsorb upon surfaces and it is not very soluble in fluids. These properties may account for the fact that helium does not induce narcosis when it is breathed by divers at high pressures and that it does not form clathrates. (Clathrates—from a Latin word meaning "enclosed by a lattice"—are substances in which the molecules of one compound, called the host, form a cagelike crystalline lattice within which there are open spaces that may be occupied by molecules of a second compound or element, called the guest, although no chemical bonds are formed between the host and the guest. The stability of a clathrate is influenced by the closeness with which the guest particles fit into the holes as well as by the polarizability of the guest.) The properties of small size and low polarizability make helium a highly mobile gas; *i.e.*, one that is not trapped by various structures of other atoms. The greater polarizability of argon and the heavier noble gases accounts for their formation of clathrates, including those in which water is the host—hydrates—and also accounts for the ready adsorption of these gases on surfaces. The narcotic activity of the noble gases also parallels their polarizability. The noble gases reduce the sensitivity of living cells to attack by oxygen under the influence of radiation; the heavier gases are more effective than the lighter ones in this action, perhaps simply as a result of their greater solubility in fluids, but possibly also as a consequence of their greater ability to absorb neutrons.

Clathrates

THE INDIVIDUAL NOBLE GASES

Helium. Because the first evidence of the existence of helium was the presence of certain wavelengths in the yellow region of the spectrum of the light emitted from the sun, the name of the element was derived from the Greek word *hēlios*, "sun." Helium constitutes about 23 percent of the mass of the universe, but only about eight parts per 1,000,000,000 of the Earth's crust; ordinary air contains about five parts per 1,000,000 of helium.

The nucleus of every helium atom contains two protons (that is, the atomic number of helium is two) but, as is the case with all elements, isotopes of helium exist. (Isotopes are different forms of the same element that vary in the number of neutrons present in the nuclei of their atoms.) The known isotopes of helium contain from one to six neutrons, so their mass numbers (the sums of the numbers of protons and neutrons) range from three to eight. Of these six isotopes, only the two having mass numbers of three (helium-3, symbolized ^{3}He) and four (helium-4, ^{4}He) are stable; all the others are radioactive, decaying very rapidly into other substances. Helium-4 is by far the more plentiful of the stable isotopes: helium-4 atoms outnumber those of helium-3 about 700,000 to one in atmospheric helium and about 7,000,000 to one in certain helium-bearing minerals.

Most of the helium of commerce is derived from natural gas produced in the southwestern United States. The helium, which comprises between 1.5 and 7 percent of the gas, is separated by continuous processes involving several steps. A portion of the helium market is supplied from plants that liquefy air on a large scale; the

amount of helium obtainable from 1,000 tons of air is about 1.3 pounds (about 112 cubic feet, measured at room temperature and atmospheric pressure).

Liquid and solid helium

The boiling and freezing points of helium are lower than those of any other known substance. Helium is the only element that cannot be solidified by sufficient cooling at normal atmospheric pressure; it is necessary to apply pressure of 25 atmospheres at a temperature of 1° K (−272.15° C, −457.9° F) to convert it to its solid form.

The isotope helium-4 is unique in having two liquid forms. The form designated I exists at temperatures from its boiling point of 4.21° K down to 2.18° K (−270.97° C, −455.75° F); form II exists at still lower temperatures. Liquid helium II exhibits the property called superfluidity: its viscosity, or resistance to flow, is so low that it has not been measured. This liquid spreads in a thin film over the surface of any substance it touches, and this film flows without friction even against the force of gravity.

A liquid mixture of the two isotopes helium-3 and helium-4 separates at temperatures below about 0.8° K into two layers. One layer is practically pure helium-3; the other is mostly helium-4, but retains about 6 percent helium-3 even at the lowest temperatures achieved. Dissolving helium-3 in helium-4 is accompanied by a cooling effect that has been utilized in the construction of cryostats (devices for production of very low temperatures) that can attain—and maintain for periods of days—temperatures as low as 0.01° K.

Neon. Neon, the name of which is derived from the Greek word *neos*, "new," is present in some minerals, but its only commercial source is the atmosphere, of which it comprises 18 parts per 1,000,000 by volume. Because its boiling point is −246° C (−411° F), neon remains, along with helium and hydrogen, in the small fraction of air that resists liquefaction upon cooling to −196° C (−321° F). Neon is isolated from this cold, gaseous mixture by bringing it into contact with activated charcoal, which adsorbs the neon and hydrogen; removal of hydrogen is effected by adding enough oxygen to convert it all to water, which, along with any surplus oxygen, condenses upon cooling.

Neon was the first element shown to consist of more than one stable isotope. In 1913 application of the technique of mass spectrometry (*q.v.*) revealed the existence of neon-20 and neon-22, which comprise 90.92 and 8.82 percent, respectively, of the naturally occurring mixture. The third stable isotope, neon-21, which makes up 0.26 percent of natural neon, was detected later. Five radioactive isotopes of neon also have been identified.

Argon. The first of the noble gases to be discovered, argon was named from the Greek word *argos*, "lazy," because of its chemical inertness. In cosmic abundance, argon ranks approximately 12th among the 100-odd chemical elements; although the stable isotopes argon-36 and argon-38 make up all but a trace of this element in the universe, the third stable isotope, argon-40, comprises 99.60 percent of the argon found on Earth. The terrestrial preponderance of argon-40 presumably arose from the formation of this isotope by the radioactive decay of potassium-40.

Argon is the most plentiful of the noble gases on Earth, comprising 0.934 percent by volume, or 1.288 percent by weight, of the atmosphere. The element is obtained from air by liquefaction and fractional distillation; although the boiling points of argon, oxygen, and nitrogen all lie within a few degrees of each other, efficient processing provides each gas in a purity of more than 99.9 percent.

Krypton. Krypton is named from the Greek word *kryptos*, "hidden." Traces of krypton are present in minerals and meteorites, but the usual commercial source is the atmosphere, which contains 1.14 parts per 10^6 by volume. Krypton also is formed by the nuclear fission of uranium triggered by slow neutrons: this source may be expected to become increasingly important because of the growing number of fission-power plants. Krypton has isotopes of every mass number from 74 through 95; six, with mass numbers 78, 80, 82, 83, 84, and 86, are stable. After it has been stored a few days, krypton obtained by nuclear fission contains only one radioactive isotope, krypton-85, which has a half-life of about 10 years, because all the other radioactive isotopes have half-lives of three hours or less. (The half-life is the length of time during which one-half of any original amount of an unstable substance decays.)

Because its boiling point is about 30° C higher than those of the major constituents of air, krypton is readily separated from liquid air by fractional distillation; it accumulates along with xenon in the least volatile portion. These two gases are further purified by adsorption onto silica gel, redistillation, and passage over hot titanium metal, which removes all impurities except other noble gases.

Krypton is the lightest of the noble gases that has been converted into chemical compounds (see below *Compounds of the individual noble gases: Krypton*).

Xenon. The name xenon is derived from the Greek word *xenos*, "strange" or "foreign." Like several other noble gases, xenon is present in meteorites and certain minerals, but its only useful source has been the atmosphere, of which it composes 86 parts per 10^9 by volume. Nuclear reactors may become an important source of xenon, because the fission of uranium produces several isotopes of xenon.

The mass numbers of the known isotopes of xenon range from 118 to 144; nine of these numbers correspond to stable isotopes. The xenon isotopes produced in the greatest amount by nuclear fission are xenon-131, -132, -134, and -136, which are stable, and xenon-133, which is radioactive, having a half-life of 9.2 hours.

Xenon is the least volatile of the noble gases obtainable from the air. Its purification has been mentioned above (see *Krypton*). Numerous compounds of xenon have been prepared since the discovery of the noble-gas compounds in 1962 (see below *Compounds of the individual noble gases: Xenon*). Although xenon itself is an unusually safe anesthetic, its compounds appear to be toxic.

Radon. Radon was originally called radium emanation: it is the radioactive gas formed, along with helium, as radium decays. The term radon sometimes has been restricted to the isotope of mass 222, other isotopes being referred to as thoron (symbolized Tn, now called radon-220), formed from thorium; and actinon (An, now radon-219), formed from actinium. The name emanation, symbolized Em, has been used to denote the entire family now recognized as radon isotopes. Radon has no stable isotope; radioactive isotopes having masses ranging from 204 through 224 have been identified, the longest-lived of these being radon-222, which has a half-life of 3.82 days. All the radon isotopes undergo radioactive decay, the stable end-products being helium and isotopes of heavy metals, in most cases lead.

COMPOUNDS OF THE NOBLE GASES

General considerations. In the combination of atoms to form chemical compounds, the atoms become bound together by forces arising from changes in the arrangements of their outermost electrons. In simplified terms, the outright transfer of electrons from one atom to another creates oppositely charged particles (ions) that attract one another to form ionic, or electrovalent, compounds. If, instead, the electrons are shared by the atoms, the bond is called covalent, one covalent bond involving a pair of electrons. In actuality, there exists a practically continuous gradation between these two extreme types, because even in compounds called ionic, certain aspects of their behaviour suggest a small degree of electron sharing, and in covalent compounds, inequality in the electron affinities of the bonded atoms causes the electron distribution to be more or less unsymmetrical, giving the effect of partial electrical charges on the bonded atoms.

Ionic and covalent compounds

The first known compounds of the noble gases were discovered in 1962 and, although none of them yet has found commercial application, they have attracted much

attention as representatives of a class of substances previously thought incapable of existence. In much of their chemical behaviour the noble gases closely resemble their neighbours in the periodic table, the halogens, but because the gain or loss of electrons would disrupt the stable configuration of the outermost shell of a noble-gas atom, these elements conform most closely to covalent, rather than ionic, bonding patterns.

By the early 1970s only the heavier noble-gas elements krypton, xenon, and radon were known to form any compounds at all. Even in these compounds, only the most electronegative atoms or groups of atoms are capable of bonding to the noble-gas atoms. These binding units, often called ligands, include fluorine, oxygen, and combinations of oxygen with sulfur, fluorine, tellurium, or chlorine. In their compounds, the noble-gas atoms display a variety of bonding arrangements depending on the number of valence electrons that become involved in the bond-forming process. These numbers of electrons often are referred to as oxidation numbers and are specified by roman numerals in the names of the compounds; thus, in xenon(VI) fluoride (xenon hexafluoride) six electrons of the xenon atom participate in bond formation. The established oxidation states of the noble-gas elements harmonize with the pattern shown by other nontransition elements; *i.e.*, those in the a-groups (see figure, Groups Ia, IIa, etc.). According to this pattern, any element prefers those oxidation states that are numerically equal to the number of its periodic group or that number less two, four, or six. Assignment of VIIIa as the group number of the noble gases is amply justified by the existence of compounds in which xenon exhibits oxidation states of +8 (*e.g.*, xenon tetroxide), +6 (xenon hexafluoride), +4 (xenon tetrafluoride), and +2 (xenon difluoride).

Oxidation numbers

Each of the five elements preceding xenon in the fifth period of the table forms a fluoride in which the element has its maximum number: indium binds three fluorine atoms; tin, four; antimony, five; tellurium, six; and iodine, seven. It might be anticipated, therefore, that xenon octafluoride could exist; actually, this compound has not been prepared, but the failure is thought to be due to the difficulty of arranging eight fluorine atoms within the necessary distance of the xenon atom, rather than to an incapability of engaging eight electrons in chemical bonding. In the fourth period of the table, ending with krypton, the fluorides of gallium, germanium, arsenic, selenium, and bromine create a pattern that indicates, correctly, that the compounds of krypton are likely to be less stable than those of xenon.

Stability of noble-gas compounds. The stabilities of the noble-gas compounds are best expressed in terms of the energies required to break the bonds present, forming either ions or neutral atoms. Generally, across each period of the periodic table, the mean bond energy (the amount of energy needed to break all the bonds divided by the number of bonds) of fluorides decreases from Group IVa to Group VIIIa; *i.e.*, the fluorides of the noble gases are the least stable. Thus, for the xenon period, the bond energies per ligand decrease from tin to iodine, indicating a relatively low value for the bond energies in xenon octafluoride, a compound that has resisted all attempts to prepare it. A similar extrapolation in the same series for the +6 oxidation state, however, predicts that xenon hexafluoride should be a stable compound, as it is. In the krypton period, the figures correctly predict low bond energies for all the krypton fluorides. Similarly, it is evident that the bond energies of argon fluorides will be less than those of krypton fluorides. It seems unlikely, therefore, that argon compounds can be prepared.

The noble-gas fluorides are oxidizing agents of extraordinary power. This property is consistent with the value of the enthalpy of atomization (a measure of the energy change in a chemical reaction that breaks up a molecule into its component atoms) of the krypton difluoride molecule, which is only two-thirds that of the diatomic fluorine molecule itself (fluorine being the strongest oxidizing agent of all the elements).

The noble-gas compounds (those of xenon, in particular) may eventually prove to be useful reagents, because the by-product of their reactions is always a noble gas, a relatively inert substance. The noble-gas fluorides are, therefore, much less likely to yield unwanted products than are the halogen fluorides long employed in the preparation of metal fluorides and in the fluorination of aromatic compounds, because the halogens liberated in the use of the halogen fluorides are themselves very active and therefore apt to bring about undesired reactions. Xenon difluoride already has proved to be useful in fluorinating aromatic hydrocarbons, and it is a powerful oxidizer in aqueous alkali.

Noble-gas compounds as reagents

Bond structures in noble-gas compounds. The close relationship of noble-gas compounds to compounds of the elements of Groups VIIa and VIa, as indicated by the bond-energy data, also is shown by the structural similarities of the molecules.

Because each of the noble-gas cations has the name electron arrangement as its neighbouring halogen atom, it is to be expected that the cations consisting of single atoms of fluorine and noble gas—for example, the xenon monofluoride cation (formula XeF^+) and the krypton monofluoride cation (KrF^+)—would each show a close relationship to the corresponding halogen monofluoride molecule. This is so for iodine monofluoride (IF) and the xenon monofluoride ion, as shown by various measurements made of the structures.

The chemical bonding should be similar in these related species, and it is appropriate to represent the bonding in terms of a classical electron-pair bond, each element achieving an octet configuration. The bonding of two fluorine atoms to a neutral noble-gas atom, as in the generation of compounds of xenon and krypton, should be very similar to the bonding of two fluorine atoms to a halogen monofluoride.

Theories of bonding in noble-gas compounds. Since 1913, when Bohr proposed the existence in atoms of shells occupied by particular numbers of electrons, considerable advances have been made in understanding how energy states of atoms are related to the distribution of electrons in these shells and how the electronic structures of atoms account for the patterns expressed by the periodic table of the chemical elements. It has been impossible to provide exact mathematical descriptions of these energy states—because there is no way to evaluate all the interactions among the electrons and between them and the nucleus—but methods have been developed that make it possible to estimate, with a fair degree of accuracy, many of the observable properties of atoms, such as their absorption of light.

The union of two or more atoms to form a stable molecule is possible only because the energy of the combination is less than the sum of the energies of the separated atoms. Calculation of energies of molecules poses an even more formidable problem than does the calculation of atomic energies; comparison of estimates with observed properties also is more difficult, because the behaviour of molecules—for example, their absorption of light—is much more complex. Two widely used methods of estimating molecular energies, called the molecular-orbital theory and the valence-bond theory, may be regarded as attempts to relate the electronic energy levels of molecules to those of the atoms making them up. The two techniques differ in detail, and although they yield similar results when applied to certain molecules, they lead to divergent predictions in other cases.

Calculation of molecular energies

One of the theories embodies the viewpoint that covalent bond formation represents the presence of electrons in new energy states common to both bonded atoms. Bonding of atoms that do not have electrons in such energy levels in their ground states requires excitation, or promotion, of valence electrons into suitable states. For the molecule to be stable, the energy decrease resulting from bonding must more than offset the excitation energy; in the case of noble-gas compounds, a controversy has arisen over the validity of this approach because the excitation energies appear to be too large to be balanced by bonding.

Compounds of the individual noble gases. *Helium.* The great difficulty of removing an electron from an atom of helium to form a helium ion is a clear indication of the chemical inertness of the element. In the gaseous state, however, ions represented by the formula HeR^+ (*R* being another atom or group) are well known. The helium atom has an affinity for the hydrogen ion, or proton (H^+), and the chemical bonding in the positive ion formed (HeH^+) is presumably similar to that between hydrogen atoms in molecular hydrogen (H_2). It is unlikely, however, that this ion or its relatives can be stabilized enough to exist in the solid state.

Neon. As would be expected from the high ionization potential of neon, compounds of the element are unknown. Ions, such as Ne_2^+ and NeH^+, may be generated in electrical discharges, but salts incorporating these ions have not been prepared.

Argon. The outer (valence) electrons of argon are tightly bound, and it has not been possible to prepare chemical compounds of this element. The argon fluoride ion (ArF^+) has the same electron arrangement as the compound chlorine fluoride (ClF); its electron affinity, however, may be so high that it will be hard to find a suitable ion partner for salt formation. Numerous gaseous positive ions of argon have been observed by mass spectroscopy.

A number of clathrates of argon are known. One, a hydrate, is composed of eight argon atoms and 46 water molecules; clathrates with organic compounds include several formulated $Or{\cdot}2Ar{\cdot}17H_2O$, in which the symbol *Or* represents acetone, methylene chloride, chloroform, or carbon tetrachloride.

Krypton. In its known chemical behaviour, krypton exhibits only the oxidation state +2, which appears in the compound krypton difluoride (KrF_2) and its derivatives. Because of the ease with which it gives up fluorine atoms, krypton difluoride is the most powerful oxidative fluorinator known. It is a colourless solid, decomposing at room temperature; the compound interacts with water to produce krypton gas, oxygen, and hydrogen fluoride. Krypton clathrates include a hydrate consisting of eight krypton atoms and 46 water molecules.

Xenon. Compounds have been prepared in which xenon has the oxidation states +2, +4, +6, and +8; these substances represent the greatest range of oxidation states and the largest variety of compounds known for any of the noble gases.

The xenon atoms with oxidation state +2 forms the stable difluoride and its derivatives, as well as an unstable dichloride. Xenon monoxide has not yet been prepared.

The only compound of xenon with oxidation state +4 that is available in appreciable quantities is xenon tetrafluoride; xenon tetrachloride is stable only at extremely low temperatures. The fluorine ligands of the tetrafluoride may be successively replaced by other highly electronegative ligands, leading to a series of compounds with the general formula $XeF_{4-n}L_n$, in which *L* represents the ligand and *n* is an integer from one to four. The ready spontaneous conversion of xenon(IV) compounds, in part to xenon(VI) compounds and in part to the free element, a reaction known as disproportionation, parallels the more familiar example of the iodine-containing ion, hypoiodite.

The xenon atom with an oxidation state of +6 is present in xenon hexafluoride, xenon trioxide, and several derivatives of these compounds. Interaction of xenon hexafluoride with oxides produces, initially, xenon oxide tetrafluoride. Xenon dioxide difluoride may be prepared from the interaction of xenon trioxide and xenon hexafluoride. Xenon oxide tetrafluoride and xenon dioxide difluoride are colourless, somewhat volatile solids. Xenon trioxide, also colourless, is readily soluble in water and is a powerful explosive. Successive replacement of the fluorine atoms in xenon hexafluoride by other ligands leads to a series of compounds formulated $XeF_{6-n}L_n$. The pair of nonbonding valence electrons in these Xe(VI) compounds evidently occupies an appreciable volume of space, because the shape of the molecule of xenon hexafluoride is not that of a regular octahedron, as it would be expected to be if the six fluorine atoms were packed as closely as possible around the xenon atom. The molecular arrangements of other Xe(VI) compounds also indicate the spatial requirements of the nonbonding electron pair.

Xenon octafluoride is unknown, as mentioned previously (see above *Compounds of the individual noble gases: General considerations*). In strongly alkaline solutions, xenon hexafluoride disproportionates into gaseous xenon and xenon in the +8 oxidation state in the form of the perxenate, or xenate (VIII) ion XeO_6^{4-}. In this ion, the six oxygen atoms are located at the vertexes of a regular octahedron with the xenon atom at the centre, in contrast to the distorted arrangement of the atoms of xenon hexafluoride. The perxenate ion is present in several solid compounds, combined with metal ions such as sodium or barium; these compounds are salts of perxenic acid, H_4XeO_6, which is a relatively weak acid. The perxenate ion oxidizes water to oxygen, itself undergoing reduction to the xenate(VI) ion, $HXeO_4^-$; this process occurs slowly in alkaline solutions but almost instantaneously in acids. Despite this instability in solution, the crystalline perxenates possess considerable thermal stability, the anhydrous sodium salt being unaffected by heating to temperatures up to 360° C.

Perxenate ion

Xenon tetroxide may be prepared from anhydrous perxenates; great care must be taken in the preparation because the tetroxide is even more unstable than the explosive trioxide.

Radon. Radon interacts spontaneously with fluorine and with any of the interhalogen fluorides except iodine pentafluoride to form a colourless fluoride that is considered to be (by analogy with the similar behaviour of xenon difluoride) radon difluoride. Radon fluoride may be an ionic compound; it decomposes without vaporizing at temperatures above 250° C. No complex compounds, oxides, or chlorides of radon have yet been reported, probably because the high radioactivity and short life of radon make research upon it very difficult. Several clathrates of radon, including a hydrate, have been described.

BIBLIOGRAPHY. G.A. COOK (ed.), *Argon, Helium and the Rare Gases*, 2 vol. (1961), authoritative accounts of the history, occurrence, properties, production, analytical determination, and uses of the noble gases (published one year before the discovery of noble-gas compounds); M.W. TRAVERS, *Life of Sir William Ramsay* (1956), an authoritative account by Ramsay's principal co-worker on the discovery of the noble gases and the early controversy surrounding the argon discovery; H.H. HYMAN (ed.), *Noble-Gas Compounds* (1963), an account of the work done on noble-gas chemistry prior to April 1963, the feverish activity in the field during the previous year, and the discovery of argon; N. BARTLETT and F.O. SLADKY, "The Chemistry of Krypton, Xenon and Radon," in A.F. TROTMAN-DICKENSON (ed.), *Comprehensive Inorganic Chemistry* (1971).

Nordrhein-Westfalen

Nordrhein-Westfalen (North Rhine-Westphalia) is a state, or *Land,* of the Federal Republic of Germany and plays a leading role in the national economy. The Rhine–Ruhr area—the nation's most important industrial area—runs through the centre of the state and dominates it economically and culturally. It is one of the most heavily populated areas of the world, and its rich natural resources support a wide variety of burgeoning industries. Outside of this complex, the state exhibits a contrasting character of rural charm grounded in the German past. It covers an area of 13,144 square miles (34,044 square kilometres) and is located in the west central section of the country. The state is bordered by The Netherlands and Belgium on the west and the states of Niedersachsen (Lower Saxony) on the north, Hessen on the east, and Rheinland-Pfalz on the south. The state capital is at Düsseldorf, and Bonn is the national capital. With a population of nearly 17,000,000 inhabitants, it is the most populous state of West Germany, containing almost 30 percent of the national total.

The landscape. *Relief.* The state includes the upland regions of North Eifel in the south and the mountains of

the Sauerland in the southeast. Their plateaus—cut sharply by valleys—cover layers of schist, graywacke (dark gray rock composed of fine particles), and sandstone. The areas composed of limestone are pocked with caves and crevices, and volcanic rock occurs in the region of the Siebengebirge (the Seven Hills) on the eastern bank of the Rhine River. In the east, the Weserbergland—the mountainous region bordering the Weser River—is composed of layers of limestone, clay, and sandstone and is characterized by several escarpments and by the narrow, elongated ridges of the Wiehengebirge (Wiehen Mountains), the Teutoburger Wald, and the Eggegebirge (Egge Mountains).

Mountains and lowlands

The northwest is composed of lowlands that gradually merge with the upland regions. The region of the lower Rhine consists of terraced surfaces of gravel, sand, and loam and some hills formed by glacial activity. In the south, the lowlands extend in a baylike formation covered with loess (a brownish mixture of fine particles deposited by the wind) to the vicinity of Bonn, and, in the east, the Westphalian extension, or bay, with bowl-shaped layers of limestone, marl, and sand, partly covered by glacial formations, pushes forward between the Weserbergland and the Sauerland. Fertile areas of loess are also found along the southern border of the Westphalian bay around the Hellweg.

Drainage. The Rhine River drains the largest physical region of the state. Areas bordering on the west are drained by the Maas River, on the north by the Ems, and on the northeast by the Weser. Ultimately, the drainage system leads to the North Sea.

Climate. The climate of the area is influenced by the proximity of the North Sea to the Gulf Stream. The lowland zones are mild in the winter, with mean January temperatures of about 34° F (1° C), while July temperatures average about 63° F (17° C). Precipitation is often less than 30 inches in the Rhine Valley. The mountainous regions, however, are cool and receive high precipitation. On Kahler Asten Berg, in the Sauerland, the average winter temperature does not rise above 32° F (0° C), and the average temperature in July barely reaches 55° F (13° C). The annual precipitation exceeds 55 inches, and for more than 110 days of the year Kahler Asten Berg is covered with snow.

Vegetation and animal life. Forests predominate in the higher mountain regions, where the ancient beechwoods have been replaced by pines. In the lowlands, large forests are generally found only in barren, sandy areas. In parts of the Westphalian bay extensive moorlands are maintained as small nature reserves or protected areas. Red deer and wild boars are found in the larger forested regions, and roe deer are widespread. Wild horses still live in the Westphalian bay. The largest bodies of water —Altwässer Reservoir and Staubecken (a pond)— are renowned for their numerous types of birds, including thousands of migratory or hibernating ducks and geese.

Settlement patterns. The settlement pattern of the state reflects the influence of both physical features and historical development. The principalities that existed until about 1800 and played an important role in commercial development are recalled by the regional names of Bergisches Land (the Duchy of Berg), Märkisches and Kurkölnisches Sauerland (the Mark and the Cologne Archbishopric-Electorate of Sauerland), and Münsterland (the princely bishopric of Münster). In the former religious territories—especially in the Kurkölnisches Sauerland and in the princely bishoprics of Münster and Paderborn—economic development was slow and quiet, and large areas still possess a predominantly rural character. In Westphalia, many of the great farmhouses have been preserved, with their large gates at the gabled front and the stables along both sides of the long halls. In the Münsterland, the widely scattered or loosely grouped farms are surrounded by oak groves, forest plots, and hedges, giving the impression of a landscaped park. Along the lower Rhine, the architectural form known as the T-house is found, the barn area is opened from the gable front, and the dwelling area is erected across like a crossbar. The Frankish farm—with buildings grouped at right angles around a courtyard—predominates in the southern section of the Rhine lowlands.

Historic development

Industrial development, particularly in the case of the iron and metal industries, centred in the mountains, as it was dependent upon the existence of ore, the production of charcoal, and the supply of water power from rivers and brooks. The large, industrially concentrated region of the Rhine–Ruhr between Bonn and Hamm contains many large cities and metropolitan areas. It is one of the most densely populated regions of the world, with a population exceeding 11,000,000 inhabitants by the early 1970s. In addition, there are smaller urban areas around Aachen, Münster, Bielefeld, and Siegen.

The people. *Cultural identities.* The Rhineland–Westfalen boundary—running through the state from northwest to southeast—corresponds to the old border between the Saxons and the Franks and is reflected in certain dialectical variations. The language differences involve the conjugation of verbs and the shifting of sounds in many words.

During the Reformation, the principalities in the Nordrhein-Westfalen area were divided between Roman Catholicism and Protestantism. These differences still exist, and the regional predominance of one religion over the other still exists in many areas. More than half of the total state population is Roman Catholic, but local dominance of either religion often claims more than 75 percent of a region's inhabitants.

These cultural differences are strongest in the rural areas. In the urban centres, social displacement and the mixture of peoples has dissipated the strength of original cultural identifications. Since 1870 the heavy immigration of East Germans and Poles—as well as Italians, Turks, Greeks, Yugoslavs, and the Dutch—has created a more cosmopolitan atmosphere in the cities. Most of the cultural mingling has occurred in the Rhine–Ruhr area where almost two dozen major cities are located (*e.g.*, Cologne, Essen, Düsseldorf, Dortmund, and Duisburg).

Demography. The average population density of the state is about 1,300 persons per square mile. Densities are highest in the urban Ruhr region, where they reach over 9,000 inhabitants per square mile. In the agrarian areas, however, densities range between 400 and 650 persons per square mile, while the heavily forested mountains support only 200 to 250 persons per square mile. Immediately after World War II, population increase was largely the result of immigration, but this trend decreased after 1960. The excess of births of less than one percent during the 1960s has also decreased.

The economy. *Agriculture and forestry.* Almost two-thirds of the state's total land area are utilized for commerical farms, gardens, or orchards. The most important fertile regions are located in the southern lowlands, where wheat and sugar beets are grown. In the vicinity of the Rhine near Bonn, Cologne, and Düsseldorf, fruits and vegetables are cultivated, while in Münsterland and in the regions of the lower Rhine cattle raising and pig breeding play considerable roles. The most important regions engaged in forestry are the montane spruce-forest and the lowland pine-forest areas.

The total forested area of 3,100 square miles represents 24 percent of the state's total land area. Agriculture and forestry together earn about 3 percent of the gross domestic product.

Mining and industry. The state is the most significant mining and energy-producing area of West Germany. Bituminous coal deposits are located in the regions of the Ruhr and Aachen, and lignite is mined west of Cologne. In the late 1960s, 100,000,000 tons of bituminous coal and about 90,000,000 tons of lignite were mined annually, representing almost 90 percent of national production. Petroleum refineries—connected with the North Sea ports of Wilhelmshaven and Rotterdam by a system of pipelines—are concentrated in the region of the Ruhr and on the Rhine. The water supply necessary to heavy industry and urban concentration is maintained by almost 60 dams, located mainly in the mountains of the Sauerland, the Bergisches Land, and in the north Eifel region.

Coal, oil, and water

Industrial, mining, and energy production comprise

over half of the state's gross domestic product. On the basis of coke production, blast furnaces, steel mills, and rolling mills have developed in the Ruhr region, mainly in Duisburg and Dortmund. Raw steel production exceeds 30,000,000 tons annually, amounting to more than 70 percent of national production. In the territories of the Berg-Mark bordering on the south, the iron and metalware industries and textiles play a leading role. The textile industry is also extensive around Krefeld, Mönchengladbach, and in northwestern Münsterland. Leverkusen, Marl, and a few locations along the Rhine are noted for their chemical industry, and the area of Lippe is well-known for its wood and furniture industries. The areas around Düsseldorf, Cologne, Aachen, and Bielefeld have a highly diversified industrial structure.

Services. The service industry produces about 45 percent of the state's gross domestic product. There are many commercial enterprises, trading houses, loan societies, and banking houses. The system of municipal savings banks is extended throughout the state.

Transportation. The close-meshed transportation network meets the heavy demands of urban concentration and industrialization. In the densely settled areas of the Rhine and Ruhr rivers, a streetcar system is maintained for passenger traffic, and the most important lines of the 4,700 miles of railway track have been electrified.

There are 660 miles of federal superhighways, as well as nearly 17,700 miles of federal, provincial, and county roads. The Rhine is one of the world's most heavily travelled water routes. Together with the state's important canals, it serves primarily as a means of transport for mass-produced goods. Duisburg-Ruhrort, at the mouth of Ruhr River, is the largest inland harbour in Europe. The two largest airports of the region are located at Düsseldorf-Lohausen and Cologne-Wahn.

Administration and social conditions. *Government.* According to the state constitution of 1950, representatives to the state parliament, or Landtag, are elected by direct, secret ballot. Political activity operates within a multiparty system, and state government is formed by coalition. The *Land* is presided over by the minister-president, who appoints his cabinet. The state is divided into the six major administrative districts of Düsseldorf, Cologne, Aachen, Münster, Detmold, and Arnsberg; these are further divided into local administrative units, municipalities, and unincorporated cities. In the local districts and municipalities, the chief administrator shares responsibility with the head of the council or the mayor.

Administrative divisions

The regional leagues of the Rhineland and Westphalia-Lippe have their seats in Cologne and Münster, respectively. Specific duties—including welfare, traffic, and cultural affairs—have been transferred to them from the local administrative districts and unincorporated cities. A nonmunicipal association for the coal-mining territory of the Ruhr has its seat at Essen.

Justice. There are three superior courts of appeal, in Düsseldorf, Hamm, and Cologne; 19 provincial courts; and 160 magistrates' courts with approximately 2,700 judges. In addition to the regular courts, there are administrative courts and courts of finance, labour, and social relations.

The police. The police are under the authority of the national Ministry of the Interior. The areas of state police jurisdiction coincide with the governmental districts, while the areas of local police jurisdiction agree generally with the local administrative districts and the unincorporated cities.

Education. Free public schools for general education are divided into primary schools, or *Grundschulen;* upper primary schools, or *Hauptschulen;* advanced training schools, or *Fachoberschulen;* secondary schools, or *Realschulen;* and grammar schools, or *Gymnasien.* In addition, there are special schools for those with learning disabilities. Evening classes are available for professional people desiring to obtain a diploma for completion of grammar or secondary school studies.

The system of colleges has been greatly expanded to include universities in Bonn, Münster, Aachen, Cologne, Bochum, Düsseldorf, Dortmund, and Bielefeld. There are also several education colleges. The total college enrollment—including theological and denominational colleges—numbered about 100,000 in the early 1970s.

Health and welfare. In the late 1960s many hospitals were newly constructed or enlarged. There were about 750 hospitals, with 110 beds and six staff doctors for every 10,000 persons. Public welfare services include sickness, accident, revenue, and unemployment insurance, the funds for which are raised through premiums paid by the insured and their employers.

The standard of living index has been steadily rising since the early 1960s, and gross weekly earnings of those employed in industry had risen about 58 percent by 1969. Durable consumer goods, such as television receivers, refrigerators, and automobiles are common, but less than one-third of the population owns telephones. Multiple housing units, as opposed to single-family dwellings, are prevalent in the cities.

The common economic interests of employers in the industrial sector are represented by some 20 industrial federations and chambers of commerce. There are also eight trade associations that operate on the principle of self-administration in the same manner as the two agricultural associations.

Industrial associations

Over 2,000,000 employees are represented through labour unions that are affiliated with the German Federation of Trade Unions (Deutscher Gewerkschaftsbund); the German Federation of Office-workers (Deutsche Angestellten-Gewerkschaft) and the German Federation of Officials (Deutscher Beamtenbund) also have considerable memberships.

Cultural life and institutions. Cultural institutions, such as the theatre, orchestras, museums, and libraries are to be found in considerable numbers. An attempt was being made in the 1970s to organize them into a unified and coordinated system. There are 21 cities with municipal and state theatres, 13 private indoor theatres and studios, and from 15 to 20 summer, open-air theatres. Symphony orchestras and choirs are cultivated at the state and municipal levels, and further cultural development is promoted by the Folkwang School (Folkwangschule) in Essen and the state music conservatories in Cologne and Detmold.

Many cities possess museums noted for their excellent exhibits of the history of art and culture. There are also a number of small local museums and the national gallery for the state collection of works of art in Düsseldorf. In 1963 the Prize for Art of the State of Nordrhein-Westfalen was established as a distinction for excellence in artistic endeavours. There are state archives in Düsseldorf, Münster, Detmold, and Brühl, and the state's 75 libraries possess a total of 9,000,000 volumes.

Closer attention is being given to the maintenance and care of architectural and cultural monuments as a result of destruction during World War II. The older industrial buildings, some of them of considerable interest to industrial archaeologists, are also maintained as monuments.

There are about 100 daily newspapers with almost 250 local editions and a total circulation of approximately 4,000,000. The West German Radio in Cologne broadcasts three daily musical programs and is involved in 25 percent of the community programming on German television. It also broadcasts a regional program.

Future prospects. As a result of its highly developed industrial economy, Nordrhein-Westfalen is faced with the problems of balanced economic advance. Growing industry places a high demand on the area's natural resources, and the extraction industries must face the problems of reserve depletion. Improvement of the environment is of particular importance. Measures against air and water pollution have been enacted, and recreational facilities are being expanded. In the Ruhr region, new parks have been opened, and entertainment centres are under construction or planned. In the east, the health baths have been promoted and 13 nature parks have been established, especially in the richly forested mountains. These facilities are close to the industrial cities and are planned to promote an aesthetic as well as physical relief from urban problems. (W.v.K.)

North Africa, History of

North Africa as here used means the area of present-day Morocco, Algeria, Tunisia, and Libya. In ancient times the Greeks used the word Libya, derived from a tribal name on the Gulf of Sidra, to describe the land north of the Sahara, the territory whose native peoples were subjects of Carthage; and also the whole continent. The Romans applied the name Africa (of Phoenician origin) to their first province in the northern part of Tunisia, to the whole area north of the Sahara, and to the whole continent. The Arabic term Jazīrat al-Maghrib, meaning island of the west, applied first of all to land west of Egypt but more particularly to present-day Tunisia, Algeria, and Morocco. French scholars regularly use Maghrib in this sense, and it is so used here. The Arabs used the word Barbar (Berber), perhaps derived from the Latin *barbari*, to describe the non-Latin-speaking peoples of the Maghrib at the time of the Arab conquest, and it has been used in modern times of the non-Arabic-speaking population (French Berbères, Berbers). A frequent usage refers to the non-Phoenician and non-Roman inhabitants of classical times, and their language, as Berber. It should be stressed that the theory of a continuity of language and race between ancient inhabitants and the modern Berbers is not proved; consequently the word Libyan is used in this article of these peoples in ancient times. This article is outlined as follows:

I. Ancient North Africa

GEOGRAPHY AND PREHISTORY

The description of the area as an island refers to its isolation, bounded by the sea to the north and west, by the Sahara to the south and east. The desert has been the dominant factor in the North African environment, though it has not always been as dry as it is today. At various times during the last million years there have been periods of abundant rainfall, the last occurring about the 6th millennium BC at the beginning of the Neolithic Period (see below). A major trade route connecting the Mediterranean with the African world existed along the Hoggar-Tibesti ridge in the central Sahara, and it is probable that communications existed across the western Sahara also. Nevertheless, the Sahara always constituted a formidable barrier to the movement of techniques and peoples. In ancient historical times, much of North Africa was evergreen forest or scrub of Mediterranean type, and the fauna included animals such as the elephant, zebra, and ostrich, now extinct in the area.

The mountains were of the utmost importance in the historical development of the area. They run generally from east to west, parallel to the coast, with their highest elevations in the Atlas range. They are not continuous but constitute separate blocks, especially in the coastal areas. Although it was in the mountains that the rainfall was highest, the forest was intractable, and early settlements tended to opt for the plains and valleys between or south of the mountains. The Mediterranean coast, for much of its length, is extremely inhospitable, offering few natural harbours and still fewer natural lines of communication into the interior. Even the major rivers, such as the Majardah and Cheliff, are unnavigable. Only in northeastern Tunisia is the coastline more favourable, and it is not surprising that the main movement of culture and conquest have moved from there westward.

The coastal strip in the area of Tripoli (Ṭarābulus) in western Libya is an extension of that of Tunisia. To the east, some 800 miles of the Sirtic Desert separates it from Barqah (Cyrenaica) at the eastern end of modern Libya, which thus has had a very different history from the Maghrib. Settlement there was effectively confined to the elevated plateau of al-Jabal al-Akhḍar and did not extend more than about 70 miles south of the coast. Barqah's contact with Egypt was limited by an intervening 600 miles of semidesert.

The Maghrib provided the paradox of an area in which various cultures have imposed some measure of uniformity, while political unity has been rare; for this geography is largely responsible. The area of settlement is of vast length but little breadth and has no natural centre from which political uniformity could be imposed; its natural communications have never been easy, and the mountain blocks have been large enough to maintain populations to a greater or lesser degree independent of and hostile to those that controlled the plains.

Earliest inhabitants

Although there is uncertainty about some factors, Ain Hanech (in Algeria) is the site of one of the earliest traces of human occupation in the Maghrib. Somewhat later but better attested are sites at Ternifino and Sidi Abd er-Rahmane (both in Morocco). The former produced hand axes associated with *Atlanthropus mauretanicus*, a hominid more primitive than other Neanderthaloids of the same period. Sidi Abd er-Rahmane also produced evidence of *Atlanthropus*, with a date of 200,000 BC at least.

Succeeding these early hand-ax remains are the Levalloiso-Mousterian industries similar to those of the Levant. It is claimed that nowhere did the middle Paleolithic evolution of flake-tool techniques reach a higher state of development than in North Africa. Its high point in variety, specialization, and standard of workmanship is named Aterian, after the type site of Bir el-Ater in southern Tunisia; assemblages of this material occur all over the Maghrib and the Sahara. Radiocarbon testing from Morocco indicates a date for early Aterian material of *c.* 30,000 BC. Its spread appears to have taken place during one of the periods of desiccation, and the carriers of the tradition were clearly adept desert hunters. The few asso-

ciated human remains are Neanderthaloid, with substantial differences between those found in the west and those in Cyrenaica. In the latter area, a date of *c.* 43,000 BC for the Levalloiso-Mousterian has been obtained (at Haua Ftoah). The tools and a fragmentary human fossil of Neanderthaloid type are almost identical to those of Palestine.

The earliest blade industries of the Maghrib, associated as in Europe with the final supersession of Neanderthaloids by *Homo sapiens*, are named Oranian (type site, La Mouilla, near Oran in western Algeria), a culture of obscure origin, which seems to have spread along all the coastal areas of the Maghrib and Cyrenaica between *c.* 15,000 and 10,000 BC. Following the Oranian was the Capsian, the origin of which is also obscure. Its most characteristic sites are in the area of the great salt lakes of southern Tunisia, the type site being al-Maqṭa' (el-Mekta), near Qafṣah (Capsa). The climate during both Oranian and Capsian appears to have been relatively dry and the fauna one of open country, ideal for hunting. Between *c.* 9,000 and 5,000 BC, upper Capsian spread northward to influence the Oranian and also eastward to the Gulf of Sidra, and by the 3rd millennium BC, if not earlier, a uniform human type appears to have been established through the Maghrib. Since there is much evidence that the Neolithic culture of the Maghrib was not introduced by invasion but through the acceptance of new ideas and technologies by the Capsian peoples, it is possible that they were the ancestors of the Libyans known in historic times.

Neolithic culture

The spread of early Neolithic culture in Libya and the Maghrib occurred during the 6th and 5th millennia BC and is characterized by the domestication of animals and the shift from hunting and gathering to self-supporting food production (often still including hunting). The pastoral economy, with cattle the chief animal, remained dominant in North Africa until the classical period. Although the new type of economy may have originated in Egypt or the Sudan, the character of the flint-working tradition of the Maghribian Neolithic argues in favour of the survival of much of the earlier culture, which has been called Neolithic-of-Capsian tradition. Accordingly, the transition, if not of independent local origin, is best explained by the gradual diffusion of new techniques rather than by massive immigration of new peoples.

The Neolithic-of-Capsian tradition in the Maghrib persisted at least into the 1st millennium BC with relatively little change and development; there was no great flourishing of late Neolithic culture nor little that can be described as a Bronze Age. North Africa was wholly lacking in metallic ores other than iron, hence most tools and weapons continued to be made of stone until the introduction of ironworking techniques.

Prehistoric rock carvings flourished in the southern foothills of the Atlas south of Oran and in the Hoggar and Tibesti ranges. While some are relatively recent, the great majority appear to be of the Neolithic-of-Capsian tradition. Some show animals locally or even totally extinct, such as the giant buffalo, elephant, rhinoceros, and hippopotamus in areas now covered by desert. While Egyptian influence may be discerned, the character of the rock art is so different from that of Egypt that it can hardly be said to derive from it. On the other hand, it is very much later than the rock paintings of Paleolithic times in southwest Europe, and an independent development is not excluded. The art is primarily that of a culture that remained largely, though not exclusively, dependent on hunting and that survived on the Saharan fringes until historical times.

There are many thousands of large, stone-built surface tombs in North Africa that appear to have no connection with earlier megalithic structures found in northern Europe, and it is unlikely that any of them is earlier than the 1st millennium BC. Large structures such as the tumulus at Mzora (54 metres in diameter) and the mausoleum known as the Medracen (40 metres in diameter) are probably of the 4th and 3rd centuries BC and show Phoenician influence, though there is much that appears to be purely Libyan.

THE CARTHAGINIAN PERIOD

The Phoenician settlements. North Africa (with the exception of Cyrenaica) entered the mainstream of Mediterranean history with the arrival in the 1st millennium BC of Phoenician traders on its coast. The Phoenicians were not looking for land to settle but for anchorages and staging points on the trade route from Phoenicia to Spain, a source of silver and tin. Points on an alternative route by way of Sicily, Sardinia, and the Balearic Islands also were occupied. The Phoenicians lacked the manpower and the need to found large colonies as the Greeks did, and few of their settlements grew to any size. The sites chosen were generally offshore islands or easily defensible promontories with sheltered beaches on which ships could be drawn up. Carthage (from the Phoenician Kart-Hadasht, New City), destined to be the largest Phoenician colony and in the end an imperial power, conformed to the pattern.

Tradition dated the foundation of Gades (modern Cádiz; the earliest known Phoenician trading post in Spain) to 1110 BC, Utica (Utique) to 1101 BC, and Carthage to 814 BC. The earlier dates appear legendary, and no Phoenician object earlier than the 8th century BC has yet been found in the west. At Carthage some Greek objects have been found, datable to about 750 or slightly later, which comes within two generations of the traditional date. Little can be learned from the romantic legends about the arrival of the Phoenicians at Carthage transmitted by Greco-Roman sources. Though individual voyages doubtless took place earlier, the establishment of permanent, or at any rate seasonal, posts is unlikely to have taken place before 800 BC, but they antedate the parallel movement of Greeks to Sicily and southern Italy.

Material evidence of Phoenician occupation of 8th-century-BC date comes from Utica, and of the 7th or 6th century BC from Hadrumetum (Sūsah, Sousse), Tipasa (east of Cherchell), Siga (Rachgoun), Lixus, and Mogador, the last being the most distant Phoenician settlement so far known. Finds of parallel date have been made at Motya (Mozia) in Sicily, Nora (Nurri), Sulcis, and Tharros (Torre di S. Giovanni) in Sardinia, and Cádiz and Almuñécar in Spain. Unlike the Greek settlements, however, those of the Phoenicians long remained politically dependent on their homeland, and only a few were situated where the hinterland had the potential for development. The emergence of Carthage as an independent power, leading to the creation of an empire based on the secure possession of the North African coast, resulted less from the weakening of Tyre, the chief city of Phoenicia, by the Babylonians than from growing pressure from the Greeks in the western Mediterranean; in 580 BC some Greek cities in Sicily attempted to drive the Phoenicians from Motya and Panormus (Palermo) in the west of the island. The Carthaginians feared that if the Greeks won the whole of Sicily they would move on to Sardinia and beyond, isolating the Phoenicians in North Africa. The successful defense of Sicily was followed by attempts to strengthen limited footholds in Sardinia; a recently discovered fortress at Monte Sirai is the oldest Phoenician military building in the west. The threat from the Greeks receded when Carthage, in alliance with Etruscan cities, checked the Phoenicians off Corsica in 535 and succeeded in excluding the Greeks from contact with southern Spain. A further success occurred in Africa itself; in 514 BC a Spartan named Dorieus attempted to found a settlement at the mouth of the Cinyps River (Oued Oukirri) in Libya, but the Carthaginians, regarding this as an intrusion into their own territory, were able to expel the Spartans with the help of native Libyans.

Relations with the Greeks

Carthaginian supremacy. By the 5th century BC, active military participation by Tyre in the west had doubtless ceased; from the latter half of the 6th century it was under Persian rule. Carthage thus became the leader of the western Phoenicians, and in the 5th century formed an empire of its own, centred on North Africa, which included existing Phoenician settlements, new ones founded by Carthage itself, and a large part of modern Tunisia. The actual stages of the growth of Carthaginian power are not known, but the process was largely completed by

the beginning of the 4th century. The whole of the Cap Bon (Jazīrat Sharīk) peninsula was occupied early, ensuring Carthage a fertile and secure hinterland. Subsequently, penetration extended southwestward as far as a line running roughly from al-Kāf to the coast at Thaenae (now the ruins of Thīnah, or Tina). Penetration occurred south of this line later, Theveste (Tabassah, Tébessa) being occupied in the 3rd century BC. In the Cap Bon peninsula, where the Carthaginians developed a prosperous agriculture, the native population may have been enslaved, but elsewhere these people were obliged only to pay tribute and furnish troops.

Carthaginian settlements

Carthage maintained an iron grip on the entire coast, from the Gulf of Sidra to the Atlantic coast of Morocco, establishing many new settlements to protect its monopoly of trade. These were mostly small places, probably of only a few hundred inhabitants. The Greeks called them *emporia*, markets where native tribes brought articles to trade, which could also serve as anchorages and watering places. Permanent settlements in modern Libya were few and dated after the attempt of Dorieus to plant a Greek colony there. Though in time fishing and agriculture played a part in their wealth, Leptis with its neighbours Sabratha and Oea (Tripoli) became rich through trans-Saharan trade; Leptis was the terminus of the shortest route across the Sahara linking the Mediterranean with the Niger. A Carthaginian named Mago is said to have crossed the desert several times, but doubtless much of the trade (in precious stones) came through intermediate tribes. Other stations on the Gulf of Gabes included Zouchis, known for its salted fish and purple dye, Gigthis (Bū Ghirārah), and Tacapae (Qābis, Gabes). North of Thaenae was Acholla, traditionally an offshoot of the Phoenician settlement on Malta, Thapsus (Rass Dimas), Leptis Minor, and Hadrumetum, the largest city on the east coast of Tunisia. From Neapolis (Nabeul), a road ran direct to Carthage across the base of the Cap Bon peninsula.

West of Carthage there have been changes in the course of the Bagradas (Majardah) River; as a result, Utica, a port in Carthaginian and Roman times, is now some seven miles from the sea. Utica was second only to Carthage in importance among the Phoenician settlements and always maintained at least a nominal independence. Beyond Cape Farina (Ra's Sīdī ʿAlī al-Makkī) as far as the Straits of Gibraltar, the coast offered a number of anchorages, but few of the stations reached anything like the prosperity of those on the Gulf of Gabes and the east coast of Tunisia. One of the more important was Hippo Diarrhytos (Banzart, Bizerte), whose natural advantages as a port were utilized at an early date; another Hippo, later called Hippo Regius (Bône), was also probably of Carthaginian origin. Along the same stretch of coast were Rusicade (Philippeville, Skikda) and Collo (Chullu). Still farther west, a number of place-names known from the Roman period betray an earlier Phoenician interest through the incorporation of a Phoenician element, *rus*, meaning "cape"; *e.g.*, Rusuccuru (Dellys) and Rusguniae (Matifou). Tingi (Tangier) was already settled in the 5th century BC.

Trade. Ancient sources agree that Carthage had become perhaps the richest city in the world through its trade, yet very few traces of its wealth have been discovered by archaeologists. This is no doubt because most of it was in perishables—textiles, unworked metal, foodstuffs, and slaves; its trade in manufactured goods was only a small part of the whole. There can be no doubt that the most profitable trade was that inherited from the Phoenicians in the western Mediterranean, in which tin, silver, gold, and iron were obtained in exchange for manufactures and consumer goods of small value. Carthage ruthlessly maintained its monopoly of this trade from the late 6th to the end of the 3rd century BC by sinking intruders and exacting recognition of its position from other states. Its wealth is attested by the vast mercenary armies it was able to raise and the mintage of gold coins in the 4th century far in excess of that known for other advanced states.

It was apparently in connection with this trade that during the 5th century there occurred two voyages of exploration and trade, evidently of particular importance since reports of them were known to later generations of Greeks and Romans. One was along the Atlantic coast of Morocco, the other northward along the Atlantic coast of Spain. They were led by Hanno and Himilco, respectively, both members of a leading family in Carthage. Hanno's voyage is generally associated with an account in Herodotus, writing about 430 BC, of Carthaginian trade on the Atlantic coast of Morocco. Herodotus describes a system of dumb barter with primitive peoples, by which the Carthaginians exchanged manufactured goods for gold. It is not known where the exchanges took place; the Río de Oro is a possibility, and it is probable that Hanno's expedition went beyond Cape Verde. Nevertheless, the "gold route" did not survive the fall of Carthage and was unknown to the Romans. This has led some scholars to argue that the Carthaginians' interest in the Atlantic coast of Morocco was stimulated by the more prosaic attraction of the abundant fish.

Voyages of exploration

Himilco's voyage also was known to the Greeks and Romans. He sailed north along the Atlantic coast of Spain, Portugal, and France and reached the territory of the Oestrymnides, a tribe living in Brittany. The purpose of this voyage was apparently to consolidate control of the trade in tin along the Atlantic coast of Europe. It followed the route used by the Tartessians, a people of southern Spain in the area where Cádiz had been founded, who knew of Ireland and Britain. This trade was no doubt the latest phase of contact between the various areas of the Atlantic seaboard that went back to Late Neolithic times. There is no evidence that Himilco reached Britain, nor indeed has any Phoenician object ever been found in the island, but probably Cornish tin was obtained through the tribes of Brittany. Tin was also obtained from northwestern Spain. It is notable that, at Cádiz, the Carthaginian tombs found at intervals over the past century have produced nothing earlier than the 5th century, which would indicate that it was not until that date that it became a large and permanent base for the exploitation of trading opportunities in the west.

Trading contacts with the Greek world had been substantial from the earliest period of Phoenician colonization, in spite of the intermittent wars with the Greeks of Sicily. Pottery from Corinth, Athens, Ionia, Rhodes, and other Greek centres has been found at Carthage, Utica, and many other sites, as well as imports from Phoenicia itself and from Egypt. It is known that Selinus, a Greek city in Sicily, grew wealthy from trade with Carthage, probably in foodstuffs, before Carthage conquered the neighbouring hinterland. During the 5th century there appears to have been a decline in imports from the Greek world. One factor that inhibited trade was the lack of a Carthaginian coinage before the early 4th century, though most important Greek states had had their own coinages for at least a century before that. Carthaginian merchants, however, did not cease to frequent Greek ports, and a number of them were established at Syracuse in 398. From that date, economic contacts with advanced states seem to have revived, especially after the conquests of Alexander the Great in the eastern Mediterranean created a new market for the cheap Carthaginian manufactured goods. The Carthaginian merchant became such a familiar figure in such economic centres of the Greek world as Athens and Delos that there were Greek comedies in which the central figure was the Carthaginian trader.

Revival of Economic Contacts

Wars outside Africa. Except in backward or thinly populated areas Carthage's foreign policy was nonexpansive. One major departure was a disaster: in 480 Carthage intervened in intercity struggles among the Greeks of Sicily and suffered a heavy defeat at Himera. After a long period of peace, it went in 410 to the help of Segesta, an ally in Sicily, and turned the war into one of revenge for the earlier defeat. After initial successes, including the destruction of Himera, a treaty confirmed Carthage's control of the west of the island. During the 4th century most of the wars were caused by the attempts of various rulers of Syracuse to drive the Carthaginians out of Sicily; three of these (398–392, 382–375, and 368) were with

Dionysius I of Syracuse. Most of the time the eastern limit of Carthaginian power in the island was recognized as the Halycus (Platani) River. The only occasion in which Carthage suffered directly (since its armies were largely mercenary) was in 310, when the ruler of Syracuse, Agathocles, under heavy pressure in Sicily, launched a daring invasion of Africa, the first experienced by Carthage. Over a period of three years he caused great devastation in Carthaginian territory in eastern Tunisia but in the end was defeated.

Treatment of subject peoples. Carthage was notorious in antiquity for oppressing and exacting excessive tribute from its subjects. There were, however, different categories of subject community, the most favoured being the original Phoenician settlements and the colonies of Carthage itself. There is little evidence of opposition among them to Carthaginian control. Similar institutions and laws may be attributed to a common cultural background rather than to an attempt to impose uniformity. Carthage exacted dues on imports and exports and levied troops and probably sailors. Carthaginian subjects of various nationalities in Sicily also received favourable treatment, at least in economic matters. Relatively free trade was allowed until the end of the 5th century, and a number of cities had their own coinage. In the 4th century, some Sicilian Greek states became subject to Carthage, paying a tribute amounting apparently to one-tenth of their produce. It was the Libyans of the interior who suffered most, though few were reduced to slavery. During the First Punic War, Libyans had to pay one-half of their crops as tribute, and it is supposed that the normal exaction was a quarter, a burdensome imposition. They were also required to provide troops, and from the early 4th century they formed the largest single element in the Carthaginian army; it is unlikely that they received pay except in booty before the Punic Wars. The Carthaginians are said to have "admired not those governors who treated their subjects with moderation but those who exacted the greatest amount of supplies and treated the inhabitants most ruthlessly." This judgment (by the Greek historian Polybius) was made in connection with the Libyans, and a destructive revolt—one of a number known—that followed the first Punic War. In this revolt (241–237 BC), mercenaries, unpaid after the Carthaginian defeat in the First Punic War, revolted and for a while controlled much of Carthage's North African territory. It was fought with great atrocities on both sides, and the Libyans were among the most fervent of the rebels. They even issued coins on which the name Libyan appears (in Greek), which probably indicates a growing ethnic consciousness. Notwithstanding this relationship, Carthaginian civilization had profound effects on the material culture of the Libyans (see below).

Political and military institutions. Hereditary kingship prevailed in Phoenicia down to Hellenistic times, and Greek and Roman sources refer to kingship at Carthage. It appears not to have been hereditary but elective, though in practice one family, the Magonid, dominated in the 6th century. The power of the kingship was diminished during the 5th century, a development that has its parallels in the political evolution of Greek city-states and of Rome. Roman sources directly transcribe only one Carthaginian political term—*sufet*, etymologically the same as the Hebrew *shofet*, generally translated "judge" in the Old Testament but implying much more than merely judicial functions. At some stage, probably in the 4th century, the *sufets* became the chief magistrates of Carthage and other western Phoenician settlements. Two *sufets* were elected annually by the citizen body, but all were from the wealthy classes. The chief power rested with an oligarchy of the wealthiest citizens, who were life members of a council of state and decided all important matters unless there was serious disagreement with the *sufets*. A panel of judges chosen from among its members had obscure but formidable powers of control over all organs of government.

Carthaginian magistrates

During the 6th and 5th centuries, most military commands were held by kings, but later the generalship was apparently dissociated from civil office. Even in the time of the kings, military authority appears to have been conferred upon the kings only for specific campaigns or in emergencies. The generals are said to have been regarded as potential overthrowers of the legal government, but in fact there is no record of any army commander's having attempted a coup d'etat.

Up to the 6th century, the armies of Carthage were apparently citizen levies similar to those of all city-states of the early classical period. But Carthage was too small to provide for the defense of widely scattered settlements, and it turned increasingly to mercenaries, officered by Carthaginians, with citizen contingents appearing only occasionally. Libyans were considered particularly suitable for light infantry, Numidians and Mauretanians for light cavalry; Iberians and Celtiberians from Spain were used in both capacities. In the 4th century the Carthaginians also hired Gauls, Campanians, and even Greeks. The disadvantages of mercenary armies were more than outweighed by the fact that Carthage could never have stood the losses incurred in a whole series of wars in Sicily and elsewhere. Very little is known about the manning of the Carthaginian fleet; technically, it was not overwhelmingly superior to those of the Greeks, but it was larger and had the benefit of experienced sailors from Carthage's maritime settlements.

The city. The Romans completely destroyed Carthage in 146 BC, and a century later built a new city on the site, so that little is known of the physical appearance of the Phoenician city. It is almost certain that the ancient artificial harbour—the Cothon—is represented today by two lagoons north of the bay of al-Karm (el-Kram). In antiquity it had two parts, the outer rectangular part being for merchant shipping, the interior, circular division being reserved for warships; sheds and quays were available for 220 warships. Its small size probably means that it was chiefly used in winter when navigation almost ceased. The city walls were of great strength and 22 miles in length; the most dangerous section, across the isthmus, was over 40 feet high and 30 feet thick. The citadel on the hill called Byrsa was also fortified. Between Byrsa and the port was the heart of the city, its marketplace, council house, and temples. In appearance it may have been not dissimilar to towns in the eastern Mediterranean or Persian Gulf before the impact of modern civilization, with narrow winding streets and houses up to six stories high. The exterior walls were blank except for a solitary street door, but they enclosed courtyards. A figure of 700,000 for the city population is given by the geographer Strabo, but this probably included the population of the Cap Bon peninsula. A more reasonable figure could be about 400,000, including slaves, similar to that of Athens.

Physical appearance of Carthage

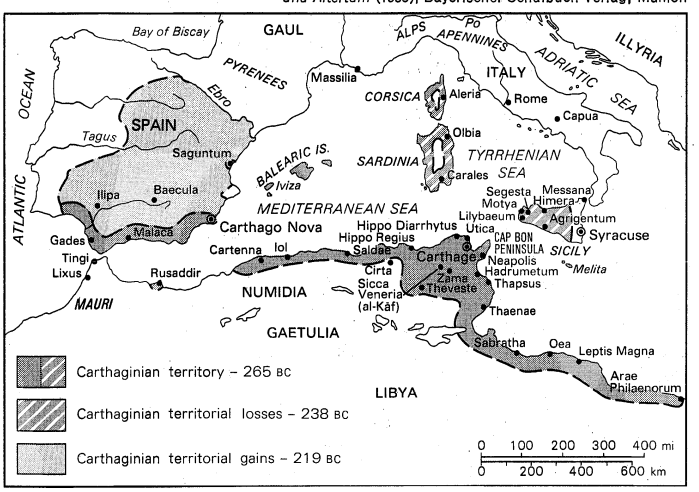
Religion and culture. The Carthaginians were notorious in antiquity for the intensity of their religious beliefs, which they retained to the end of their independence and which in turn influenced the religion of the native peoples. The chief deity was Baal Hammon, the community's divine lord and protector. A sombre god, he was identified by the Greeks with Cronus and by the Romans with Saturn. During the 5th century a goddess named Tanit came to be widely worshipped and represented in art. It is possible that her name is Libyan and that her popularity was connected with the acquisition of land in the interior, as she is associated with symbols of fertility. These two overshadow other deities such as Melkart, principal deity of Tyre, identified with Heracles, and Eshmoun, identified with Aesculapius. Human sacrifice was the element in Carthaginian life most criticized; it persisted in Africa much longer than in Phoenicia. The child victims were sacrificed to Baal (not to Moloch, an interpretation based on a misunderstanding of the texts), and the burnt bones buried in urns under stone markers, or stelae. At Carthage thousands of such urns have been found in the "Sanctuary of Tanit," and similar burials have been discovered at Hadrumetum, Cirta, Motya, Calaris, Nora, and Sulcis. The whole character of Carthaginian religion was sombre, being one of the weakness of human beings in the face of the overwhelming and capricious power of the gods. The great majority of Carthaginian personal names, unlike those of Greece and Rome, were of reli-

gious significance; *e.g.*, Hannibal, "Favoured by Baal," or Hamilcar, "Favoured by Melkart."

Carthaginian artistic and intellectual achievements

In comparison with the extent of its power and influence, the artistic and intellectual achievements of Carthage were small. What limited remains of buildings survive—mostly in North Africa and Sardinia—are utilitarian and uninspired. In the minor arts—pottery, jewelry, metalwork, objects in terra-cotta, and the thousands of carvings on stelae—a similar lack of inspiration may be felt. The influence of Phoenician, Egyptian, and Greek artistic traditions can be observed, but they failed to stimulate as they did, for example, in Etruria. There is no evidence that Greek philosophy and literature made much impact, though certainly many Carthaginians in the city's later history knew Greek and there were libraries in the city. One work is known, a treatise on agriculture by a certain Mago, but this may have been based on Hellenistic models. On the whole, the Carthaginians adhered to traditional modes of thought, which no doubt gave them a sense of solidarity amid more numerous and hostile peoples. Their fanatical patriotism enabled them to offer a more prolonged resistance to Rome than any other power. Their influence on North African history was, in the first place, to bring it into the mainstream of the advancing civilization of the Mediterranean world; more particularly it introduced into North Africa more advanced techniques leading to agricultural progress, which implied in turn a change by many Libyans from a seminomadic to a stable way of life, and the possibilities of urbanization, which were fully realized in the Roman period.

From *Grosser Historischer Weltatlas*, vol. 1, *Vorgeschichte und Altertum* (1953); Bayerischer Schulbuch-Verlag, Munich



Carthaginian Empire.

Carthage and Rome. In the 3rd and 2nd centuries BC, Carthage was weakened and finally destroyed by Rome in the three Punic Wars (*q.v.*). Treaties between Carthage and Rome had been made in 508 and 348, and for a long period the two powers had no conflicting interests. But by the 3rd century Rome dominated all southern Italy and thus approached the Carthaginian sphere in Sicily. In 264 Rome accepted the submission of Messana (Messina), though this state had previously had a Carthaginian garrison, partly because of exaggerated fears of a possible Carthaginian threat to Italy and partly because of hopes of gain in Sicily. For Carthage, a Roman foothold in Sicily would upset the traditional balance of power on the island. The ensuing First Punic War, which lasted until 241, was very costly in human life, with losses of tens of thousands being recorded in some naval engagements. Contrary to expectation, the Carthaginian fleet was worsted on several occasions by the newly built Roman; on land the Romans failed to drive the Carthaginians out of Sicily, and a Roman invasion of Tunisia ended in catastrophe. Carthage made peace after a final naval defeat off the Aegates Islands, surrendering its hold on Sicily. Sardinia fell to Rome in 238.

In response to the defeat, Carthage, under the leadership of Hamilcar Barca and his successors (usually described as the Barcid family), set about establishing a new empire in Spain. The object appears to have been to exploit the mineral wealth directly rather than through intermediaries and to mobilize the manpower of much of Spain into an army that could match that of Rome. Hamilcar and his son-in-law Hasdrubal built up an army of over 50,000 Spanish infantry and occupied half of the Iberian Peninsula. Finally, in 219, Hannibal, Hamilcar's son, ignored Roman threats designed to prevent the consolidation or extension of the new empire. His invasion of Italy and the crushing defeats he inflicted on the Romans at Lake Trasimene (217) and Cannae (216) were the gravest danger Rome had ever faced. The majority of Rome's allies and subjects in Italy remained loyal, however, and Hannibal found increasing difficulty in getting supplies and reinforcements. In 204 Scipio Africanus landed near Utica with a Roman army, and in 203 Hannibal was recalled from Italy; but he was defeated by Scipio at Zama (Sabʿ Biʾār) in 202. Carthage made peace soon afterward, surrendering its fleet, its overseas possessions, and some of its African territory. During the next 50 years it retained some measure of prosperity, although frequently under pressure from the Numidians under King Masinissa. From 155 BC irrational fears of a Carthaginian revival were stimulated at Rome by Cato the Elder, and in 149, on flimsy pretexts, the Carthaginians were forced to choose between evacuating their city and settling inland, or a doomed resistance. They chose the latter, and, after a three-year siege, the city was destroyed and its site ceremonially cursed by the younger Scipio.

New empire in Spain

The Greeks in Cyrenaica. The natural contacts of Cyrenaica were northward with Crete and the Aegean world. Its coast was visited by Cretan fishermen in the 7th century, and the Greeks became aware that it was the only area in North Africa still available for colonization. A severe overpopulation on the small Cyclades island of Thera (Santorini) led to the foundation of Cyrene (*c.* 630) on a site in easy reach of the sea, well watered, and in the fertile Jabal al-Akhḍar plateau. The founder's name was, or was changed to, Battus, a Libyan word meaning king. For some time friendly relations existed with the local peoples, and there was more intermarriage with non-Greek women than was usual in colonies. Later, when more colonists were attracted by the increasing prosperity, hostilities occurred in which the settlers were successful. Cyrene also repulsed an invasion by the Egyptians (570) but in 525 submitted to Persia. Meanwhile, Cyrene had established other Greek cities in the area—Barca (al-Marj), Taucheira (Tūkrah), and Euhesperides (Benghazi), all of which were independent of their founding city. During the 6th century, Cyrene rivalled the majority of other Greek cities in its wealth, manifested in part by substantial temple building. Prosperity was based on grain, fruit, horses, and, above all, on an apparently extinct plant, *Silphium*.

The dynasty of Battus ended *c.* 440 BC with the establishment of a democratic constitution like that of Athens, and the general prosperity of Cyrenaica continued through the 4th century in spite of some political troubles. Cyrenaica submitted to Alexander the Great in the late 4th century and subsequently became subject to the Ptolemies. The cities nevertheless enjoyed a good deal of freedom in running their own affairs. The constitution of Cyrene was elaborated as a fairly liberal oligarchy, with a citizen body of 10,000 and two councils. During the 3rd century a federal constitution for all the Cyrenaican cities was introduced. Apollonia, the port of Cyrene, became a city in its own right; Euhesperides was refounded as Berenice, and a new city, Ptolemais (Ṭulmaythah), was founded, while Barca declined; the term Pentapolis came to be used of the five cities Apollonia, Cyrene, Ptolemais, Taucheira, and Berenice. In 96 BC Ptolemy Apion bequeathed it to Rome, which annexed the royal estates but left the cities free. Disorders led Rome to create a regular province in 74 BC, to which Crete was added seven years later. After Mark Antony temporarily granted it to his daughter Cleopatra Selene, Augustus re-established it with Crete as a senatorial province.

Constitutional changes in Cyrene

THE RISE AND DECLINE OF NATIVE KINGDOMS

Between the destruction of Carthage and the establishment of effective Roman control over the Maghrib there was a period that saw a brief flourishing of native kingdoms. Amid the shifting tribal nomenclature of the sources of various periods, two main groups of relatively sedentary tribes may be distinguished: the Mauri, living between the Atlantic and the Moulouya or perhaps the Chelif rivers, who gave their name to Mauretania; and the Numidae, in the area to the west of that formerly controlled by Carthage. A third group, the Gaetuli, was a largely nomadic people of the desert and its fringe. The various tribes first emerge into history in the late 3rd century BC, no doubt after a period of evolution resulting from contact with Carthaginian civilization. This is difficult to trace, as Carthaginian products were scarce in the interior of the Maghrib before the 2nd century BC, but the large tumuli at Mezora, Sīdī Sulaymān, Souk el-Gour, and the Medracen, apparently royal tombs of the 4th and 3rd centuries BC, testify to a developing economy and society. No doubt service in the Carthaginian mercenary armies was a major stimulus to native Libyan progress. This was most noticeable in Numidia and reached a high point under Masinissa. The son of a chief of the Massyli, a tribe dominating the area between Carthaginian territory and the Ampsaga River (Wādī al-Kabīr), he had been brought up at Carthage and was 20 years old at the outbreak of the Second Punic War. At first his tribe was at variance with Carthage, but in 213 BC it became reconciled when its powerful western neighbours, the Masaesyli, under Syphax deserted Carthage. From 213 to 207 BC Masinissa commanded Numidian cavalry in Spain for the Carthaginians against Rome. On the latter's victory at Ilipa in 207, he returned to Africa where Syphax, now reconciled with Carthage, had occupied some of his tribal territory, including Cirta (Constantine), and his own claims to succession to the chieftainship were disputed. When the Romans landed in Africa in 204 he rendered them invaluable assistance. Recognized by the Romans as king, he annexed the eastern part of Syphax' kingdom and reigned with success until 148 BC. The Greek geographer Strabo said that he "turned the nomads into a nation of farmers." This is exaggerated, since cereal culture had long been established in parts of Numidia, yet there is no doubt that the area of grain production was much enlarged. This was achieved by deliberate encouragement of Carthaginian civilization. Along with new techniques, Carthaginian language, religion, and art penetrated rapidly inland, and Masinissa's capital, Cirta, took on the aspects of a Carthaginian city; incipient urbanization of a number of native villages is also possible. Masinissa issued copper, bronze, and lead coinage for local use, as did some of the Carthaginian coastal towns under his rule.

King Masinissa of Numidia

On his death in 148, his kingdom was divided among his three sons, possibly on the insistence of the Romans, but the latter did not prevent its reunification under Micipsa (148–118 BC). The progress begun under Masinissa continued as refugees from the destruction of Carthage fled to Numidia. Meanwhile, the Romans had formed a province of the area of Tunisia northeast of a line from Thabraca to Thaenae but showed little interest in exploiting its wealth. The attempt by the Roman reformer Gaius Gracchus in 122 BC to found a colony on the site of Carthage failed, though individual colonists who had taken up allotments remained. When Micipsa died, another division of Numidia among three rulers took place, in which Jugurtha (118–105 BC) emerged supreme. He might have been recognized by Rome, but he provoked war when he killed some Italian merchants who were helping a rival to defend Cirta. After some successes due to the incompetence of Roman generals, he was surrendered by Bocchus I, king of Mauretania. The kingdom was again reconstituted under other descendants of Masinissa. The boundaries of the Roman province were slightly enlarged in the area of the upper Majardah, where veterans of the army of Gaius Marius received lands. During the next 50 years there was further immigration by individual Roman settlers and merchants but no deliberate state action. The last relatively formidable king of Numidia was Juba I (*c.* 60–46 BC), who supported the Pompeian side in the Roman civil war between Pompey and Julius Caesar but fell to the dictator in 46 BC at Thapsus. A new province, Africa Nova, was formed from the most developed part of the old Numidian kingdom east of the Ampsaga; it was subsequently (before 27 BC) amalgamated with the original province of Africa by the emperor Augustus. In 33 BC Bocchus II of Mauretania died, bequeathing his kingdom to Rome, but Augustus was unwilling to accept responsibility for so large and relatively backward an area. In 25 BC he installed Juba II, son of Juba I, as king; he ruled until his death about AD 24. He was married to Cleopatra Selene, daughter of Mark Antony and Cleopatra, and under them Iol, renamed Caesarea (Cherchel), and also Volubilis, near Fès(Fez), a secondary capital of the rulers of Mauretania, became centres of late Hellenistic culture. Juba himself was a prolific writer in Greek on a number of subjects, including history and geography. His son Ptolemy succeeded but was executed for unknown reasons by the Roman emperor Caligula in AD 40. A brief revolt followed but was easily suppressed, and the kingdom was divided into two provinces, Mauretania Caesariensis, with its capital at Caesarea, and Mauretania Tingitana, with its capital at Tingi.

Formation of Africa Nova

ROMAN NORTH AFRICA

Administration and defense. For over a century from its acquisition in 146 BC the small Roman province was governed by a minor Roman official from Utica; but changes were made by Augustus, reflecting the growing importance of the area. The governor was henceforth a proconsul, residing at Carthage, after its refounding by Augustus as a Roman colony, and responsible for the whole territory from the Ampsaga in the west to the border of Cyrenaica. The proconsul also commanded the army of Africa and was thus one of the very few provincial governors in command of an army and yet formally responsible to the Senate rather than to the emperor. This anomaly was removed in AD 39 when Gaius entrusted the army to a *legatus Augusti* of praetorian rank. Although the province was not formally divided until 196, the army commander was *de facto* in charge of the area later known as the province of Numidia and also of the military area in southern Tunisia and along the Libyan Desert. The proconsulship was normally held for only one year; like the proconsulship of Asia, it was reserved for former consuls and ranked high in the administrative hierarchy. In the 1st century it was held by several men who subsequently became emperor; *e.g.*, Galba and Vespasian. The commanders of the army normally held the post for two or three years. In the 1st and 2nd centuries AD it was an important stage in the career of a number of successful generals, but in the 3rd century Africa became a military backwater. The two Mauretanian provinces were governed by men of equestrian rank who also commanded the substantial numbers of auxiliary troops in their areas. In times of emergency the two provinces were often united under a single authority.

Tribes on the fringe of the desert and beyond constituted more of a nuisance than a threat as the area of urban and semi-urban settlement gradually approached the limit of cultivable land. A number of minor conflicts with nomadic tribes are recorded in the 1st century, the most serious of which was the revolt of Tacfarinas in southern Tunisia, suppressed in 23. As the area of settlement extended westward as well as south, so the headquarters of the legion moved also; from Ammaedara (Ḥaydarah) to Thevеste under Vespasian, thence to Lambaesis under Trajan. Tribal lands were reduced and delimited, which compelled the adoption of sedentary life, and the tribes were placed under the supervision of Roman "prefects." A southern frontier was finally achieved under Trajan with the encirclement of the Aurès and Nemencha mountains and the creation of a line of forts from Vescera (Biskra) to Ad Majores (Besseriani). The mountains were penetrated during the next generation but were never developed or romanized; nevertheless, they rarely appear to have been a source of disturbance. During the 2nd century, stretches of continuous wall and ditch—the *fossatum*

Westward movement of settlement

Africae—in some areas provided further defense against desert nomads and also marked the division between the settled and nomadic ways of life. To the southwest of the Aurès a fortified zone completed the frontier defensive system or *limes*, which extended for a while as far as Castellum Dimmidi (Messad), the most southerly fort in Roman Algeria yet identified. South of Leptis Magna in Libya, forts on the trans-Saharan route ultimately reached as far as Cydamus (Ghudāmis).

In the Mauretanias the problem was more difficult, because of the rugged nature of the country and the distances involved. The encirclement of mountainous areas, a policy followed in the Aurès, was again pursued in the Kabylie ranges and the Ouarsenis. The area round Sitifis (Sétif) was successfully settled and developed in the 2nd century, but farther west the impact of Rome was for long limited to coastal towns and the main military roads. The most important of these ran from Zarai (Zraia) to Auzia (Aumale) and then to the valley of the Chelif. Subsequently, the frontier ran south of the Ouarsenis as far as Pomaria (Tlemcen). West of this area, it is doubtful whether a permanent road connected the two Mauretanias, sea communication being the rule. In Tingitana, the frontier ran south of a line from Meknès to Rabat. The tribes of the Rif must have lived in virtual independence, and they were probably responsible for a number of wars recorded in Mauretania under Domitian, Trajan, Antonius Pius (which lasted six years), and others in the 3rd century. The effect of these was limited; they did little or no damage to the urbanized areas and never necessitated a permanent increase in the African garrison. For Numidia and the military district in the south of Tunisia and Libya, this amounted to one legion and auxiliaries, about 13,000 men; the Mauretanias had auxiliary units only, totalling some 15,000. This may be contrasted with the position in Britain, where three legions and auxiliaries were required. From the middle of the 2nd century AD the African garrison was largely recruited locally.

The growth of urban life. The most notable feature of the Roman period in North Africa was the development in Tunisia, northern Algeria, and some parts of Morocco of a flourishing urban civilization. This was due in the first place to control of nomadic and pastoral movements, which opened large areas of thinly settled but potentially rich land to consistent exploitation. In addition, there was the incipient urbanization of some parts due to the Carthaginians and the ambitions of Libyan rulers such as Masinissa; and, lastly, the settlement in Africa of Italian immigrants, who, though relatively few in comparison with the population as a whole, provided the impetus to expansion. Julius Caesar settled many veterans in colonies, mostly coastal towns; and, equally important, he established a military adventurer named Publius Sittius along with many Italians at Cirta, beginning the romanization of Numidia. Caesar planned to refound Carthage, and this was effected by Augustus. The number of his original settlers was 3,000, but the colony grew with quite remarkable rapidity because of its favourable geographical position in relation to contact with Rome and Italy. A number of other colonies were founded in the interior of Tunisia and at widely separated places on the Mauretanian coasts. There was also immigration from Italy by private individuals at this time. Colonial foundations of veterans in Mauretania occurred under Claudius (*e.g.*, Tingi, Caesarea, Tipasa). Cuicul (Djemila) and Sitifis were founded by Nerva, and Thamugadi and a number of places nearby, in the area north of the Aurès, by Trajan. The army was a potent vehicle in the spread of Roman civilization and played a major part in the romanization of the frontier regions.

Italian settlers

Though at first inferior in status to the Roman towns, native communities enjoyed the local autonomy that was the hallmark of Roman administration. Between 400 and 500 such units were recognized, the majority of them villages or small tribal fractions. Many, however, advanced in wealth and standing to rival the Roman colonies, acquiring in the process the grant of Roman citizenship, which put the seal of imperial approval on the prosperity, stability, and cultural evolution of developing communities. Naturally, the earliest to show signs of increasing prosperity were the surviving Carthaginian settlements on the coast and places, particularly in the Majardah Valley, where the Libyan population had been much influenced by Carthaginian culture and which now also had numbers of Roman immigrants. Leptis Magna and Hadrumetum received Roman citizenship and the status of a colony from Trajan, and Thubursicu Numidarum (Khamissa) and Calama (Guelma) probably the rank of a municipality. But it was under Hadrian, the first emperor to visit Africa, that the flood tide of such grants occurred; Utica, Bulla Regia (Hammam Derradji), Lares (Lorbeus), Thaenae, and Zama achieved colonial rank, and the process continued throughout the 2nd century. Finally, Septimius Severus, who originated from a rich family of Leptis Magna, and of largely mixed descent, became emperor in AD 193 and showed a great deal of favour to his native land.

In the Maghrib, Roman rule replaced no civic oligarchy, as in the hellenized provinces of Asia Minor, nor a strongly based tribal aristocracy, as in Celtic Gaul. The creation of new wealth and a new social leadership depended on the activity of individuals, and Roman African society became a notable example in antiquity of a self-made bourgeoisie, the wealth being largely in ownership of land. In the 1st century AD there were a few very large estates owned by absentee Roman senators, but they were subsequently absorbed in the extensive imperial estates; the general pattern was of landowning on a more moderate, though still substantial, scale by residents. These landowners made their homes in the towns, not on their estates, and provided the local municipal aristocracies. There was also a numerous class of smaller landowners, but the majority were small tenant farmers with considerable security and traditional rights, on a sharecropping basis. The proportion of slaves is unknown but may well have been smaller than in Italy.

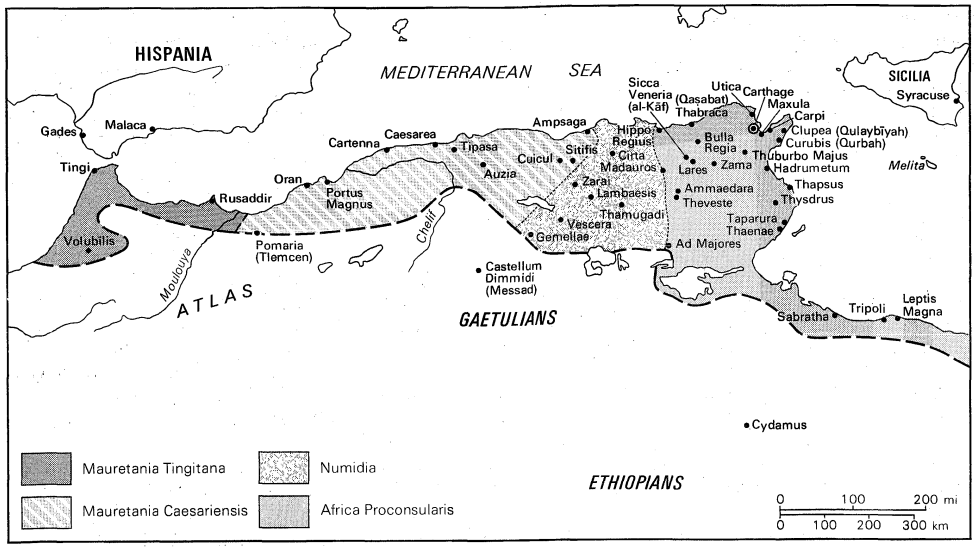
Growth of new social elite

Many of the wealthier Africans entered the imperial administration. The first African consul held office in the reign of Vespasian; at the beginning of the 3rd century, men of African origin held one-sixth of all the posts in the equestrian grade of the administration and also constituted the largest group of provincials in the Senate. It is uncertain what proportion were of native Libyan or mixed origin, but in the 2nd century they were certainly the majority.

During the 2nd and early 3rd centuries the wealthy classes in the towns expended vast sums on their communities in gifts of public buildings such as theatres, baths, and temples; statues; public feasts; and distributions of money. This was a general phenomenon due partly to the lack of alternative profitable investment opportunities and also to a strongly felt social obligation, but the Africans adopted it with particular avidity.

Economy. The density of the towns in no way implies the predominance of trade or industry; all but a few were residences of both landowners and peasants, and their prosperity depended on agriculture. By the 1st century AD, African exports of grain provided two-thirds of the needs of the city of Rome. Some of this, for distribution by the emperors to the urban proletariat, came from the imperial estates and from taxes, but much went on to the open market. An estimate of grain production of something over 1,000,000 tons, of which 25 percent was exported, has been made. Areas of grain production were the Cap Bon peninsula, the valleys of the Miliana and Majardah, and tracts of relatively level land north of a line from Sitifis to Madauros (M'Daourouch). Cereal crops were the most important in the above areas, but fruit, figs, grapes, and beans were also produced.

The production of olive oil became almost as important as cereals by the 2nd century AD, particularly in southern Tunisia and along the northern slopes of the Aurès and Bou Taleb mountains. By the 4th century Africa exported oil to all parts of the empire. Successful cultivation of olives demanded careful management of available water, and the archaeological evidence indicates that much attention was paid to irrigation in the Roman period.



Roman Africa in the 3rd century AD.
Adapted from J.D. Fage, *An Atlas of African History* (1966); Edward Arnold Publishers

Livestock was an important part of the economy of Roman Africa, though direct evidence is slight. African horses were used in racing and no doubt also in the Roman cavalry. Cattle, sheep, pigs, goats, and mules were also raised. Africa was the major source of the wild animals for shows in Rome and other major cities of the empire, in particular panthers, lions, elephants, and monkeys. Fishing, which had been developed along the coast as far as the Atlantic in the Carthaginian period, continued to flourish. Timber from the forests along part of the north coast, and marble, the most important North African source being Simitthu (Shimṭū), were also exported.

There were no large-scale industries even by ancient standards in North Africa, though traditional arts and crafts were practiced. Pottery flourished, and lamps of North African manufacture were marketed in the northern provinces of the empire. Mosaic pavements were extremely popular among the wealthy throughout North Africa, and well over 2,000 have been discovered, with enormous variations in quality. The majority were made by local craftsmen, though some of the designs originated elsewhere. It is also clear that the building trades were major consumers of both skilled and unskilled labour.

Population increases

Prosperity under the Romans undoubtedly led to a rise in the population of the Maghrib in the first two centuries AD; in the absence of reliable statistics, estimates have varied between 4,000,000 and 8,000,000 (the latter being also the population about a century ago). The most recent study proposed about 6,500,000, of whom about 2,500,000 were in present Tunisia. Some 40 percent (but perhaps more) lived in the towns. Of these, Carthage was in a class of its own, reaching at least 250,000. The next largest was Leptis Magna (80,000), followed by Hadrumetum, Thysdrus (el-Djem), Hippo Regius, and Cirta, with between 20,000 and 30,000 each. Many towns in close proximity to each other, especially in the Majardah Valley, averaged between 5,000 and 10,000.

The road system in Roman Africa was the most complete of any western province; a total of some 12,500 miles has been supposed, though only a small proportion was fully surfaced. In origin most roads were military but were open to commerce, and a number of minor roads linking towns off the main routes were built by the local communities. The main arteries were: Carthage to Theveste; Carthage to Cirta through Sicca Veneria; Theveste to Tacapae through Capsa; Theveste to Lambaesis; Cirta to Sitifis; Cirta to Rusicade; and Cirta to Hippo Regius. Carthage handled by far the greatest volume of overseas official traffic and trade, being the natural port for the wealthiest area of North Africa. Nevertheless, most of the ports originally founded by Phoenicians and Carthaginians expanded during the Roman period; in view of the high costs of land transport, it was natural that agricultural products would go to the nearest port for shipment.

Later Roman Empire. The whole Roman Empire underwent a grave crisis between the death of Alexander Severus (235) and the accession of Diocletian (284), resulting from outside attacks, internecine wars, and a total collapse of the monetary system. Africa suffered less than most parts of the empire from the first two factors, though there was an unsuccessful revolt by landowners in AD 238, against the fiscal policies of the emperor Maximinus Thrax, which ended in widespread pillage. There were tribal revolts in the Mauretanian mountains in 253–254, 260, and 288, and the situation finally brought a visit from the emperor Maximian in 297–298, but the revolts had little effect on urbanized areas. On the other hand, the towns were injured by economic difficulties and inflation, and building activity almost ceased. There was some return of confidence under Diocletian (284–305) and Constantine (312–337). Administrative changes introduced at this time included the division of the province of Africa into three separate provinces: Tripolitania (capital Leptis Magna), covering the western part of Libya; Byzacena, covering southern Tunisia and governed from Hadrumetum; while the northern part of Tunisia retained the name Africa. In addition, the eastern part of Mauretania Caesariensis beame a separate province (capital Sitifis). In the far west, the Romans gave up much of Mauretania Tingitana, including the important town of Volubilis, apparently because of pressure from the tribe of the Baquates. In the general reorganization of the Roman army by Diocletian and Constantine, the field army (*comitatenses*) in Africa, numbering on paper some 21,000 men, was put under a new commander, the *comes Africae*, independent of the provincial governors. Only the governors of Tripolitania and of Mauretania Caesariensis also had troops at their disposal, but these were second-line soldiers, or *limitanei*. The whole frontier region along the desert and mountain fringes was divided into sectors and garrisoned by *limitanei*. The latter were locally recruited and closely identified with the farming population of their areas. The Tripolitanian Jabal, which was increasingly exposed to nomad attacks by the Austuriani, is notable for a large number of fortified farms.

Reforms under Diocletian

Africa, like the rest of the empire, experienced the economic difficulties and governmental pressures that were a feature of the later Roman Empire. The power of the landowners increased at the expense of their tenants and of smaller farmers, both of whom the imperial government sought to bind to the soil in a state of quasi-serfdom. In the cities, the tasks of local government that had earlier been eagerly undertaken by the wealthy became burdensome, and again the imperial government sought to make them compulsory and hereditary; while the

councillors themselves sought by any means to enter the imperial administration or professions that provided immunity. The process is well attested in Africa. Nevertheless, urban life withstood these pressures better in Africa than in the west generally.

Christianity and the Donatist controversy. Christianity grew much more rapidly in Africa than in any other western province. It was firmly established in Carthage and other Tunisian towns by the 3rd century and had produced its own local martyrs and an outstanding apologist in Tertullian (*c.* 160–240). During the next 50 years there was a remarkable expansion; over 80 bishops attended a council at Carthage in 256, some from the distant frontier regions of Numidia. Cyprian, bishop of Carthage from 248 until his martyrdom in 258, was another figure whose writing, like that of Tertullian, was of lasting influence on Latin Christianity. During the next half-century, its spread was primarily in Numidia (at least 70 bishops known in 312). The reasons for its exceptionally rapid growth are disputed. In northern Tunisia, urban communities provided a similar social and economic environment to that in which it had first spread in Anatolia and Syria, but this hardly accounts for Christianity's early acceptance in distant parts of Numidia. The existing religious situation may be part of the explanation; in the intermingling of religious currents of Libyan, Carthaginian, and Roman origin in the first two centuries, the cult of Baal Hammon, now under the Roman name of Saturn, became increasingly prominent and may have facilitated a transition to a monotheistic religion. It is certain also that African Christianity always included a vigorous and fanatical element that must have had its effect in spreading the new religion, even though there is little evidence of positive missionary efforts.

Christians were still a minority at the end of the 3rd century, particularly among the wealthy and educated classes; but they were in a good position to benefit from Constantine's adoption of the religion and his grants of various privileges to the clergy. At this time (AD 313), a division occurred among the African Christians that lasted over a century. Some Numidian bishops objected to Caecilian, a newly chosen bishop of Carthage, on the ground that his ordination had been performed by a bishop who had weakened during Diocletian's persecution of the church and hence was invalid. They consecrated a rival bishop, and, when he died, another named Donatus, who gave his name to the ensuing schism. The imperial government recognized Caecilian as the true bishop of Carthage, and those in communion with him as the Catholic Church in Africa and hence alone the recipients of imperial favour. A series of appeals by the Donatists to Constantine resulted in investigations and judgments, all of which went against them. Confiscation of their churches led to some deaths, the victims being honoured as martyrs, but in 322 Constantine rejected further pressure. The Donatists increased rapidly, and for the rest of the century probably equalled the Catholics. The Donatists were strongest in Numidia and Mauretania Sitifensis, the Catholics in the proconsular province; the position in the Mauretanias was more even, but Christianity did not spread rapidly there until the 5th century. In 347 the emperor Constans exiled a number of bishops and took strong measures against the *circumcelliones*, an obscure group of either wandering religious fanatics or seasonal farm workers who were particularly enthusiastic Donatists. But in 362 Julian the Apostate allowed the return of the exiles, who were welcomed with enthusiasm, and the movement proved as strong as ever. Some Donatists appear to have been associated with the revolt of a Mauretanian chieftain, Firmus, and in 377 the first of a series of laws condemning Donatism was issued. Nevertheless, these laws were only sporadically enforced, partly because the provincial governors and many of the local magistrates were still pagan and, at a time of growing weakness in the imperial government, were inclined to ignore instructions they found unwelcome. Donatism was further supported by Gildo, brother of Firmus and *comes Africae* 387–397. Then Augustine of Hippo Regius applied his enormous powers of leadership and persuasion in stimulating the Catholics to resolute action, evolving at the same time a theory of the right of orthodox Christian rulers to use force against schismatics and heretics. In 411 an imperial commission summoned a conference at Carthage to establish religious unity; the Donatists had to obey, though the decision against them was a foregone conclusion. The laws that followed their condemnation were more generally enforced and, though there was some resistance, broke the heresy as a powerful movement; some communities still existed in the 6th century, however.

Beginning of schism

Much controversy surrounds the interpretation of the significance of Donatism; an important view considers it in some sense a national or social movement. It is said to have been particularly associated with the rural population of less romanized areas and with the poorer classes in the towns, whereas Catholicism was the religion of the romanized upper classes. The identification of the imperial government with the Catholics would have intensified the strength of the movement, and the *circumcelliones*' violence constituted incipient peasant revolt. Thus the movement is claimed as analogous to Monophysitism in Egypt and Syria, which produced a vernacular literature and a passive rejection of Greco-Roman culture. The hostility of the Donatists to the existing society was typified by Donatus' remark: "What has the emperor to do with the church?" Against this view it may be said that Donatism in the strictly unromanized tribal areas was certainly weak, and the relationship of the sect with Firmus and Gildo was of little importance. In Numidia it was at least as strong in the towns as in the rural areas, and in any case the distinction between the two is exaggerated. The entire controversy was conducted in Latin, and no vernacular literature was produced; in fact, until the time of Augustine, most of the educated class, of the same social background as Augustine himself and fully imbued with Roman tradition, were Donatists if they were not pagan. It was precisely the reluctance of the landowners to have their peasants disturbed, and the negligence of many provincial governors (both attacked by Augustine), that long protected the Donatists. Lastly, in spite of the remark attributed to Donatus, there is no evidence that the movement attacked the imperial system as a whole, as opposed to individual emperors and officials, and it made full use of its many opportunities to defend itself at law both against the Catholics and against divisions in its own ranks.

Donatism as national movement

Nevertheless, although it is difficult to sustain the view that Donatism, especially in Numidia, represented in some way a resurgence of local pre-Roman culture, and still more hypothetical to suggest that something similar led to the emergence of heretical movements of Islām in the same areas, Donatism certainly appealed to deep-seated traditions of African Christianity. Its fanatical devotion to the memory of martyrs, its doctrinal conservatism and total refusal to compromise on its claim to be the true church while its opponents were contaminated by the stain of weakness in the persecutions, were fully in line with the heroic days of Tertullian and Cyprian.

Extent of romanization. The question whether Roman civilization in the Maghrib was a superficial phenomenon affecting only a small minority of the population who were economically successful, or whether it had profound effects on a majority, is similarly disputed. A priori the former view may be supported by the fact that whereas Gaul and Spain emerged from the Dark Ages with a language and religion derived from their Roman past, in the Maghrib both disappeared, arguably because they were superficial. It is not disputed that in the mountainous areas, such as the Aurès, Kabylie, and Atlas, native Libyan language and culture continued little affected by Roman civilization, though the majority appear to have been Christian by the 7th century; nor that Libyan and Carthaginian traditions survived in other areas and affected the modes of acceptance of Roman civilization. As regards language, the late form of Phoenician known as Neo-Punic was still spoken fairly widely in the 4th century; for example, in the hills near Hippo Regius. Inscriptions in the language and script occurred often at

the beginning of the Roman period but are very rare after the end of the 1st century AD. An exception may be in Tripolitania, where a form of Neo-Punic was inscribed in Latin perhaps as late as the 4th century. There was also a Libyan script known solely from funerary stelae and akin to the script of the present Tuareg; it was known in some sense over most of the Maghrib but may not have been used later than the 3rd century. On the other hand, there is no evidence that the languages were ever literary languages, and the inscriptions are negligible in number compared with the Latin. It may also be observed that the areas in which Libyan inscriptions occur do not correspond with the later areas of Berber dialects. The Latin language unquestionably became general through the whole Maghrib, though to a limited extent in the mountains; it is impossible to define any precise social level at which it was unknown. There is a good deal to be said for the view that the spread of Christianity, whether Catholic or Donatist, completed the victory of Latin among elements which up to that time had perhaps still not used it. Further, there is no evidence for the survival of Neo-Punic or Libyan after the 4th century outside the mountains. Similarly in Gaul, the native Gallic language does not seem to have lasted beyond the 5th century. The fact that there and in Spain Romance languages and Christianity survived, while in North Africa they did not, was perhaps not so much due to a greater depth of romanization as to subsequent historical developments.

Victory of Latin language

The Vandal conquest. The effect of the Donatist controversy on the economy and administration of the African provinces cannot be measured. At the very moment of the effective victory of the Catholics, the rest of the Roman Empire was crumbling to ruin. In 406 the Rhine was crossed by Vandals, Alani, Sueves, and others who overran most of Gaul and Spain within the next few years. In 408 Alaric and the Visigoths invaded Italy and in 410 sacked Rome. Although the empire in the west survived for some time longer, the emperors were increasingly at the mercy of their barbarian generals. Meanwhile, large tracts of imperial territory were lost by the settlement of the invading tribes. Africa escaped for a while, though only death prevented Alaric from leading his tribe across the Mediterranean. Its retention became ever more vital to the survival of what was left of imperial authority. In this situation the *comites Africae* were increasingly tempted to intrigue for their own advantage. One of them, Bonifacius, is said to have invited the Vandals, who at the time were occupying Andalusia, to his aid, but it is more likely that the Vandals were attracted to Africa by its wealth and needed no such formal excuse. Led by their king Gaiseric, the whole people, 80,000 in all, crossed into Africa in 429 and in the next year advanced with little opposition to Hippo Regius, which they took after a siege during which St. Augustine died. After defeating the imperial forces near Calam, they overran most of the country, though not all of the fortified cities. An agreement made in 435 allotted Numidia and Mauretania Sitifensis to the Vandals, but in 439 Gaiseric took and pillaged Carthage and the rest of the province of Africa. A further treaty with the imperial government (442) established the Vandals in Africa Proconsularis, Byzacena, Tripolitania, and Numidia as far west as Cirta.

Vandal impact on the economy

Although the Vandals were probably no more deliberately destructive than other German invaders (the notion of "vandalism" stems from the 18th century), they accelerated the economic decline that had already set in. The imperial authorities had to reduce the taxes of Mauretania by seven-eighths as a result of their devastation. Over much of northern Tunisia, landowners were expelled and the properties handed over to Vandals. Although the agricultural system remained based on the peasants bound to the soil, the expulsions had a serious effect on the towns with which the landowners had been connected. The Vandals, like other invading tribes except the Franks, were divided from the mass of their subjects by their Arian heresy. Although their persecution of the Catholics was exaggerated by the latter, Vandal kings certainly exercised more pressure than others. This was no doubt in reaction to the vigour of African Christianity, which kept the loyalty even of those who had little to lose by the substitution of a Vandal for a Roman landlord.

Gaiseric was perhaps the most perceptive barbarian king of the 5th century in realizing the total weakness of the empire. He rejected the policy of formal alliance with it and from 455 used his large merchant fleet to dominate the western Mediterranean. Rome was sacked, the Balearics, Corsica, Sardinia, and part of Sicily were occupied, and the coasts of Dalmatia and Greece were plundered. Although there was no general cessation of trade, this activity accelerated the breakup of the economic unity of the western Mediterranean, already threatened by the creation of the barbarian kingdoms. Gaiseric's successors were less formidable; Huneric (477–484) launched a general persecution on the Catholics, apparently from genuine religious fanaticism rather than for political reasons, but his successor adopted a milder policy. Under Thrasamund (496–523) there is evidence that many Vandals adopted Roman culture, but the tribe retained its identity until the Byzantine reconquest.

A significant development of the Vandal period was the emergence of independent kingdoms in the mountainous and desert areas, largely of Libyan character. They appeared first in the Mauretanias, where the Roman frontier, already drawn back under Diocletian, receded further under the Vandal kings. By the end of Vandal rule, independent kingdoms existed in the region of Altava (Lamoricière), in the Ouarsenis and the Hodna. After 480, towns to the north of the Aurès, such as Thamugadi, Bagai, and Theveste, were sacked by the inhabitants of another kingdom in the Aurès. All of the names of the known chieftains are Libyan in character, though the survival of romanized elements within some of the kingdoms is attested by epitaphs in Latin and the use of Roman names. Finally, as a harbinger of a serious threat to settled life, whether Roman or Libyan, tribes that had retained a nomadic life on the borders of Cyrenaica and Tripolitania and caused much damage in the 4th century, began to push westward and were already a serious threat to the southern parts of Byzacena by the end of Vandal rule.

The Byzantine period. North Africa held an important place in the emperor Justinian's scheme for reuniting the Roman Empire and destroying the Germanic kingdoms. His invasion was undertaken against the advice of his experts (an earlier attempt in 468 had failed disastrously), but his general Belisarius succeeded, though partly through the incompetence of the Vandals. He landed in 533 with only 16,000 men and within a year the Vandal kingdom was destroyed; some of its people were transported to the Orient, others soon merged with the existing population. A new administrative structure was introduced, headed by a praetorian prefect with six subordinate governors for civil matters and a master of soldiers with four subordinate generals. It required some 12 years, however, to pacify Africa, due partly to the resistance of the Mauri to the re-establishment of an ordered government, partly to lack of support to the army in men and money, leading to frequent mutinies. A remarkable program of fortifications, many of which survive, was rapidly built under Belisarius' successor Solomon. Many were garrison forts in the frontier region, which again seems to have run, at least for a while, south of the Aurès, and then northward from Tubunae to Saldae. But many surviving towns in the interior were also equipped with substantial walls; *e.g.*, Thugga and Vaga (Béja, Bājah). After the death of Justinian (565) there were further difficulties with the Mauri, but the most serious damage was done by the nomadic Louata from the Libyan Desert, who on several occasions penetrated far into Tunisia.

Administrative reorganization

Africa shows a number of examples of the massive help given by Justinian in the building, and particularly the decoration, of churches, and the re-establishment of Catholic orthodoxy was widely welcomed, though it inevitably brought the persecution of surviving Donatists. Seriously weakened though it had been under the Vandals, some traces of the vigour of the African Church remained when it led the opposition of the Western churches to the theological policies of emperors at Con-

stantinople; *e.g.*, those of Justinian and also of Heraclius and Constans II immediately before the Arab invasions.

Little is known of the Byzantine period in the Maghrib after the death of Justinian. The power of the military element in the provinces grew, and between 585 and 591 a new official, the exarch, was introduced whose powers were almost viceregal. The economic conditions continued to decline because of the increasing insecurity and also the notorious corruption and extortion of the administration, though whether this was worse in Africa than in other parts of the Byzantine Empire it is impossible to say. It is certain that the population of the towns was only a small proportion of what it had been in the 3rd or even 4th century. The court of Constantinople tended to neglect Africa because of the more immediate dangers on the eastern and Balkan frontiers. Only once in its latest phase was it the scene of an important historical event; in 610 Heraclius, son of the African exarch at the time, sailed from Carthage to Greece in a revolt against the unpopular emperor Phocas, and succeeded him the same year. That Africa was still of some economic importance to the empire was shown in 619; the Persians had overrun much of the east including Egypt, and Africa alone appeared able to sustain an empty treasury and provide recruits. Heraclius thought of leaving Constantinople for Carthage but was prevented by popular feeling in the capital.

Maghrib on the eve of Arab conquest

In view of the lack of evidence for the Byzantine period, and the still greater obscurity surrounding the period of Arab conquest (643–698) and its immediate aftermath, conclusions on the state of the Maghrib at the end of Byzantine rule are speculative. Much of it was in the hands of tribal groups, among which the level of Roman culture was in many cases no doubt negligible. The Byzantine administration was in a sense foreign to the Latin population also, which may have added to the dislike felt for its notorious corruption. Few traces of the former prosperity of the settled areas remained. The Byzantine military forces were inadequate and incompetent, and it is significant that in the period of Arab conquest the most determined resistance came from the Libyan tribes, which frequently appear not so much as allies as leaders in the struggle. The total decay of town life (except in a few places on the east Tunisian coast) and of the ordered agricultural system occurred within a century after the end of Byzantine rule, though some scholars consider that a modicum survived until the invasions of further nomads, the Banū Hilāl, in the 11th century. Latin was still in use on Christian epitaphs at en-Ngila (Tripolitania) and al-Qayrawān (Kairouan) in the 10th and 11th centuries, and other Christian communities are known to have survived as long. The Arab conquest has its part in the widely discussed theory of Henri Pirenne, that the essential break between the ancient and medieval worlds came with the destruction of the unity of the Mediterranean world not by the Germanic but the Arab invasions. It is obvious that this is true of the Maghrib, but there is much to be said for the view that its earlier conquerors, the Vandals, were the major agent in the destruction of the economic unity of the Mediterranean. Whether this is the key to the problem is another matter.

Roman Cyrenaica. Much of the Roman period in Cyrenaica was peaceful. Some Roman immigrants resided there at an early date, and some of the Greeks received Roman citizenship. A famous inscription of 4 BC contains a number of edicts of the emperor Augustus regulating with great fairness the relationship between Roman and non-Roman. The character of its civilization, however, remained entirely Greek. Jews formed a considerable minority group in the province and had their own organizations at Berenice and Cyrene. They took no part in the great revolt of Judaea in AD 66 but in 115 began a formidable rebellion in Cyrene that spread to Egypt. No reason for it is known. It caused great destruction and loss of life, and Hadrian took special measures to reconstruct Cyrene and also sent out some colonists. Peaceful conditions returned, but in 268–269 the Marmaridae, inhabiting the coast between Cyrenaica and Egypt, gave trouble. In the reorganization of the empire by Diocletian, Cyrenaica was separated from Crete and divided into two provinces, Libya Superior, or Pentapolis (capital Ptolemais), and Libya Inferior, or Sicca (capital Paraetonium, Marsā Maṭrūḥ). A regular force was stationed there for the first time under a *dux Libyarum*. At the end of the 4th century, the Austuriani, a nomad tribe that had earlier raided Tripolitania, caused much damage, and Cyrenaica began to suffer from the general decline of security throughout the empire, in this case from desert nomads. A notable phenomenon of the 5th and 6th centuries, as in Tripolitania, is the number of fortified farms, most frequent in the Jabal itself and south of Boreum (Bū Qurādah) and also apparently in the region of Benghazi.

Christianity no doubt spread to Cyrenaica from Egypt. In the 3rd century the bishop of Ptolemais was metropolitan, but by the 4th century the powerful bishops of Alexandria consecrated the local bishops. The best known Cyrenaican is Synesius, a citizen of Cyrene of philosophic tastes who was made bishop of Ptolemais in 410 partly because of his ability to obtain help for his province from the imperial authorities. Under Justinian a number of defensive works were constructed as elsewhere in Africa; *e.g.*, Taucheira, Berenice, Antipyrgos (Tobruk), and Boreum. Recent excavations of a series of churches in the province also reveal the expenditure he devoted to their beautification, in what was a province of minor importance. On the eve of the Arab conquest (AD 643) the general condition of Cyrenaica would appear to have been on a par with most of the other eastern provinces of the empire, or perhaps slightly better, since it was not as greatly disturbed by the divisions within the church.

(B.H.W.)

II. From the Islāmic conquest to 1830

STRUCTURE AND MENTALITY OF THE PRE-ISLAMIC BERBER WORLD

In order to understand the Islāmic conquest of North Africa and how Islāmic life in the country was organized, it is necessary to examine the social structure of the North African world.

In the 7th century, the vast majority of the North African people were probably Berbers who, though of rather diverse physical types, possessed a common background of language and civilization. Society was tribal in structure. The homeland was not the land per se but rather the race; every social grouping conceived itself only as an assembly of men stemming from a common ancestor, frequently fictitious, whose name they continued to bear.

Berber society

Berber society was, and remained, remarkably fragmented within itself; the basic grouping was the segment, which, among a sedentary population, brought together 200 or 300 households in a territory with a radius of about four to six miles. This was a minuscule state—a tiny republic managed by a council of family heads and magistrates of temporary tenure who maintained order and assured the fair distribution of the profits and of the burdens of the small community. Among the nomads the douar (*dūwār*)—the collection of tent dwellers who lived and migrated together—was frequently of even more limited dimensions. The tribe, which comprised about ten subunits, had an assembly, and in time of war it often selected for itself a single chief. The confederations of tribes—except for moving livestock and for the nomads who needed to have living space assured them—were little more than a framework of regrouping for common defense by force or through alliances.

In political formations of variable size and cohesiveness there was profound democratic sentiment and an extreme mistrust of personal power. What mattered was the preservation of freedom and the autonomy of the small social cells by which the Berbers ordered their existence. Opposing factions grouped themselves into two opposite leagues, or *leffs*; thus, in case of oppression by a chief or by neighbours, they were able to call upon their *leff* brethren for help, and the latter would usually impose arbitration. Armed conflicts remained rare and limited. Without attaining the level of a nation-state or of a nation, Berber society attained equilibrium and peace in a fragmentation that did not facilitate the accomplishment of great collective tasks.

ISLAMIC NORTH AFRICA TO C. 1250

The Islāmic conquest and domination by the Umayyad (Banū Umayyah) caliphs. After a swift conquest of Syria, Egypt, Mesopotamia, and Iran, Islām halted before the steppes of Central Asia and the boundaries of India. Byzantium was organizing a belated but effective resistance on the frontier of Anatolia. Thus, blocked on the east and on the south, Islām lost no time in launching new conquests in the western Mediterranean. In spite of difficulties, this task was undertaken again and again, and it finally succeeded. Beginning in 642, Islāmic troops launched attacks from Egypt into Cyrenaica and into Tripolitania; an expedition reached the southern part of Tunisia, where it defeated the Byzantine patrician Gregory. The northern part of the country was not occupied, however, and the Byzantines finally induced the invaders to leave after paying them tribute.

Muslim penetration into North Africa

Conflicts between the caliphs ʿAlī (ruled 656–661) and Muʿāwiyah I (ruled 661–680), and the founding of the Umayyad caliphate by the latter, halted the Islāmic expansion toward the West until 670, when ʿUqbah ibn Nāfiʿ penetrated Tunisia and there founded al-Qayrawān (Kairouan). Toward the year 683 ʿUqbah undertook a large-scale expedition toward the West that brought him as far as Sūs (Sous) in southern Morocco. North Africa appeared to be adapting itself readily to the authority of Islām; but on his return, ʿUqbah was beaten and slain by a Berber chief, Kusaylah, and the Muslim troops were forced to fall back through Cyrenaica.

For 15 years, despite the incessant efforts of the Umayyad caliphate, then at the apogee of its power, the Berbers, under the command of Kusaylah and later of a woman, Kāhinah, offered stubborn resistance in Tunisia and in Constantine (Qusṭanṭinah, in what is now eastern Algeria). Finally, from 703 to 711, Mūsā ibn Nuṣayr, following somewhat the route taken by ʿUqbah, encountered only local resistance and brought all North Africa under the Islāmic domain.

This Muslim conquest was not really an Arab invasion. North Africa received only about 100,000 Arab militiamen, most of them stationed in Ifrīqīyah (Tunisia and eastern Algeria). But these Arab elements proved to be tyrannical. Their cupidity—their exactions and slave levies—led to a mass revolt in the country. The motivation and leadership for this revolt were furnished by a Muslim heresy—Khārijism, which in the East rallied to its side all the enemies of the Umayyad dynasty. The rebellion, originating in Morocco in 740, conquered all North Africa in two years; the caliphate armies were defeated in Morocco, and the Umayyads were barely able to save al-Qayrawān.

Rebellion against Arab occupation

The Umayyad caliphs, already in a stage of decline, could not reconquer North Africa. In 750 they were replaced as caliphs in the East by the ʿAbbāsids, who became the new masters of the Muslim world and transferred the seat of the caliphate to Baghdad.

The North Africans were unable to use this respite to consolidate their victory and to fortify their independence; from 742 to 746 they continued fighting among themselves in the name of the various sects of Khārijism. The ʿAbbāsids succeeded in conquering only Tunisia and the eastern part of Constantine; their governor in al-Qayrawān, Ibrāhīm I ibn al-Aghlab, while apparently remaining the faithful vassal of the caliphs of Baghdad, became independent in fact. Throughout the rest of North Africa, Khārijī kingdoms were founded, but a great many tribes remained independent.

Although the Khārijī revolt brought about the dismemberment of the caliphate in the Western lands, the people in the West remained faithful to Islām. Thenceforth there existed a Muslim West, which intended to seek the paths of political life and of Islāmic civilization.

Formation of North African Islām. After the 9th century, eastern North Africa remained under the dynasty of the Aghlabids (Banū al-Aghlab), a theoretical vassal of the ʿAbbāsids of Baghdad. The rest of the Muslim West —the Maghrib and Spain—was solving by itself and without foreign interventions the problems posed by its political organization in the framework of the new faith.

The Maghrib was inhabited by various political and religious groups. The most important of the new independent kingdoms was that of the Rustamids of Tāhart, which had an annex in the Jabal Nafūsah in Tripolitania; it was Khārijī, as indeed were the kingdoms of Tilimsān (Tlemcen) and of Sijilmāssah. In the Atlantic plains of Morocco, the confederation of the Barghawāṭah rallied to the cause of a particular heresy, diverging from Khārijism as well as from orthodoxy. The Ghumārah of northern Morocco followed the precepts of their prophet Hā-Mīm and constituted a new sect. Idrīs I, a descendant of ʿAlī, founded a Shīʿah kingdom in northern Morocco. This diversity of sects, theoretically and unequally Islāmic, attests to the fact that North Africa was an exemplar of great religious tolerance. Outside these more or less stable and organized states, there often lived a good many tribes and independent confederations that seem to have practiced Khārijism. The three foremost kingdoms of that time—the Idrīsid, the Rustamid, and the Aghlabid—represent three different forms of Muslim organization.

Political and religious fragmentation

The Idrīsid (al-Idrīsiyūn) kingdom (789–926). Idrīs was a descendant of ʿAlī and hence a *sharīf* (noble, or illustrious one). Fleeing the persecution of his people by the ʿAbbāsids, he succeeded in gaining support in northern Morocco, where he founded the Idrīsid kingdom. His son, Idrīs II, founded the capital city of Fez (Fās, modern name Fès). Upon the death of Idrīs II, the kingdom was divided into a series of principalities.

The Idrīsid kingdom was the first in Morocco to derive support from a mixed Arab and Berber central government and militia. The founding of Fez—an Islāmic city inhabited in part by people from al-Qayrawān and by Andalusians that rapidly developed its own mode of life outside the world of tribes and where influences coming from the Orient by way of al-Qayrawān and later by way of Córdoba were welcomed and diffused—played an important role in the history of Morocco. The Idrīsids combatted Khārijism and particularly strove to convert the tribes that had remained pagan. The political efforts of the Idrīsids were confined to the western and eastern parts of Morocco and were ephemeral, but their civilizing task bore vast and enduring consequences.

The Rustamid kingdom of Tāhart (787–911). ʿAbd ar-Raḥmān ibn Rustam entrenched himself in the central Maghrib, where he founded the city of Tāhart. In 787–788, by concluding peace with the ʿAbbāsids, he succeeded in consolidating his kingdom, which, despite the presence of an oriental colony, remained Berber. Frequent conflicts among the clans gave the kingdom a turbulent history. Ibn Rustam showed himself to be quite generous toward the adherents of orthodoxy and of other Khārijī sects. The Rustamid state lived in peace with its neighbours; thus Tāhart was able to become the centre of an active caravan trade and to maintain economic relations with the Orient.

The Aghlabids (Banū al-Aghlab) of Ifrīqīyah (800–909). Whereas the greater part of North Africa sought its path in diverse sects or heresies, Ifrīqīyah was actively organizing its Muslim life within the realm of orthodoxy. The Aghlabids gloriously fulfilled the obligations of a holy war by inflicting upon the Christians the conquest of Sicily. They did not penetrate the Maghrib, however; they contented themselves with ruling over old lands of sedentary population, abounding in cities—namely, Tunisia and the eastern fringe of Constantine. These were the only lands populated extensively by the Arabs. Al-Qayrawān, an Islāmic centre, became an active point of religious life and soon had its own learned men, for the most part of the Mālikī juridical school; the city soon extended its intellectual and religious influence over all North Africa and paved the way for the triumph of orthodoxy.

The influences proceeding from Baghdad and the ʿAbbāsid world were felt in all domains. Aghlabid art—which left a major imprint in the great mosque of al-Qayrawān—mingled more and more oriental forms with the inherited traditions of Rome and the Byzantines. Following the example of the ʿAbbāsid caliphs, whose luxury they wanted to imitate, the Aghlabid emirs ordered the

Aghlabid art and architecture

construction, in the suburbs of al-Qayrawān, of veritable governmental cities, where they erected their palaces. Within the culture of Ifrīqīyah, Aghlabid was a bridgehead from the Orient.

North Africa in the 10th century. After two centuries of Islām, North Africa appeared profoundly transformed. Along with substantial strata of the population, Islāmization won to its side the independent tribes themselves—the minor non-Islāmized groups were only exceptions to the rule. A considerable portion of Berbers still belonged to Islāmic sects, especially to Khārijism, but these divergent attempts at religious proselytizing finally redounded to the advantage of orthodoxy.

Arab colonization remained at a low level. Except for Ifrīqīyah, which had absorbed about 100,000 Arabs in race and language, no small "Eastern" groups existed. Virtually the entire region remained inhabited by Berbers. Linguistic Arabization remained confined to four regions: the Tunisian Sāḥil (Sahel); northern Constantine, between the city and the sea; the country between the coast and Tlemcen; and Fez and a part of northern Morocco in the Idrīsid domain. But the region experienced the attraction of two great centres of Muslim civilization: from the 9th century, that of al-Qayrawān; and after the 10th century, that of Córdoba. Large cities adopted and disseminated a civilization that was authentically Islāmic.

Predominance of sedentary population

The sedentary population greatly exceeded the great camel-riding nomads, who were found only in a pre-Sahara or Sahara zone. The horseback-riding nomads migrated toward the steppes or plains zones and to the high plateaus; but they seemed to be in particular search of good land for settling. All the fertile plains and mountains were inhabited by sedentary peasants, clustered in villages that had fields, gardens, and orchards. The greater part of these rustics lived comfortably and were of a pacific disposition.

Social fragmentation permitted only limited conflicts. Large-scale wars and their ruinous aftermaths came only with the founding and actions of the great empires.

The Fāṭimid crisis. Shīʿism, a branch of Islām, claimed to be the legitimatist sect of the religion; its followers—the Shīʿites—believed that the head of the Muslim community could only be a descendant of the Prophet through his daughter Fāṭimah, spouse of ʿAlī, and rejected the claims of the Sunnī majority to orthodoxy. Under the ʿAbbāsids the ʿAlids became an opposition and were persecuted; Shīʿism organized itself into secret sects.

By its doctrine, Shīʿism was opposed to Khārijism; it had not spread into North Africa except in the realm of the Idrīsids, who undoubtedly were practicing only a mitigated form of Shīʿism. But in Little Kabylie, Abū ʿAbd Allāh, a missionary of ʿUbayd Allāh, the Shīʿī pretender, was warmly received by the Kutāmah and their brethren of the Kabyle race. Abū ʿAbd Allāh then led them into battle against the Aghlabids, who represented both Muslim orthodoxy and ʿAbbāsid power. In 910 he entered al-Qayrawān, putting an end to the Aghlabid dynasty, and shortly afterward he installed there his master, ʿUbayd Allāh.

The new Fāṭimid Empire had its base among the Kabylie Ṣanhājah, who constituted the best element in its military forces. But the sovereigns were "orientals" in source and in mentality; their constant ambition was the conquest of the Muslim "Orient." Several attempts against Egypt failed before the conquest was finally accomplished; in 972 the Fāṭimid ruler al-Muʿizz entrenched himself in his new city of Cairo, leaving the government of his North African domain to his Berber lieutenant, Yūsuf Buluggīn ibn Zīrī.

Before emigrating eastward, however, the Fāṭimids needed to develop an African policy. On many occasions they sought to enlarge their western possessions. They succeeded at times in imposing their authority on central Maghrib and on northern Morocco; but they came into conflict with the Zanātah tribes, traditional enemies of the Ṣanhājah. The Zanātah were supported by the Umayyad caliphs of Córdoba, defenders of Muslim orthodoxy, who had dreaded a Fāṭimid attack against the Andalusian coast and who possessed the skill of constantly regrouping in the Maghrib and establishing a protective curtain of Berber allies. The struggle between the Ṣanhājah, fighting for the Fāṭimids, and the Zanātah partisans of the Umayyads of Córdoba was the 100 Years' War of North Africa.

But, even in their own realm, the Fāṭimids were unable to banish orthodoxy; they were little more than a Shīʿite government at the head of a country that remained, for the most part, Sunnī. Furthermore, the Zanātah who joined the Umayyad alliance rallied to the side of the Sunnī.

The thrust of orthodox opinion was felt by the Berber dynasties of the Zīrids and of the Ḥammādids, who occupied the North African domain of the Fāṭimids. They were led to reject obedience to the caliphs of Cairo and to return officially to Sunnah. This religious conflict was bound to entail burdensome economic and social consequences for North Africa.

The Arab invasion of North Africa. The Fāṭimids, more and more absorbed in eastern affairs, did not possess the means for reconquering the country from which they had departed and from which they had always drawn their best military elements. In 1051 they launched an attack against North Africa, using entire Arab tribes then stationed in Upper Egypt, namely, the Banū Hilāl and the Sulaym, which were followed by the Maʿgil. But these Bedouins could not found an Arab dynasty. They overwhelmed the country, however, which they often ravaged, and they imposed tributes on both the cities and the countryside.

The coastal area of Tunisia having been reduced, the Zīrids maintained themselves until 1148. In 1090–91 the Ḥammādids abandoned their capital, Qalʿah, only to regroup themselves at the port of Bejaïa (Bougie).

In the middle of the 12th century the Bedouin inundation had reached the region of Ṣatīf (Sétif) in the West. Then and there the North African dynasties collided with the Arab problem without achieving satisfactory solutions.

Reunification of the Maghrib

The great Moroccan empires. A Berber confederation, which comprised the Lamtūna, the Gudālah, and the Massūfah, occupied Mauritania and the countries south of the "loop" of the Niger; it controlled a part of the caravan routes of the Atlantic Sahara and in the 9th century achieved victories over the black kingdoms. This Mauritanian confederation then experienced an eclipse but revived again in the 11th century. Its chief, Yaḥyā ibn Ibrāhīm, made a pilgrimage to Mecca and brought back a scholar from Morocco, ʿAbd Allāh ibn Yāsīn, to improve his people's rather summary Islām. The missionary attempt of ʿAbd Allāh ibn Yāsīn at first achieved success; but when opposition developed he resorted to force, and was supported by a Ṣanhājah chief, ʿAbd Allāh, whose partisans, called the Almoravids (al-Murābiṭūn, or saints), considered their action a holy war.

The Almoravids, in response to an appeal from their racial brethren of the Tāfīlālt against the Zanātah, who imposed their authority on the Moroccan oases, fled back toward North Africa, whence they had departed a century before; and there they were going to change the destiny of the Muslim West.

The Almoravid (al-Murābiṭūn) empire. The Almoravids, between 1054 and 1059, conquered southern Morocco, where they founded Marrakesh, base of operations for future conquests and thenceforth capital of the Almoravid empire. ʿAbd Allāh ibn Yāsīn had been slain in combat. Yūsuf ibn Tāshufīn (Tāshfīn) became the sole chief of the movement, and from 1062 to 1092 he conquered northern Morocco and the Maghrib as far as Algiers (al-Jazāʾir). For the first time since the Khārijī revolt, the Maghrib territories were united under a single power of Saharan origin but rooted in Morocco. In theory the Almoravids represented a religious reform that was nothing more than an "outbidding" of orthodoxy; they were the champions of Sunnism and of Mālikism (a school of Islāmic law), which had already attained a dominant role in the Maghrib.

Muslim Spain was beginning to knuckle under to the local Christians; the kings from the north were imposing costly protectorates upon the *amīr*s, who had divided

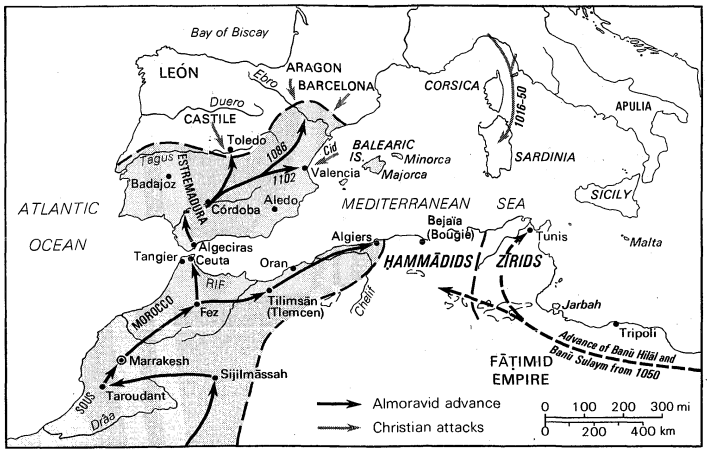


Figure 3: The Almoravid Empire.
Adapted from J.D. Fage, *An Atlas of African History* (1966); Edward Arnold Publishers

among themselves the domain of the Caliphate of Córdoba. In 1085 Alfonso VI of Castile and León entered Toledo and annexed the major portion of what was to constitute the new Castile. The *amīr*s appealed to the Almoravids, and Yūsuf ibn Tāshufīn disembarked with a Berber army and, in 1086, won a victory over Alfonso VI at az-Zallāqah, in Estremadura. In 1068 Tāshufīn tried to check a Castilian thrust into eastern Andalusia; feebly supported by the *amīr*s, he tried unsuccessfully to seize the Christian fortress of Aledo and was obliged to retreat to Morocco. Several of his old allies fell again under Christian vassalage. Having arrived on the scene as the saviour of Spanish Islām, Yūsuf ibn Tāshufīn became its conqueror in 1091 by deposing all the local rulers. The Cid, the famous Spanish captain, halted the Almoravid forces before Valencia until his death in 1099; by 1103 the Almoravids were masters of all of Muslim Spain and several times had triumphed over the Castilians.

Relations with Muslim Spain

From Algiers to Castile, Yūsuf ibn Tāshufīn was the undisputed sovereign of the Muslim West. The Christian reconquest was halted for 20 years. The Almoravids achieved completely the type of domination analyzed subsequently by the Arab historian Ibn Khaldūn (1337–1406)—namely, a clan at the head of an empire, of which it was the vital force and the beneficiary.

A close symbiosis was thenceforth established between the Maghrib and Muslim Spain. Undoubtedly the Spanish Muslims were never thoroughly reconciled to African domination, but they had to bend to the service of their new masters. The great beneficiaries of this political union were the cities in Morocco, which were won over by the Andalusian civilization, of which they became secondary focuses. The Almoravid sultans erected in the Maghrib very beautiful monuments, where Spanish Muslim art reigned unrivalled.

The empire of the Almoravids, founded in Morocco as in Spain on a minority of conquerors, maintained itself under a paradoxical formula—to array Berber armies against the Christians from Spain, but in Morocco to reinforce their own troops with Christian mercenaries. In spite of the minor advances by the Christian forces, Almoravid power maintained itself in the Iberian Peninsula; but after 22 years of fierce fighting in the Maghrib, it succumbed in the face of the revolt of another family of Berber tribes—the Maṣmūdah of the Moroccan Atlas, who themselves fought in the name of religious reform.

The Almoravids had implanted traditions, however, which subsequent dynasties inherited: they paralleled a Berber army staff with a central government of the Hispanic type. But orthodox Islām as well as Mālikī Islām, whose triumph they consolidated, were not transformed and improved by their action—they had Mālikī jurists of an extremely conservative frame of mind, under whose influence the Almoravids enforced in Spain a veritable Muslim inquisition and sought to banish movements that were in quest of genuine spirituality, thus remaining outside what was best in the religious life of their times.

The Almohad revolt

The Almohad (al-Muwaḥḥidūn) empire. The revolt that crushed the Almoravid empire arose in the mountain that borders on the horizon of Marrakesh, the Great Atlas. A Berber from the Atlas, Muḥammad ibn Tūmart, permeated by an intensely religious fervour, was an avowed enemy of Mālikī formalism, which was triumphing under the Almoravids. From the Tunisian port of Mahdīyah, he began a slow "trek" to Morocco; he built around himself a group of faithful disciples and began to preach his own doctrine. He demanded above all a rigorous conception of divine unity, whence was derived the name al-Muwaḥḥidūn (the "unitarians"), which his partisans adopted; at the same time he imposed puritanical reforms in the customs. Upon his return to Marrakesh, he lost no time in coming into conflict with official circles. On the verge of being arrested, he fled into the Atlas, where he raised the standard of rebellion against the Almoravids, whom he regarded as heretics who must be fought with fire and sword by all the tribes of his race—the Maṣmūdah from the mountain. Muḥammad ibn Tūmart preached the Almohad doctrine as well as urging rebellion, which soon encompassed the Anti-Atlas and the Great Atlas.

The Almoravids proved incapable of subduing this revolt, which was shielded by the mountains, but a first thrust by the Almohads against Marrakesh failed. Muḥammad ibn Tūmart died in 1130; his favourite disciple and successor, ʿAbd al-Muʾmin, after a 17-year campaign, finally succeeded in gaining possession of both northern and eastern Morocco. He defeated the Almoravids at Tilimsān (Tlemcen) and returned to Marrakesh, which he took by storm, thus putting an end to Almoravid domination.

The apogee of Berber Islām

ʿAbd al-Muʾmin was bent on expanding his domain. In two campaigns (1151 and 1158–59) he brought all North Africa under his control. This constituted the apogee of Berber Islām—a Berber was ruling with all the dignity of a caliph over all his racial brethren.

Upon returning from his first campaign, however, ʿAbd al-Muʾmin collided in the Ṣatīf region (in Constantine) with a powerful Arab coalition of tribes. He defeated them but found himself incapable of putting an end to the creeping invasion that, for an entire century, was undermining eastern Barbary. He attempted to rally the Arabs for a holy war in Spain and also in order to consolidate his dynasty. But the Arab problem became increasingly critical for his successors. Under the third Almohad ruler, Abū Yūsuf Yaʿqūb al-Manṣūr (reigned 1184–99), Almoravid chiefs from the Balearic Islands, who had disembarked at Bougie in 1184, rallied all the Arabs around them. In spite of vigorous resistance, repeated bloody revolts ravaged the east and the centre of the empire for 20 years. It became necessary to send to Tunis, with sovereign powers, a great Almohad personality, Abū Muḥammad ibn Abī Ḥafṣ, to end the uprising. But deportations of Arab tribes to Morocco served only to open up the whole Barbary to those who proved themselves to be the worst dissidents and malcontents.

By succeeding the Almoravids, the Almohads were obliged to lead the conflict against the Christians and,

Adapted from J.D. Fage, *An Atlas of African History* (1966); Edward Arnold Publishers

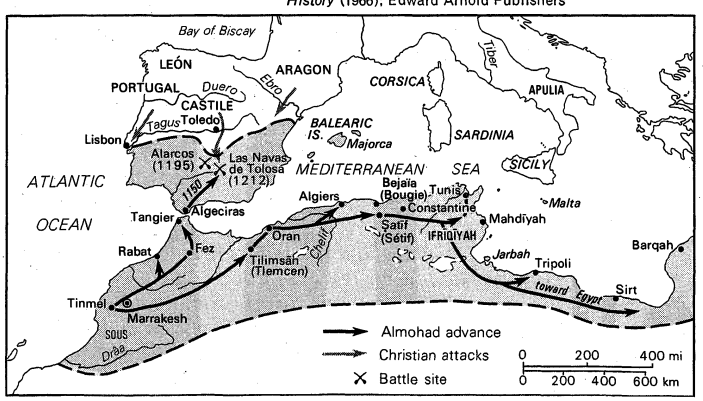


Figure 4: The Almohad Empire.

consequently, become in turn the masters of Muslim Spain. The Christian thrust came to be felt more and more. On several occasions the Almohads dispatched into Andalusia numerous troops from Morocco. Despite the magnitude of this effort, the Christians recorded gradual advances. A great Almohad victory at Alarcos (1195) by Ya'qūb al-Manṣūr led the Christian kings to organize themselves for a crusade. In 1212, at Las Navas de Tolosa, the Almohads were badly beaten. This was followed by a swift decline of the dynasty; and the great Christian reconquest, delayed for a moment, was unloosed after 1230.

The Almohads made no attempt to redress the situation in Spain. Anarchy reigned in Marrakesh, while the Arabs, by that time scattered throughout the plains and plateaus of Barbary, were ruining the country and participating in all the internal conflicts. It was a Berber-Zanātah push, however, by the Banū Marīn, that engaged the last Almohads in battle and in 1269 gained possession of Marrakesh.

Achievements of the Almohads

The efforts of the Almohads had achieved a scope and at times a success that no other dynasty of the Muslim West ever attained. And the Almohad hereditary traditions—their errors as well as their successes—continued to have an impact on the Barbary for several centuries to come. Their reign signalizes the apogee of Berber Islām. But they did not settle the Arab problem, nor were they able to utilize to good advantage the Bedouin tribes, whose expansion toward the West they helped. On the religious plane the Almohad reform movement failed. Still, the Almohads caused Spanish and North African Islām to move in a more liberal milieu, and they favoured a great mystic movement, Ṣūfism, which had penetrated the West before their reign.

Like the Almoravids, they were the devoted servants of Andalusian civilization; by virtue of their struggle against the Christians they gave it a respite of a century. In the arts, these ancient puritans at times invested their mosques with remarkable decorative restraint, which made possible the development of a style of classic purity; and they created some of the most beautiful monuments of Islāmic art.

Adapted from J.D. Fage, *An Atlas of African History* (1966); Edward Arnold Publishers

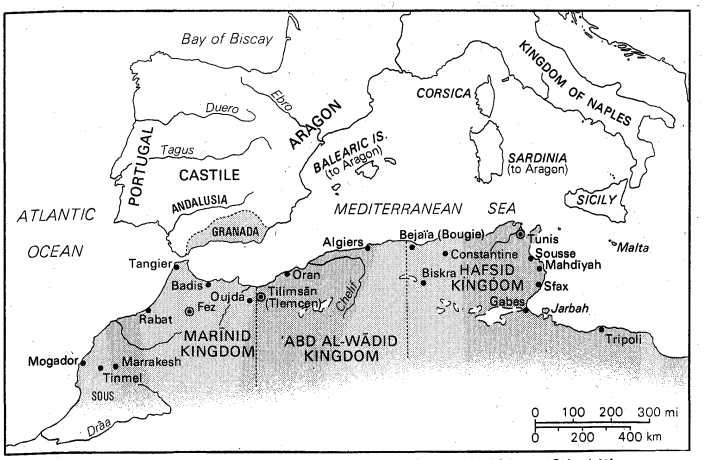


Figure 5: Last medieval dynasties in North Africa, 13th–14th centuries.

FROM THE 13TH CENTURY TO THE BEGINNINGS OF EUROPEAN DOMINATION (1830)

The east medieval dynasties of North Africa (13th–15th centuries). The decline of the Almohad empire can be attributed largely to the mediocrity of its rulers and its government. Anarchy reigned in Marrakesh, where the viziers were fighting each other, making and unmaking sultans amidst the mounting ruins of empire. Three kingdoms divided North Africa among themselves. In Tunisia a kingdom developed under the dynasty of the Ḥafṣids (Banū Ḥafṣ), descendants of the governor, Muḥammad ibn Abī Ḥafs. The founding of the two other kingdoms was due to Zanātah-Berber groups; the 'Abd al-Wādids (Banū 'Abd al-Wād) organized the kingdom of Tilimsān,

The Ḥafṣids, 'Abd al-Wādids, and Marīnids

while the Marīnids first settled in east Morocco, accomplished the conquest of the north, took Fez as their capital, and crushed the last Almohads at Marrakesh (1269).

The Arab problem. These three kingdoms had Berber dynasties at their head; but their politics were constantly dominated by the Arab problem. The Bedouins had by this time invaded all the plains and all the transitional areas; Ma'qil groups had spread throughout the pre-Sahara zone. Although undisciplined and militant, the Arab tribes, from the 13th to the 15th centuries, furnished the bulk of the military forces available to the sultans. Their top leaders joined the government and, through marriage or adoption, were often allied with the royal families. Above all, the Arabs exerted a disastrous influence on the economic life. Even in the most fertile plains they practiced a paradoxical nomadism, into which they dragged along the Berber population, which gradually amalgamated with them.

The Arabs' arrival was frequently accompanied by devastation; in every possible way they undermined the sedentary peasant way of life, which had constituted the strength and good fortune of North Africa in the first centuries of Islām as in ancient times. Roving tents replaced villages and rural hamlets. Gardens and orchards were abandoned while the forests retreated. Only the mountains escaped these disastrous transformations. The ancient and normal economy of North Africa was perverted into swift impoverishment.

It is against this background of progressive ruin, frequently one of disorder, that the histories of the Marīnids, the 'Abd al-Wādids, and the Ḥafṣids are inscribed. The 'Abd al-Wādids were the most feeble in the face of the Arabs. The old towns of the Ḥafṣid kingdom presented a certain counterweight to the Bedouin instability. The Marīnids first adopted a policy of resistance and fell back to the Atlantic Sahara of the Ma'qil tribes, who attempted to sweep back into Atlantic Morocco; but they too were submerged by the Bedouin tidal wave.

Political life. In the 12th–14th centuries no outside danger threatened North Africa, which dwelt in almost complete isolation. Economic relations with Europe remained open, but maritime commerce never flourished. Only the Marīnids intervened overseas; but their incursions into Andalusia, at the end of the 13th century and again in the first half of the 14th, served only to aid in the consolidation of the kingdom of Granada (Arabic, Gharnāṭah) without ever setting back the Christian reconquests.

Conflicts were frequent among the three kingdoms. A stubborn racial hatred was incessantly brewing between the 'Abd al-Wādids and the Marīnids. Many a time the Marīnids invaded the kingdom of Tilimsān. On two occasions the Marīnids, at the apogee of their dynasty, became masters of the central Maghrib for a few years; but they were unable to maintain the conquest of Tunisia. Their grand design to remold the Almohad empire failed, and Morocco remained thenceforth within its present boundaries. In the 15th century the military hegemony passed into the hands of the Ḥafṣids, but they too were no longer able to achieve any enduring conquests.

Religious life. The failure of the Almohad reform consolidated the victory of the Mālikī juridical school. But under this entrenchment of official Islām, religious life became transformed profoundly. Muslim mysticism, or Ṣūfism, which had come from the East and had undergone a rapid development in the 12th century in Muslim Spain, won all North Africa and, after the 13th century, resulted in the formation of numerous brotherhoods: holy personages, both living and dead, and the chiefs of the brotherhoods played the role of intercessors. This development of Ṣūfism of the marabout type (from the Arabic word *murābiṭ*) was in general more intense in the Berber lands than in the "Arabianized" plains.

Influence of Ṣūfism

Intellectual life. Intellectual life remained closely linked with Islām, but it was not locally vigorous; for the most part it consisted of an adoption of the genres and style of the Arab literature of Spain. Jurists and chroniclers contented themselves with compiling the works of their predecessors. All this intellectual activity imported

from Spain thrived in rather restricted circles, mainly those of the large cities and the courts of the rulers. Enlivened for a brief time by the arrival of Andalusian refugees in the 13th century, it was not long before it became a victim of ossification. In this entire process there was one magnificent exception—the historian and statesman Ibn Khaldūn, who composed a veritable sociology of Muslim kingdoms of the Middle Ages.

Artistic life. In the 13th century, artistic workshops modelled after those of Muslim Spain succeeded in taking root in Morocco; thus, at the beginning of the 14th century, Hispano-Moorish architecture and decoration appeared to be fully acclimatized in the large cities of the Maghrib. In Tilimsān and in Fez this art can be distinguished only by slight nuances from that of Granada. In Tunisia, the arrival of Andalusian refugees was reflected in Ḥafṣid art through important and enduring Hispanic contributions, which intermingled with local traditions.

This art underwent no further development, and that of Granada itself declined. But North Africa had nevertheless participated in the classical age of Hispano-Moorish art. Morocco preserved this art faithfully without breathing new life into it.

Ottoman North Africa. The extension of the Ottoman Empire into North Africa in the 16th century was not the result of a planned conquest but is attributed to particular and genuinely African factors. At the beginning of the 16th century the Portuguese had already established themselves in various ports of the Straits of Gibraltar and of northern Morocco and were extending their undertakings to Morocco's Atlantic coast. Meanwhile, the Spaniards launched a vigorous effort on the rest of Barbary and occupied Melilla, Peñon de Velez, Mers-al-Kebir, Oran, and the Peñon of Algiers (an islet in the harbour of Algiers). The ʿAbd al-Wādids and the Ḥafṣids at times were obliged to accept the protectorate of Spain.

Khayr ad-Dīn Barbarossa and his successors

The founding of the regency of Algiers. Five brothers, including ʿArūj and Khayr ad-Dīn (nicknamed "Barbarossa"), who came from the island of Lesbos and engaged in piracy in the neighbouring archipelago, decided around 1505 to transfer their undertakings to the western Mediterranean. They achieved swift success, and the Ḥafṣid sultan permitted them to settle on the island of Jarbah (Djerba). The populations conquered or threatened by the Spaniards appealed to these new leaders. ʿArūj established himself in the city of Algiers, then recaptured Tlemcen from the Spaniards. But the latter reoccupied the city and slew ʿArūj.

The possessions of the pirates were seriously threatened. Barbarossa, in order to procure aid, decided to join his possessions with the Ottoman Empire. He rendered homage to the sultan Selim I, who confirmed him in his command with the title of *beylerbey* and dispatched 6,000 men with artillery to him. Thus, in a quasifortuitous fashion, the greater part of North Africa, after centuries of isolation, again became a dependency of an Eastern caliphate.

Thanks to this support, Barbarossa succeeded in extending his dominion toward the east by occupying Bône (Annaba), Collo, and Constantine. In 1529 he drove the Spaniards from the Peñon of Algiers, and Algiers became the capital of Turkish Algeria. In 1534 Barbarossa gained possession of Tunis. But in the following year the Holy Roman emperor Charles V (Charles I of Spain) recaptured the city, erecting a fortress at La Goulette, and again imposed a protectorate over the Ḥafṣid sovereign (Tunisia came under Turkish control again in 1574); Barbarossa regained his prestige by an incursion into Mahon. The sultan then invited him to Istanbul and appointed him as admiral in chief of the Ottoman fleet (*kapudan paşa*).

From 1536 to 1587 the *beylerbeys* (Barbarossa and his successors), residing in Algiers, governed Turkish North Africa; they piled up success after success over the Spaniards and solidified Ottoman domination. After repelling an attack by the first Saʿdī sultan of southern Morocco, the *beylerbeys* intervened in Morocco, where for a certain time they imposed the Waṭṭāsid ʿAlī Abū Ḥassūn, enemy of the Saʿdī, upon Fez. The conquest of Morocco remained among the designs of the *beylerbeys*, but their struggles against the Spaniards absorbed their best fighting units. In 1587 the title of *beylerbey* was abolished, and the three regencies of Algiers, Tunis, and Tripoli were entrusted to different pashas. Turkish North Africa thus merged into the general organization of the Ottoman Empire.

Ottoman administration of Algeria

The structure of the Algerian state was of a military type; it rested at the same time on local privateers and the Janissaries (the Ottoman standing army), both enjoying considerable privileges. The Janissaries constituted an excellent professional army, stationed in barracks, equipped with modern weapons, and thoroughly trained and endowed with an extraordinary esprit de corps. The army was reinforced by the native cavalry (the *sipāhīs*) and by Kabylie contingents from northern Algeria that were as sturdy as the Janissaries.

Barbary pirates were particularly powerful in the 17th century and made Algiers an important city; it became a great slave market. The city, which boasted a population of about 60,000, welcomed some Moors expelled from Spain—namely, the Tagarins.

Power soon slipped out of the hands of pashas, who now enjoyed only honorary roles; it passed into the hands of military leaders (*aghas*) and later into the hands of the *deys* (Janissary commanders). In the 18th century the *deys*, designated by the militia, maintained control of the province of Algiers and presided over the remnants of the central government. Three provincial regions (the *beyliks*) were headed by officials called *beys*. These provinces were divided into cantons commanded by *qāʾīds* (*caids*) or by *shaykhs*. Thus under a Turkish general staff there were the classic divisions of the country into tribes and segments of tribes. Conflicts with the neighbouring regency of Tunis were numerous, but the regency of Algiers did not succeed in any way in modifying in durable fashion its eastern borders.

Despite the harshness manifested by numerous *deys* of Algiers, the *deys* themselves, who were dependent on the militia, were no more than "despots without freedom." Their governments and they themselves frequently had difficulty in exacting taxes from the interior tribes.

The Ottomans belonged to the Ḥanafī juridical school; they at times constructed new mosques, and they had their own particular judges. But Mālikism remained dominant. There was no disturbing competition between the two juridical orthodox rites; likewise, in both of them doctrine and jurisprudence were stabilized. Maraboutism (a branch of Ṣūfism) had not undergone any change, and in Algeria as in Morocco the brotherhoods remained the most vital element in this North African Islām.

Architecture under Ottoman rule underwent little change but accepted a new type of decor—a floral ornament of oriental origin replaced almost everywhere the classic Hispano-Moorish decoration. The art of Turkish Algeria was thus a hybrid art that preserved some medieval forms.

Ottoman domination in Tunisia

Turkish and Ḥusaynid Tunisia. Ottoman domination in Tunisia was established with finality only after 40 years of warfare against the Spaniards, allied with the last Ḥafṣid sultans.

After 1574, Tunisia, which like Algeria had become an Ottoman province, strove to limit the influence of the Turkish government in the internal affairs of the country while at the same time availing itself of the Turks' protection against the Christian and Algerian enemies. This clever policy of a rather weak country was fairly successful. The regency of Tunis, as well as that of Algiers, was at first controlled by pashas appointed for three years, assisted and often dominated by the militia. After 1520 the militia itself appointed its chief, who took on the title of *dey*, with the pasha now playing, as in Algiers, only an honorary role.

The 17th century was likewise in Tunisia a golden age of privateering. Order prevailed throughout the country thanks to the Turkish garrisons, to biannual military rounds that assured the levying of taxes, and also to the old administrative tradition of the country transmitted by the Ḥafṣids. The arrival of Moors expelled from Spain in

1609–12 enabled the Tunisian economy to record important progress both in agriculture and in the crafts.

From 1640 to 1705 the *beys* (provincial governors) exercised effective power and constituted de facto dynasties. After 1705 the Ḥusaynid *beys* formed a veritable local dynasty recognized by the Turkish government. As far as Europe was concerned, the regencies of Algiers and Tunis came to be regarded more and more as being independent.

Whereas Turkish Algeria was deprived of deep unity and at times proved difficult to govern, Ḥusaynid Tunisia presented the aspect of a real state and was partly opened to European influences. A peace-loving country, its government rejoiced in 1830 when the French conquest of Algiers delivered it from a frequently menacing neighbour.

Sharīfian Morocco. In Morocco, from 1428 to 1459, the Marīnid sultans fell under the protection of a dynasty of viziers—the Banū Waṭṭās (Waṭṭāsids)—who finally replaced them. This new dynasty, feeble from its very beginnings, was obeyed only in parts of northern Morocco and was obliged to confront an occidental world in full flush of renaissance.

The Banū Waṭṭās

Morocco was directly affected by Portuguese incursions against its coasts. From 1415 to 1486 these attacks were directed against the ports of the Straits of Gibraltar and of northern Morocco. But from 1486 to 1550 the Portuguese settled at various points on Morocco's Atlantic coast, from Oum er-Rbia to Sous. In the face of these new problems the Banū Waṭṭās remained peaceful; they took virtually no action against the Portuguese. Resistance to these Christian assaults was offered by local chiefs as well as by the brotherhoods. Ṣūfism had continued to develop in Morocco during the entire Marīnid period and had resulted in the formation of numerous brotherhoods. The latter, however, while often enjoying the role of arbiters among the tribes, possessed no political powers. The Christian incursions against the Moroccan coasts and the attempts at penetration into southern Morocco impelled the religious leaders to become the instigators and, at times, the leaders of a holy war. From the 15th century on, marabouts (*murābitūn*; popular Ṣūfī religious leaders) became a political force, often challenging the power of the sultans.

Furthermore, the actions of the Arabs culminated in disastrous consequences; the country was impoverished, and, amidst growing distress, anarchy gained the upper hand. From the time of the Marīnids, the Atlas region and a substantial portion of the Sahara confines had evaded the authority of the sultans. Morocco split into two zones: the Bled Makhzen (land under government control) and the Bled Siba (land of dissidence), where the mountain people organized their own lives while limiting contact with the cities and the plains.

The Banū Waṭṭās struggled for 30 years against the increasing thrust of a Marabout dynasty—that of the Saʿdī *sharīf*s, who finally seized the city of Fez in 1548. A Waṭṭāsid restoration supported by the Turks in Algiers was defeated in 1554.

Saʿdī Morocco (1548–1659). The Saʿdīs had settled since the 13th century in the oasis of the Drâa. Until the 16th century they played no political role in extreme southern Morocco. After the founding of Agadir, the Portuguese attempted to develop an aggressive policy in the Sous region; the marabouts of the Sous resisted and designated a Saʿdī as leader in the holy war, but without notable results. The Saʿdī, however, were summoned against the tribes that were fighting the Portuguese of Safi. Giving priority to their personal ambitions, the Saʿdī took advantage of this situation to conquer southern Morocco and Marrakesh and then came into conflict with the Banū Waṭṭās. The Saʿdī gained new prestige by capturing the Portuguese position at Agadir in 1541, but they needed ten more years to conquer the northern part of the country and to become the sultans of Morocco.

The Saʿdī had risen out of the marabout crisis. But their struggle against the Christians had frequently been no more than a pretext for the conquest of Morocco. Their dynasty had no native support apart from some contingents from the Sous and most often had to use Arab militias.

Saʿdī foreign policy

The Saʿdī foreign policy was dominated by a fear of the Turkish peril; thus, reluctantly and with meagre results, they allied themselves with the Spaniards in Oran, who were battling against the Turks of Algiers. One Saʿdī sultan, dethroned by two of his uncles, had taken refuge in Portugal. The youthful and visionary Portuguese king Don Sebastian (ruled 1557–78) attempted to restore him to power in Morocco, but both he and his protégé were defeated and slain in the "Battle of the Three Kings" (1578). This Saʿdī victory created the impression that Morocco constituted a formidable force by virtue of having defeated in a single encounter a European power. The sultan Aḥmad al-Manṣūr (reigned 1578–1603) adroitly capitalized on this illusion. Aḥmad al-Manṣūr succeeded in reinforcing his army with mercenary Turk and Kabylie units, as well as with emigrants from Andalusia. He had a luxurious palace erected for himself at Marrakesh, called the "Badi," where he received European ambassadors with great pomp. His only expedition abroad was undertaken against a Muslim country, the Sudan, where his troops, largely Andalusian, captured Timbuktu, from which he drew gold and numerous slaves. He procured financial resources by any means within his power. The cultivation of sugarcane and the trade in sugar became his monopoly. His fiscal mismanagement, however, provoked revolts.

Though al-Manṣūr maintained friendly relations with the great European powers, which permitted Morocco to carry on a minimum of foreign trade, he did not open the country to European influences. He jealously guarded his insulation from the Christian world. Having risen out of the marabout crisis, the Saʿdī were destined to die of it. Following the death of al-Manṣūr in 1603, the Saʿdī princes began to dispute among themselves over authority; thenceforth the dynasty experienced a rapid decline. Throughout Morocco, marabout forces arose and fought each other in order to aggrandize their domain. One of them, the marabouts of Dila, seemed for a moment on the verge of restoring the unity of Morocco. But it was a second Sharīfian dynasty, that of the ʿAlawīs, that ended this anarchic fragmentation and succeeded in reuniting under its authority—actual or theoretic—the Moroccan lands.

Morocco under the ʿAlawī (or Filālī) sultans (1689–1830). The ʿAlawī *sharīf*s had for a long time been established in the Tāfīlālt, and with the decline of Saʿdī power they became the political leaders of the oasis. Mawlāy ar-Rashīd gained possession of Fez (1666), conquered all the local rulers who were disputing among themselves over the Moroccan lands, and restored the political unity of the country. He died accidentally after having consolidated his task of reorganization. Nevertheless, the country remained in a state of turbulence. His brother and successor, Mawlāy Ismāʿīl (1672–1727), had to assert his authority by force; with a militia of black slaves and Arab troops, he attempted a veritable military occupation of his own country through garrisons stationed in fortresses. Despite his efforts, Mawlāy Ismāʿīl proved incapable of subjugating effectively mountainous Morocco, which remained in a state of dissidence. He attempted incursions into Tlemcen in vain, but he did succeed in recovering some of the coastal areas still occupied by the Christians—namely, Maʿmūrah, Tangier, and Larache.

Reunification of Morocco

The crushing tax burden imposed by Mawlāy Ismāʿīl, who launched an ambitious military effort and built at Meknès immense palaces, led to numerous revolts among the tribes. Upon his death the country found itself exhausted and prey to sanguinary anarchy for a period of 30 years.

Sīdī Muḥammad ibn ʿAbd Allāh (1757–90) restored order in the country; with him the ʿAlawī dynasty confined its ambitions and defined its policy. The rulers maintained a minimum of business contacts and diplomatic relations with Europe, with the rest of North Africa, and with the Orient.

In the interior the sultans strove to check, but could not

prevent, the incursions of Berber tribes of the Central Atlas and Moyen Atlas into the Atlantic plains. Revolts were common, even in the area under subjection, and the sultans often had to undertake military expeditions in order to assure the levying of taxes.

These events were but the continuation of a series of expedients dating back to the Middle Ages. Despite the goodwill and the application by the sultans of the 18th century and the early 19th century, Morocco proved incapable of taking advantage of its last period of isolation to bring about its own reform.

Thus in Sharīfian Morocco, as in Turkish Algeria and Tunisia, there was no renaissance, no manifestation of a firm will or of clearcut evolution and progress. Revolutionary and imperialistic wars absorbed all the energies of Europe and paradoxically prolonged this North African status of isolation. Transformations could no longer occur except through foreign influences. The seizure of Algiers by the French in 1830 opened a new era; however, Europe's interventions were to be felt in the three countries at different dates and as a result of diverse methods.

Unity and diversity in Muslim North Africa. Within the scope of this article it is helpful to look back over these 11 centuries of Muslim history in North Africa by investigating in what epochs and to what degree the regions that constituted Tunisia, Algeria, and Morocco experienced political unity and a unity of civilization, or alternatively to what extent they followed their autonomous paths in these domains.

Eleven centuries of Muslim history in North Africa

In the 8th and 9th centuries, after the resistance that Barbary had offered against its conquerors (particularly in its eastern portion), the region rallied to Islām and organized its Muslim way of living. This was done either through the path of orthodoxy or through the diverse sects of Khārijism (even in its particular heresies). Except for the Aghlabid *amīr*, a theoretical vassal of Baghdad, this evolution proceeded within an entirely independent policy and even through fragmentation; but this splintering and these differences of detail must not conceal the profound unity of this great movement toward the organization and elaboration of North African Islām.

The transformation resulting from the efforts made during the 5th and 9th centuries made possible the achievements wrought in the 10th, 11th, and 12th centuries by the great African empires. The first among those—that of the Fāṭimids—was, from the Barbary point of view, an inconclusive endeavour, deficient in the long run. It ran contrary to the religious evolution of the country and devoted its utmost effort toward the conquest of Egypt, where the dynasty emigrated in 972 and became Eastern in fact, which it had been at heart previously.

To the contrary, the two great Moroccan dynasties—the Almoravids and the Almohads—laboured at the same time for the political and cultural unity of the Barbary. The Almoravids, after unifying the Moroccan lands, conquered the Maghrib. The Almohads subjected all of North Africa to their control, thus attaining the apogee of Berber Islām. Because they were also the masters of Muslim Spain, the Almoravids and the Almohads were able to further successfully the cultural unity of the Muslim West. Thus they became the *serviteurs* of the Andalusian civilization, which prevailed in the cities of the Maghrib and in Tunisia mingled with the local traditions. The political apogee of the Barbary was accompanied by the unification of its culture.

The disintegration of the Almohad empire served as the foundation, on the political level, for the tripartite division of North Africa; but it did not undermine the unity of its civilization of Spanish origin. In Tunisia the Hispanic influences were never greater than under the Ḥafṣids (13th, 14th, and 15th centuries). To this unity of Muslim culture, however, a common misfortune was added; in the three kingdoms of Tunis, Tlemcen, and Fez, the destructive depredations of the Arabs undermined the power of the dynasties and ruined the traditional economy of the country, which thenceforth became impoverished.

The beginnings of modern times, which in Europe were distinguished by profound and fruitful transformations, failed to inspire North Africa with the desire to change. The Portuguese ventures into Morocco served only to let loose a prolonged marabout crisis, which became deeply xenophobic in character. Under the Sharīfian dynasties, Morocco dwelt in isolation, adamant against innovations; it continued on its traditional path, which was prolonged until the 20th century—a paradoxical "Middle Ages."

It was a fortuitous fact—the policy of a family of pirates who had become the leaders of a holy war—that linked Algeria and Tunisia with the Ottoman Empire, which modified their civilization through a wave of Eastern influences. The political and cultural unity of the Barbary was thenceforth shattered, and the ossification of civilization in the three countries served only to consecrate the division. (H.-L.-É.T.)

III. North Africa since 1830

The restoration of European rule in North Africa was initiated in 1830 by the French conquest of Algiers city and followed in time by the occupation of the hinterland. In 1881 the French occupied Tunisia also, and in 1911 the Italians expelled the Turkish government from Libya, following this up by the progressive conquest of the Arab inhabitants. The process was completed by the Franco-Spanish occupation of Morocco, which followed the signing of the Treaty of Fès in 1912. The regimes of European colonial type that followed created entirely new conditions and by degrees gave birth to modern nationalist movements. Strengthened by the circumstances of World War II, these resulted in the establishment of Libya as an independent state in December 1951 and the independence of Morocco and Tunisia in 1956. In 1960 a new state, Mauritania, was created to the south of Morocco; and, last of all, in 1962 Algeria achieved its independence.

ALGERIA

The period of the French conquest. The French conquest of the regency of Algiers was not the result of a clearly thought out plan. Too little was known about the regency's internal constitution to have permitted such a thing, even if British opposition to the installation in North Africa of a potentially dangerous rival could have been ignored.

Beginnings of French interference

In 1830, when a blockade of Algiers imposed by the French for an alleged insult to the French consul had proved ineffective, the French government decided on a full-scale attack on the city of Algiers. For the benefit of the Arab inhabitants a proclamation was prepared in Arabic and published in French in *Le Moniteur*, stating that the French were coming not to take permanent possession of the city but to expel the Turkish foreigners and make the Arabs masters of their own country; as issued later in Algiers, however, it simply stated that the French were coming to make war on the rulers and not on the people of the country. In addressing the European powers, the French gave as their motive the intention of ending a piratical regime that practiced the enslavement of Europeans. In fact, however, the possession of European slaves had been limited from the end of the 18th century to the state, from which they could be hired for domestic purposes. In any case, the decay of privateering and a British bombardment of Algiers in 1816 had already reduced their number to virtually nothing. In fact, the French decision had been determined by the desire to gratify the officer corps by a military success and thereby to strengthen the restoration regime of Charles X.

An attempt at mediation by the nominally suzerain Ottoman power was brought to naught by a French interception of the Turkish envoy. The collapse of privateering had by then left the regency a mere shadow of its former self and had caused the *dey*'s government to provoke the Arab and Berber population by increasing taxation to make up for the lost income. In consequence, the landing at Sidi Ferruj, 16 miles west of Algiers, on July 5, 1830, of a force of 37,000 French troops led to the capitulation of the *dey*'s government within three weeks. The *dey* was banished, together with the majority of the Turkish officials and volunteer forces. Within a

fortnight of this success, the regime of Charles X was itself overthrown and succeeded by that of Louis-Philippe, which, in the circumstances, was as unwilling as its predecessor had been to define future North African policy. But this did not prevent the new minister of defense, himself a general, from instructing the new commander in chief, Gen. Bertrand Clauzel, to favour colonization. This was an example of the determining role French army officers played in deciding French action in Algeria during the next 40 years.

Left without any clear guidance from Paris, Clauzel decided to exercise control in Algeria through princes of the ruling family in Tunisia, whose ruler was allowed to understand that the whole regency might finally be entrusted to him. This prospect was gratifying to the Tunisian rulers, over whom the Algerian *deys* had made several efforts to establish their supremacy. The Tunisians briefly occupied Oran in 1830 but were unable to hold it and withdrew.

After only five months, Clauzel himself was withdrawn and was succeeded from February to December 1831 by Gen. Pierre Berthezène. Ordered to suspend the incipient colonization, of which he himself disapproved, Berthezène rapidly became unpopular with would-be settlers and the many speculators. He was succeeded by the Duke of Rovigo, who had been chief of police under Napoleon. After Rovigo's withdrawal 14 months later, because of illness, his drastic methods were criticized by a commission of inquiry, which charged some of the occupiers with barbarity. The commissioners concluded, however, that the regency of Algiers for reasons of utility, expediency, and necessity, should be definitely occupied by France. Accordingly, the officer responsible for the administration was known as "governor general of the French possessions in Africa" until 1845, when Gen. T.-R. Bugeaud, after achieving the conquest of the greater part of the country, was named "governor general of Algeria" (the latter name was a novelty introduced to describe the newly acquired territory).

Before this occurred, however, another ten years were to pass, during which the declared aim was to restrict French direct rule to the coastal area while authority inland was left in the hand of Algerians who, it was hoped, would cooperate with the invaders. In the east the outstanding opponent of the French was the *bey* Ahmad of Constantine (Qusṭanṭinah), where a first attempt to capture the city in 1836 failed disastrously. Resistance in the west, organized by the famous Abdelkader (ʿAbd al-Qādir ibn Muḥyī ad-Dīn), developed a more national character and lasted 17 years.

Resistance to the French

Abdelkader, who was a *sharīf*, had been elected in 1832 on the plain outside his native town of Mascara as leader in the struggle against the Christian invasion and greeted by popular acclaim with the title of *amīr*. In this capacity he regarded himself as *khalīfah*, or representative, of the Moroccan sultan Abdurrahman. In 1833 the French, hoping to use Abdelkader as French agent, signed with him a convention that recognized his local position. Four years later the French government made a last attempt to limit its direct control, at least in western and central Algeria, to the coastal areas. General Bugeaud, now commanding in the Oran area, accordingly concluded with Abdelkader the Treaty of Tafna, which allotted Abdelkader Tlemcen and the whole of the western province except for the cities and environs of Oran, Arzew, and Mostaganem. In central Algeria, Abdelkader was ceded Titteri, while the French were limited to the coast and the Mitija plain, and were excluded from Constantine.

The French governor general, the Comte de Damrémont, disapproved of the arrangements and made a second attack on Constantine; this proved successful, though at the cost of his own life. His successor soon came into conflict with Abdelkader, who for the moment, however, was fully occupied in establishing an orderly form of government in the West.

In 1839, however, Abdelkader denounced with some reason French violation of the Treaty of Tafna and then attacked the Mitija, destroying the settlers' farms, until Bugeaud, converted to the idea of total conquest, was given command and supplied with adequate resources from France. A six years' struggle followed, involving the wholesale destruction of Algerian villages, the spreading of desolation far and wide, and the death by hunger of thousands of refugees.

Total conquest of Algeria

In 1844 Abdelkader was forced westward into Moroccan territory, whereupon the sultan Abdurrahman, alarmed at the French advance, sent an army to the frontier. This was ignominiously defeated by Bugeaud on August 14 at Isly near Oujda. Meanwhile, French ships bombarded the Moroccan ports of Tangier and Mogador. The sultan was thus forced to agree to expel Abdelkader or hand him over to the French.

Abdelkader nevertheless fought on for more than two years. In 1847 Bugeaud, now promoted to a marshal of France and designated governor general of Algeria, decided to complete the conquest of the Kabylie Berber area, employing the same methods of devastation as in the west. Though successful, he was forced to resign for having acted without the sanction of the minister of war. His place was taken by the Duc d'Aumale, son of King Louis-Philippe. During the four months during which d'Aumale held the post, the sultan's troops expelled Abdelkader, who had again taken refuge in Morocco, and who finally surrendered with the promise that he would be allowed to live in the Orient.

In spite of this, Abdelkader was detained, first at Pau and then in the château of Amboise, until released by Napoleon III. He was then permitted to live in Damascus, where he saved the lives of a great many Christians during the massacres of 1860. Though Abdelkader's exile marked the end of what may be called resistance on a national scale, smaller operations continued, such as the occupation of the Saharan oases (Zaatcha in 1849, Nara in 1850, and Wargla in 1852). The eastern Kabylie country was only subdued in 1857, while the final great Kabylie rising of Moqrani was suppressed in 1871. The Saharan regions of Touat and Gourara, hitherto Moroccan spheres of influence, were occupied in 1900; the Tindouf area, previously regarded as Moroccan rather than Algerian, was only attached to the latter region after the French occupation of the Anti-Atlas in 1934.

French Algeria. The French thus carried on the policy, initiated by the Turks, of making Algeria the base of their rule in North Africa and of increasing their territory by bringing under Algerian control areas that had hitherto looked rather to the older established regimes of Tunis or Morocco both for spiritual leadership and for defense against outside attack. The final outcome of this process was to be the creation of what was known as French Algeria.

Administration. For the French officer corps the preservation of the conquest became a vocation and a point of honour. In the political field their influence was represented by the governor general, almost invariably drawn until 1880 from the armed forces, and by the Bureaux Arabes, whose members (officers with an intimate knowledge of local affairs and of the language of the people), having no direct financial interest, often sympathized with the outlook of the people they administered rather than with the demands of the European colonists.

European–Muslim relations. A large-scale confiscation of cultivable land, following the crushing of resistance, made colonization possible. By 1880 the coastal area had become a predominantly Christian area of mixed European origin (mainly Spanish in and around Oran; French, Italian, and Maltese in the centre and east). For long the presence of the non-French settlers was officially regarded with alarm; but, with time, the influence of French education, of the Muslim environment, and of the Algerian climate created in the non-French a European-Algerian, subnational sentiment. This would probably have resulted, in time, in a movement to create an independent state if Algeria had been situated farther away from Paris and if the inhabitants had not feared the potential strength of the Muslim majority. As it was, however, each weakening, even temporary, of the authority of the French government led to the

Influence of European settlers

increased influence of the settlers and to a renewed rising and suppression of the Muslims.

On the overthrow of Louis-Philippe's regime in 1848, the settlers succeeded in forcing the newly appointed governor general, Louis-Eugène Cavaignac, to return immediately to France. By the time he was able to come back, a few weeks later, the settlers had achieved their aim of having the territory declared French and the three former Turkish provinces converted into departments on the French model, while colonization was developed with renewed energy.

On the establishment of the French Empire in 1852, responsibility for Algeria was at first transferred from Algiers to a minister in Paris. Very soon, however, the emperor, having formed his own opinion on the problem, reversed this disposition. While expressing the hope that an increased number of settlers would forever keep Algeria French, he also declared that France's first duty was to the 3,000,000 Arabs. He envisaged the Muslims as devoting themselves to agriculture, while the colonists occupied themselves with industry and matters requiring technical skill. With considerable accuracy he declared that Algeria was "not a French province but an Arab country, a European colony, and a French camp."

This attitude aroused certain hopes in Arab minds, but they were destroyed by the Emperor's downfall in 1870. This was followed by a settler demonstration in which the new governor general, Walsin Esterhazy, was compelled to re-embark for France the day after he arrived, and authority was for six months exercised by settler committees.

Meanwhile, the abortive hopes that the Emperor's ideas had raised, followed by a French defeat in the Franco-Prussian war, resulted in the last great Kabylie rising (1871) under Muḥammad al-Moqrani. Its suppression was followed by the sequestration of another 11,000,000 acres of land and the imposition of an indemnity of 36,500,000 francs; these measures together provided land for refugees from Alsace and capital with which to exploit the land.

Domination by European population

During the 50 years that followed, the European population at last felt free to establish over the country and its native inhabitants a political, economic, and social domination, which it believed would last indefinitely. The Muslims had learned that medieval religious faith and unlimited courage could not prevail against modern education and arms. At the same time, new communications, the installation of hospitals and medical services, and modern education—dispensed to a very limited extent and in French to the Muslims, while generally available to the Europeans—created a minority of Algerians of a new type.

For Algerians, service in the army and in the factories of France during World War I was another eye-opening experience. When peace returned, some 70,000 Algerians remained in France, and by living with extreme economy they were able to support many thousands of their relations in Algeria.

The French believed that the Algerians did not want independence but to merge themselves in France. In thinking thus, the French were fixing their gaze on the tiny minority who had received a French education and who saw the salvation of the mass of their compatriots in the extension to them of a similar assimilation. But the French ignored two other groups. Algerians mainly in France, under the leadership of Messali Hajj, had formed an Algerian and nationalist movement that, from 1936, took the title of the Parti Populaire Algérien. In Algeria itself there came into being another movement, led by a man of Muslim religious learning, Shaykh Abdulhamid ben Badis; this was the Association of Algerian Ulama, which founded schools that gave education in Arabic.

The government in Paris was well disposed toward the assimilationist movement, but when in 1937 it took tentative steps in that direction the opposition of the settlers brought its efforts to nothing.

World War II and the independence movement. World War II brought with it the collapse of France and, in 1942, the Anglo-American occupation of North Africa. The occupation forces were to some extent automatically agents of emancipation, while broadcasts in Arabic both from Allied and Axis stations began to compete with promises of a brave new world for formerly subject peoples. The effect was heightened by the promise of the emancipation of Syria and Lebanon, given in June 1941 by the Free French and backed by the British authorities in the Middle East.

The Algerian Manifesto

In December 1942 the former assimilationist leader Ferhat Abbas drafted an Algerian Manifesto, for presentation to the Allied as well as to French authorities, seeking recognition of the political autonomy of Algeria as a sovereign nation. In December 1943 Gen. Charles de Gaulle declared that, because of loyalty shown, France was under an obligation to the Muslims of North Africa, and in March 1944 French citizenship was extended to certain categories of Muslims. This was by then, however, far from enough to satisfy Muslim opinion. A display of Algerian nationalist flags at Sétif in May 1945 led to an unorganized rising, in which 84 European settlers were massacred. The suppression that followed was indiscriminate, and it resulted, according to a French committee of inquiry, in not less than 1,800 deaths.

On September 20, 1947, a statute of Algeria was finally voted by the French Assembly, defining the country as "a group of departments endowed with a civic personality, financial autonomy, and a special organization." The statute created an Algerian Assembly of 120 deputies, elected in equal numbers by two electoral colleges—one composed of 370,000 Europeans and 60,000 assimilated Muslims and the other of 1,300,000 Muslims. After lengthy debates the statute was passed by a small majority, with 15 Muslim members abstaining. Muslims were at last to be considered as full French citizens with the right to keep their personal Qur'ānic status and were granted the right to work in France without further formalities. The military territories of the south were to be abolished, and Arabic was to be taught in schools at all levels.

Unfortunately, the implementation of this law was poor and the subsequent elections were "managed" by the administration, while most of the reforms laid down by the statute remained a dead letter. In spite of this, Algeria remained quiet. In reality, the principal change had been the fact that some 350,000 Algerian workers were able to establish themselves in France and, by living with extreme frugality, to remit money annually to their families in Algeria.

The Algerian War. Signs of approaching storm, however, were only too apparent. In 1950 the French police discovered that a robbery from the Oran post office had been the work of the Organisation Secrète, an offshoot of the party led by Messali Hajj, which had now taken the name of the Movement for the Triumph of Democratic Liberties (MTLD). The leader in the robbery was Aḥmad ben Bella, who had been highly commended during the fighting in Italy with the Free French. In 1952 Ferhat Abbas, when tried for a trivial offense, was defended by three lawyers—one Muslim, one Christian, and one Jewish; these combined to deliver an impressive attack on the administration. About the same time, Aḥmad Mezerna, acting head of the MTLD, took the unprecedented step of personally seeking support in Egypt. The head of the Association of Algerian Ulama toured the Arab East and secured scholarships from the Arab governments for Algerians who wished to pursue their studies in Arabic.

The Front of National Liberation

The storm burst on the night of October 31, 1954. It was organized by a few young men who, dissatisfied with Messali Hajj's leadership, had decided that justice for Algeria could only be realized by the stimulus of open rebellion. The movement took the title of the Front of National Liberation (FLN) and issued a leaflet stating that the aim was the restoration of a sovereign Algerian state. It advocated social democracy within an Islāmic framework and citizenship for any resident in Algeria, with the same rights and duties as other citizens. A preamble recognized that Algeria had fallen behind the other Arab states in emancipating itself socially and nationally, but it claimed that this could be remedied by a difficult and prolonged struggle. Two weapons would be

used—guerrilla warfare at home and diplomatic activity abroad, particularly at the UN, where the support of the Arab countries and other states would be invaluable. The FLN military objective was to make the position of the administration untenable by sudden raids, ambushes, and sabotage.

Though the first outbreak, which occurred in the region of Batna and the Aurès, was ineffective militarily, it led to the arrest of some 2,000 members of the MTLD, who had not in fact been in favour of open rebellion. In mid-February 1955 Jacques Soustelle arrived in Algiers as governor general; in June he announced a new plan, which, however, was to prove too little and too late. On August 20 a new rising at Aïn Abid, about 27 miles from Constantine, and at the mine of al-Alia near Philippeville (now Skikda) degenerated into another massacre of Europeans, followed by summary executions of Muslims. In January 1956 the electoral victory of the Republican Front in France and the premiership of Guy Mollet led to the appointment of the moderate and experienced Gen. Georges Catroux as governor general. When Mollet personally visited Algiers, however, to prepare the way for the new governor general, he was bombarded by the European populace with tomatoes. Yielding to this pressure, he allowed Catroux to withdraw and named in his place the pugnacious Socialist Robert Lacoste as resident minister. Lacoste's policy was described as pacification, but in fact it relied on forcible suppression.

French forces sent to Algeria

A French army of 500,000 men was sent to Algeria to counter the control that the rebels had managed to establish in the more out of the way portions of the country while collecting money for their cause and taking reprisals against fellow Muslims who would not cooperate with them. By May the rebels had won over the majority of previously noncommitted political leaders; and Ferhat Abbas and Tawfiq al-Madani, of the Association of Algerian Ulama, had joined FLN leaders in Cairo.

Externally, the event of 1956 was the French decision to grant full independence to Morocco and Tunisia and to concentrate on retaining "French Algeria." The rulers of the newly independent states—the Moroccan sultan and Premier Habib Bourguiba of Tunisia—hoping also to find an acceptable solution to the Algerian problem, prepared to hold a meeting in Tunis with five principal Algerian leaders who had been guests of the sultan in Rabat. French intelligence officers, however, managed to divert to Algiers the plane chartered by the Moroccan government to fly the Algerians to Tunis. The Algerian leaders were then arrested and confined in prison in France for the next six years. Far from decapitating the rising, this act provoked an outbreak in Meknès that cost the lives of 40 French settlers before the newly independent Moroccan government could restore order.

The next year, 1957, saw a rebel attempt to paralyze the administration of Algiers by terrorism. This was defeated by French parachute troops, who used torture to extract information. The French also cut Algeria off from independent Tunisia and Morocco by barbed wire fences, illuminated at night by searchlights; this separated the resistance bands within Algeria from some 30,000 Algerian armed forces who occupied positions between the fortified fences and the actual frontiers of Tunisia and Morocco, from which they drew supplies. These troops had the advantage of a friendly people and government as a base; they could not, however, penetrate into Algeria proper but could only harass the French line.

Provoked by these assaults, the French air force in February 1958 bombed the Tunisian frontier village of Sāqiyat Sīdī Yūsuf, killing a number of civilians, including children from the local school. This led to an Anglo-American mediation mission, which negotiated the withdrawal of French troops from various districts of Tunisia and their concentration in the naval base of Bizerte (Banzart).

From April 27 to 30 a meeting—the Maghrib Unity Congress—was held in Tangier under the auspices of the Moroccan and Tunisian nationalist parties and the Algerian FLN. This recommended the establishment of an Algerian government in exile and of a permanent secretariat to promote Maghrib unity. The latter proposition had little permanent result, but a government—the Provisional Government of the Algerian Republic (GPRA)—was set up on September 19.

By then, however, conditions had been radically changed by events of May 13, 1958; these began as a traditional settler rising—thousands of European Algerians sacked the offices of the governor general and, with the tacit approval of the army officers, called for the integration of Algeria with France and for the return of de Gaulle to power. The Muslims were clearly taken aback, but soon there was a relatively friendly mixing of Muslim demonstrators with the Europeans and a general hope of better times to come. In the crisis caused by this rising, de Gaulle returned to power in France.

Changes under de Gaulle

On June 4 de Gaulle visited Algiers amid scenes of great enthusiasm. He gave no clear indication, however, that he shared the settlers' enthusiasm for integration, which, in their minds, meant the submergence of the Algerians in an enlarged France. All Muslims, however, were now granted the full rights of French citizenship, and on October 30 de Gaulle announced in Constantine a plan to provide adequate schools and medical services for the Muslim population, to create employment for the vastly increasing Muslim masses, and to introduce Muslims into the higher ranks of the public services.

The European population was troubled by the lack of insistence on the theme of integration. This came to a head in September 1959, when de Gaulle, in anticipation of the opening of the UN General Assembly, declared publicly that the Algerians had the right to determine their own future. From this time it gradually became clear that while he would retain as close links between France and Algeria as he believed possible he was nevertheless prepared to go even to the length of granting independence if peace could not be secured on any other terms. The agitation of the Europeans now became extreme.

On January 24, 1960, a fresh settler rising collapsed after nine days from lack of military support. A year later, however, as the prospect of negotiations with the GPRA became more probable, there was another rising, this time organized by four generals, of whom two—Raoul Salan and Maurice Challe—had previously been commanders in chief in Algeria. De Gaulle remained unshaken, and the rising, lacking substantial support from the army, collapsed after only three days.

It became clear that the Armée d'Afrique, which had once created French Algeria, was no longer going to receive government support to fight for it. In fact, in May 1961, negotiations were opened in France with representatives of the GPRA. This body had by now long been recognized by the Arab and Communist states, from which it received aid, though it had never been able to establish itself on Algerian soil. Negotiations were broken off in July, after which the veteran Ferhat Abbas was replaced as premier by the much younger Yusuf ben Khedda. Settler opposition was meanwhile organized by a body calling itself the Organisation Armée Secrète (OAS); this began to employ terrorism as brutal as that of the rebels had sometimes been.

On March 8, 1962, negotiations were resumed, and on the 18th agreement was finally reached. Algeria would thenceforth be independent, provided only that a referendum, to be held in Algeria by a provisional government, confirmed the desire for it. In case of approval, French aid would continue; Europeans could depart, remain as foreigners, or take Algerian citizenship, as they preferred. This announcement produced a violent outburst of terrorism and attempted resistance by the OAS; but in May the terror subsided as its futility became obvious. On July 1, 1962, the referendum recorded some 6,000,000 votes in favour and only 16,000 against. After three days of unbounded Muslim rejoicing, the GPRA entered Algiers in triumph, while the departure of the Europeans, which had long since begun, assumed mass proportions.

Independence for Algeria

Meanwhile, however, the GPRA itself was torn by dissensions, and its authority was challenged by Col. Houari Boumedienne, who commanded the Algerian army on

the Tunisian side of the frontier fences. In this he was supported, after some hesitation, by Ben Bella, now released from captivity in France, and by Muḥammad Khider, secretary general of the FLN. It was not until three months later that the small-scale civil war that ensued was finally settled by the recognition of Ben Bella as premier, Boumedienne as chief of staff, and Khider as head of the party organization. By then not more than a tenth of the former 900,000 Europeans remained in the country; and the abandoned houses and apartments in the towns were rapidly taken over by Algerians from the outer suburbs.

Independent Algeria (since 1962). The Europeans who had abandoned the country included the overwhelming majority of senior administrators and managerial and technical experts. The chief exception was a group of some 10,000 French schoolteachers who remained, with great courage, often in very isolated posts.

During the six previous years some 10,000 French officers and men and possibly as many as 250,000 Muslims had lost their lives in the fighting; dozens of villages had been destroyed and 2,000,000 peasants had been moved to new sites. Nevertheless, many public services such as the post office, the railways, and the electricity supply continued to work remarkably well. On farms and in factories, however, management had largely vanished, though the workers were still there. Production fell while unemployment and underemployment reached extreme levels. Workers were able to carry out routine tasks, but matters of planning, purchasing, and marketing presented enormous problems.

Ben Bella's popularity, his style of simple living, and his courage were great assets, but his personal style of government and his reckless promises of support for revolutionary movements with which Algeria was not directly concerned were not conducive to orderly administration. The problem was complicated by the large number of leaders who had distinguished themselves in the revolutionary struggle but who now found themselves offered only posts they felt were far below their merits.

On the other hand, there were two factors of great assistance: one was the revenue from the very extensive oil fields and from the natural gas that had recently been discovered and exploited in the Sahara; the other was the improved financial position caused by the cessation of the imports required by the settlers and of the expenditure of foreign currency formerly necessitated by their journeys to France. Moreover, the Algerian workers were still able to proceed to France and to remit large sums to Algeria.

A serious problem was presented in April 1963, however, by the resignation of Muḥammad Khider and by his subsequent departure abroad, taking with him the funds of the FLN. He was subsequently assassinated in Madrid. Little by little, the gradual elimination of other dissident leaders appeared to leave Ben Bella in entire control.

Quarrel with Morocco

In the autumn of 1963 a quarrel with Morocco over the Tindouf area, which the French had attached to Algeria, led to an Algerian surprise attack on a Moroccan frontier post. Morocco, replying with vigour, had the best of the fighting. Peace was restored by the intervention of the emperor of Ethiopia and the Organization of African Unity; but the issue remained in suspense until a rapprochement between the two countries, which took place at the end of 1969, led in 1970 to an agreement for joint exploitation of the valuable Gara Jebailat iron ore deposits under Algerian sovereignty but using a Moroccan port for outlet.

Long before this, however, a totally unexpected coup in June 1965 led to the arrest and seclusion of Ben Bella and his replacement as chief of the government by Col. Houari Boumedienne. The change in government resulted in a more orderly administration.

In 1967 Algeria dispatched a force to the United Arab Republic to assist in fighting against Israel and maintained it there until 1970. Having received much assistance from the Communist states during the insurrection, Algeria gladly accepted their subsequent assistance in various forms after peace was established (for example, Bulgarian medical missions). An acute problem facing the government was the winemaking industry, which the French had established, in a Muslim country opposed by religious belief to the consumption of alcoholic liquor. With the departure of the Europeans, the internal market for wine vanished, and the French, for the benefit of French producers, were no longer interested in buying this Algerian product. The U.S.S.R. afforded relief when it undertook to buy the annual surplus.

Algeria's foreign-exchange position in general was greatly assisted by the export of oil, of which two-thirds went to France, and of natural gas, of which two-thirds went to Great Britain, in liquid form. Private investment from abroad virtually ceased, but the lack was largely made up by such international institutions as the World Bank and by other forms of international aid. By 1970 Algeria appeared to be a stable republic facing its innumerable problems with considerable determination, friendly with the Communist states but resolutely independent.

TUNISIA

European influence (1830–81). In 1830, at the time of the French invasion of Algiers, Tunis was a dominion of the Ottoman Empire but independent in its internal affairs. Educated Tunisians had not forgotten that neighbouring Algiers had in great part been created by the Turks out of Tunisian territory. Quite recently Algiers had claimed a kind of suzerainty over Tunis and even the right to tribute. In these circumstances the reigning *bey* of Tunis, Ḥusayn ibn Maḥmūd, lent ready credence to French assurances that they had no permanent claims on Algiers and welcomed the suggestion of the appointment of Tunisian princes as governors of Constantine and Oran and hints that Tunis might eventually take over the neighbouring regency. It soon became clear that this scheme had no prospect of success and it was abandoned. The next *bey*, Muṣṭafa (1835–37), would indeed have been glad openly to assist the Algerians against the French had he been in a position to do so. Meanwhile, the Ottoman suzerain had seized the opportunity of a disputed succession in Libya to depose its ruling dynasty and to re-establish direct Turkish rule. The possibility of a similar intervention in Tunis was forestalled by the dispatch of a French naval force in 1836, and a similar threat in the following year by renewed warnings. During the succeeding reign of Aḥmad Bey (1837–55), the emergence of Tunisia as a distinct state continued; and Aḥmad abolished all slavery, permitted the opening of Christian schools, emancipated the Jews, and employed European advisers to help create a modern army and navy. The latter effort, however, proved costly and produced little result. The expense of these reforms, combined with the personal extravagance of the ruler, led to increasing financial difficulties, particularly since Tunisia, like the other North African states, no longer profited from privateering. Higher taxes, imposed to make up the loss of income, provoked revolts in 1840, 1842, and 1843. As Muslim strength in Africa and in Turkey decayed, a struggle ensued between rival European powers, principally Britain and France, for a controlling interest. Britain was interested mainly in preserving weak Muslim governments along the sea route to Egypt, while France hoped to extend its control from Algeria eastward over Tunis and westward over Morocco, with a land link across the Sahara to its central and West African colonies.

Growth of financial difficulties

The influence these two powers exercised was made clear when the *bey* Muḥammad (1855–59) sanctioned the execution of a Jew for blasphemy. Joint action by the British and French consuls induced the *bey* to issue the "Fundamental Pact" (*'ahd al-amān*; September 9, 1857), a kind of declaration of the rights of man. The final collapse of the regency occurred during the reign of Muḥammad as-Sadiq (known in France as Saddok or Sadoq; 1859–82), who attempted to modernize his state and even proclaimed a constitution, or *dustūr*. Meanwhile, expenditure continued to mount and with it taxation, which the people attributed to the foreign influence. As rivals for influence, the French and English were joined after 1860 by the representative of united Italy. The

position was worsened by the foreign loans negotiated on exorbitant terms by the *khazandār* (treasurer) Muṣṭafa.

In 1864 Muhammad, faced with an armed rising, reduced the *majbā*, a new tax that had been introduced during the preceding reign, and also abolished the constitution. This was accompanied by a movement to invoke the intervention of the Turkish suzerain, which, however, was brought to naught by French opposition. Four years later (1868) the *bey* found himself compelled to accept the appointment of an international commission. By its decision Tunisia was constrained to devote half the state revenues to repayments, an arrangement for which the customs receipts served as guarantee. When the *khazandār* Muṣṭafa was at last dismissed (1873), his place was taken by the Circassian Khayr ad-Dīn, who had been a supporter of liberal measures for several years and in 1871 had secured from the Ottoman sultan a *firman*, or edict, reaffirming Ottoman sovereignty while renouncing any claim to tribute. This had successfully averted a threatened Italian naval demonstration. In 1878, however, he was dismissed by the *bey*, without having achieved any material improvement in the position. At the Congress of Berlin in 1878, the British objection to French ambitions in Tunisia was withdrawn in return for French consent to the British occupation of Cyprus. Without British support, on which Tunisia had long relied as a counterweight to other powers, it was left face to face with France and Italy, whose interests were in occupation, not in the maintenance of Tunisian independence.

The end came in 1881, when the French, on the pretext that some Tunisian tribesmen had moved into Algerian territory, landed troops at Bizerte (Banzart) and sent others into Tunisia by land from Algeria. Advancing without difficulty to the *bey*'s palace, the Bardo, at Kassar Said (modern al-Qaṣr as-Saʿīd), a short distance outside Tunis, they imposed a treaty that sanctioned a French military occupation, transferred to France the *bey*'s authority in foreign relations and finance, and provided for the appointment of a French resident minister as intermediary in all matters of common interest. This provoked a rising in southern Tunisia during which Sousse (Sūsah) was bombarded and captured in July 1881; al-Qayrawān was then captured in October, and Gafsa (Qafṣah) and Gabes (Qābis) in November. After the death of the *bey* as-Sadiq, his successor, Ali, was constrained to sign an undertaking to introduce such administrative, judicial, and financial reforms as the French government might consider useful. This agreement, known as the Convention of Marsa and signed in 1883, made French control complete.

French occupation

The protectorate (1881–1956). The arrangement thus arrived at was very different from that in Algeria. The basis was a treaty, not an outright conquest. The *bey*, known as Ṣāḥib al-Mamlakat at Tūnisīyah ("Possessor of the Kingdom of Tunis"), remained in theory an absolute monarch; two Tunisian ministers were still appointed, and the framework of the old government machinery was preserved. Tunisians continued to be subjects of the *bey*. There was no confiscation of land; mosques were not converted into churches; and Arabic remained an official language. Nevertheless, the supreme authority passed in fact into the hands of the French resident general and his staff. Under French guidance the finances were soon brought into order and a modern communications system established. Valuable phosphate mines were brought into operation near Gafsa in the south, while the extensive establishment of French and Italian colonists in the Majardah Valley resulted in a considerable export of vegetables. For a quarter of a century the country developed without serious disturbance, while the Muslim population adjusted its ways of thought, within a French framework, to the outlook and the science of the modern world. Though there was none of the wholesale confiscation of land and displacement of population that had occurred in Algeria, the most fertile portions of northern Tunisia, comprising the Majardah Valley and the Cap Bon peninsula, passed largely into the hands of Europeans. Many Europeans, especially the Italians and the Maltese, also competed with the indigenous Muslims in such humble occupations as driving taxis.

By 1906 the Tunisians had begun to demand a more active part in the government of their country, and in 1907 the Young Tunisian Party was founded, which financed a paper in the French language, *Le Tunisien*. Four years later, popular resentment of European domination was greatly intensified by the Italian invasion of Turkish Tripoli. The outbreak of World War I, however, rallied the Tunisians behind the French government. Nevertheless, the war accelerated the rate of change. Tunisians saw no reason why the Fourteen Points of the U.S. president Woodrow Wilson should not apply to themselves as much as to the subject peoples of Austria-Hungary and Germany. The Russian Revolution also made its impact. In 1920 the Destour Party, as the progress party now called itself, presented to the *bey* and to the French government a document that in effect demanded the establishment of a constitutional form of government in which Tunisians would possess equal rights with Europeans. The immediate result was the arrest of Abd al-Aziz al-Thaalibi, the Destour leader. Two years later, the aged *bey* Muhammad an-Nasir, under the influence of his sons, took the occasion of the coming visit of the French president, Alexandre Millerand, to request the adoption of the program of the Destour, failing which he threatened to abdicate within 48 hours. The resident general, Lucien Saint, responded by surrounding the *bey*'s palace with troops, with the result that the demand was withdrawn. Saint thereupon introduced repressive measures together with minor reforms that pacified Tunisian sentiment and weakened the nationalist movement for several years.

Young Tunisian Party

In 1934 a young Tunisian lawyer, Habib Bourguiba (*q.v.*), broke with the Destour Party to form a new organization, the Neo-Destour (now the Socialist Destour Party), which aimed at gaining mass support. Under Bourguiba's vigorous leadership the new party, after a sharp and often bitter struggle, completely supplanted the existing Destour and its old-fashioned leaders. Attempts by the French to suppress the new movement spread rather than reduced its popular appeal. The arrival of the Popular Front to power in France in 1936 enabled the Neo-Destour to extend its propaganda there also. When the Popular Front collapsed, renewed repression in Tunisia was followed by civil disobedience; and in 1938 serious disturbances resulted in the arrest of Bourguiba and other leading members of the party, which was officially dissolved. On the outbreak of war in 1939, they were deported, still untried, to France for greater security, until released by the Nazis when they later occupied Vichy France in 1942. Since Hitler regarded Tunisia as a sphere of Italian influence, they were then handed over to the Fascist government in Rome. Here they were treated with deference, in the hopes of eliciting a declaration favourable to the Axis. Bourguiba, however, steadily refused, in spite of the suggestions of some of his companions. In March 1943, after a noncommittal broadcast by him, they were finally allowed to proceed to Tunis, where the reigning *bey*, Muhammad al-Munsif (Moncef), formed a ministry of Destour sympathies. The subsequent assumption of power by the Free French, after the Nazi retreat, resulted in complete disillusionment. The *bey* himself was deposed, while Bourguiba escaped imprisonment only by a flight in disguise to Egypt (1945). Though ill at ease in the Middle Eastern environment, he now began a vigorous campaign of propaganda for Tunisian independence, by stages and, as far as possible, with the aid of France. In view of the emancipation of the eastern Arab states, and later of the neighbouring, relatively backward Libya, the French felt compelled to make concessions. In 1951 a government with nationalist sympathies was allowed to take office, of which the secretary general of the Neo-Destour, Salah ben Youssef, became a member. Bourguiba himself was allowed to return to Tunisia. When, however, this government wished to go farther and establish a Tunisian parliament, the result was a further bout of repression. Bourguiba was confined to the little island of Galite (now called Jāliṭah) in the Bay of Tunis, while the ministers were put under arrest, apart from Salah ben Youssef and one other who managed to escape to Cairo. This resulted, for the first time, in outbreaks of terrorism;

Emergence of Bourguiba

Outbreak of terrorism

and nationalist bands began to operate in the mountains, virtually paralyzing the country.

In July 1954 the French premier, Pierre Mendès-France, flew to Tunis, accompanied by Marshal Alphonse Juin, and promised to grant complete autonomy to Tunisia, subject to a freely negotiated convention. The negotiations were conducted in France, where Bourguiba, who was now released, was able to supervise them without directly participating. In June 1955 the convention was finally signed by the Tunisian delegates, though it imposed strict limits in the fields of foreign policy, education, defense, and finance. A mainly Neo-Destour ministry was formed, and Bourguiba himself, on returning from France, received a delirious welcome from the Tunisian people. Salah ben Youssef now came back from Cairo, denounced the agreement as too restrictive, refused to attend a specially summoned congress unanimously supporting Bourguiba, and organized armed resistance in the south, where his influence was greatest. This rising, however, was quickly put down, and Ben Youssef fled the country. (Ben Youssef was assassinated mysteriously in West Germany in 1961.)

Independence (1956–70). In March 1956, when full independence was granted to Morocco, it became inevitable that France should behave with equal liberality to the more developed Tunisia. A new Franco-Tunisian declaration was issued on March 20, and Bourguiba himself now headed the Ministry. The transfer of power was carried out successfully, Tunisians replacing Frenchmen as ministers without disturbance. In November Tunisia was admitted to the United Nations. On July 25, 1957, the rule of the *beys* was abolished and a republic declared, with Bourguiba as both president and prime minister. A major difficulty of the new government now arose from the continued fighting in Algeria and the use of the Tunisian frontier as a base by the Algerian insurgents. This problem became acute when, in February 1958, the French air force from Algeria bombarded the Tunisian frontier village of Sāqiyat Sīdī Yūsuf, inflicting many casualties, including school children. International opinion strongly supported Tunisia, and, by the intervention of a joint Anglo-American mission, French troops were withdrawn from the interior and concentrated at the naval fortress at Bizerte. Three years later, General de Gaulle was still unwilling to withdraw these remaining troops. This led to Tunisian demonstrations in the summer of 1961 during which French parachutists from Algeria inflicted heavy losses on the demonstrators. The French finally evacuated Bizerte in 1963.

The Bizerte incident

During the protectorate, Tunisia had been regarded as the North African country with the most peaceable and civilized inhabitants and has since preserved this reputation with the Western powers in general, receiving from them great financial and moral support. Its relations with France, however, at times have been extremely tense. The Bizerte affair alone resulted in 1,000 Tunisian and 20 French casualties; and in the following year all French aid was suspended when the National Assembly abruptly decided on the immediate nationalization of all land held by foreigners. The position was not modified nor were the various trade preferences restored until 1966.

Tunisia's relations with the progressive Arab states, notably the United Arab Republic and later Algeria, were also subject to violent oscillations, and both countries were on occasion accused of plots against the president's life. There was a violent dispute with the United Arab Republic and the majority of Arab states when Bourguiba in 1965 urged a softer approach toward Israel. During the 1967 Arab–Israeli war, however, Tunisian reactions were similar to those of the rest of the Arab world; and on the occasion of the Jordanian civil strife of 1970, the Tunisian prime minister was asked to mediate.

MOROCCO

Decline of traditional government (1830–1912). During the French invasion of Algeria in 1830, the sultan of Morocco, Moulay Abd ar-Rahman (1822–59) briefly sent troops to occupy Tlemcen but withdrew after French protests. The Algerian leader Abdelkader offered his allegiance to the Sultan and in 1844 took refuge from the French in Morocco. A Moroccan army was sent to the Algerian frontier; the French bombarded Tangier on August 4, 1844, and Mogador on the 15th. Meanwhile, on August 14 the Moroccan army was totally defeated at Isly near the frontier town of Oujda. The Sultan then promised to intern or expel Abdelkader if he should again enter Moroccan territory. Two years later, when again driven into Morocco, the Algerian leader was attacked by Moroccan troops and forced to surrender to the French. The dominating foreign power in Morocco at this time, however, was Britain. In 1856 the British secured a treaty granting them trade privileges, including the right of according "protection" to Moroccan citizens, making them largely independent of their national government. Immediately after Abd ar-Rahman's death in 1859, a dispute with Spain over the boundaries of the Spanish enclave at Ceuta led to a declaration of war by Madrid and to the Spanish capture of Tetuan in the following year. Peace had to be bought with an indemnity of $20,000,000, the enlargement of Ceuta's frontiers, and the promise to cede to Spain another enclave—Ifni. The new sultan, Sidi Muhammad (1859–73), attempted without any marked success to modernize the Moroccan army. After his death, his son Moulay al-Hassan I (1873–94) struggled to preserve the independence of his medieval empire. His time was consumed in unceasing journeys to restore order and collect taxes, for the machinery and means of a modern government were completely lacking. Like his predecessor, he was unsuccessful in his efforts to reform the army. After his death, his chamberlain, Ba Ahmad, ruled in the name of the young sultan Moulay Abd al-Aziz until 1901, when the latter began his direct rule. Abd al-Aziz surrounded himself with European companions and adopted their customs, scandalizing his subjects, particularly the religious leaders. His attempted introduction of a modern system of land taxation resulted in complete confusion because of a lack of qualified officials. Popular discontent and tribal rebellion became even more frequent, while a pretender, Bu Hamara, established a rival court near Melilla. European powers seized the occasion thus offered to extend their own influence. In 1904 coming events were foreshadowed when Britain gave France a free hand in Morocco in return for a French undertaking not to interfere with British plans in Egypt. Spanish agreement was secured by a French promise that northern Morocco should be treated as a sphere of Spanish influence. Italian interests were satisfied by a French undertaking not to hinder Italian designs on Libya. Possible German claims were ignored, thus enabling the Sultan to arrange an international conference at Algeciras in 1906 to discuss the whole Moroccan question. The Algeciras conference confirmed the integrity of the Sultan's domains; the right of all countries to trade in Morocco on equal terms was ensured by the imposition of customs dues at a uniform rate of 12½ percent on all imports. At the same time, however, the conference sanctioned French and Spanish policing of the ports and collection of the customs dues. Two years later, the Sultan's brother, Moulay Abd al-Hafid, led a rebellion against him, denouncing him for his departure from Muslim ways. Defeated, Moulay Abd al-Aziz took refuge in Tangier. Moulay Abd al-Hafid then made an abortive attack on French troops, which had been occupying Casablanca since 1907, before proceeding to Fez (Fès), where he was duly proclaimed sultan and recognized by the European powers (1909). The new sultan proved unable to control the country. Disorder increased until, besieged by tribesmen in Fez, he was forced to ask the French to rescue him. When they had done so, he had no choice but to sign the Treaty of Fès (March 30, 1912) by which Morocco became a French protectorate and the French government was authorized to introduce such reforms as it thought fit. In return the French guaranteed to maintain the status of the Sultan and his successors. Provision was also made to meet the Spanish claim for a special position in the north of the country; Tangier, long the seat of the diplomatic missions, retained a distinctive administration.

Attempts at modernization

French intervention

The French protectorate (1912–56). In establishing their protectorate over Morocco, the French had behind them the experience of the conquest of Algeria and of their protectorate over Tunisia; and they took the latter as model for their Moroccan policy. There were, however, important differences, owing both to the later date of its establishment and to the special circumstances of Morocco. The date was only two years before the outbreak of World War I, which was to bring with it the new attitude to colonial rule. Secondly, Morocco had a tradition of a thousand years of independence and had been strongly influenced by the civilization of Muslim Spain, while it had never been subject to Ottoman rule. These circumstances and the proximity of Morocco to Spain created a special relationship between the two countries. Morocco was also unique among North African countries in possessing a coast on the Atlantic, in the rights that various nations derived from the Act of Algeciras, and in the privileges that their diplomatic missions had acquired in Tangier. Thus, the northern tenth of the country, with an Atlantic and a Mediterranean coast, together with the desert province of Tarfaya in the south adjoining the Spanish Sahara, was excluded from the French-controlled area and treated as a Spanish protectorate. In the French sector, the French resident general became the real ruler of the country, subject to the approval of the Paris government. He worked through newly created departments manned by French officials. The outward forms of the traditional Moroccan government (*makhzan*) were preserved, though the role that it played in reality can be estimated from the fact that the grand vizier on the installation of the protectorate, Muhammad al-Moqri, still held the same post on the recovery of independence 44 years later, when he was over 100 years old. As in Tunisia, country districts were administered by *contrôleurs civils*, except in certain areas such as Fès, where it was felt necessary that officers of the rank of general should supervise the administration. In the south, certain Berber chiefs, of whom the best known was Tihami al-Glawi, were allowed to remain semi-independent.

The protectorate's government

The pre-World War II period. The first resident general, Gen. (later Marshal) L.H.G. Lyautey, was a soldier of wide experience in Indochina, Madagascar, and Algeria. He was of aristocratic outlook and possessed a deep aesthetic appreciation of the artistic qualities of Moroccan civilization. The character he gave to the administration exerted an influence throughout the period of the protectorate. As early as 1920 he submitted a report saying that "a young generation is growing up which is full of life and needs activity. . . . Lacking the outlets which our administration offers only sparingly and in subordinate positions they will find an alternative way out." Only six years after Lyautey's report, young Moroccans both in Rabat, the new administrative capital, and in Fès, the centre of traditional Arab learning and culture, were meeting, quite independently of one another, to discuss demands for reforms within the terms of the protectorate treaty. They asked for more schools, a new judicial system, the abolition of the regime of the Berber *qā'īd*s (*caid*s) in the south, for study missions in France and in the Arab East, for the cessation of official colonization, and for the suppression of licensed prostitution—objectives that would only be fully secured when the protectorate ended in 1956. Moulay Abd al-Hafid could not reconcile himself to the new regime and, after a few months, joined his brother as a French pensioner in Tangier. In his place a more amenable brother, Moulay Yusuf, was recognized as sultan and succeeded in cooperating with the French without losing the respect of his own people. A new administrative capital was created on the Atlantic coast at Rabat. At the same time, a commercial port was developed at Casablanca. By the end of the protectorate, Casablanca was a flourishing city, with nearly a million inhabitants and a considerable industrial establishment. Lyautey's policy was to build new European cities beside or at some distance from the old Moroccan towns, thus preserving the country's ancient monuments. The remarkable rhythm of innovation was little interrupted by World War I. Though the French government had proposed to retire to the coastal area, Lyautey managed to retain control of all of the occupied area. After the war one major problem was the pacification of the former Bled es-Siba outlying areas in the Atlas Mountains over which the sultan's government often had had no real control. This was finally completed in 1934. Another problem was the extension of the rising of Abd el-Krim from the Spanish to the French zone (see below *The Spanish Zone*). In 1926 Marshal Lyautey was succeeded by a civilian resident general. This marked a change to a more conventional colonial regime, accompanied by the extension of official colonization, the growth of European population, and the increasing impact of European thought on the minds of the younger generation, some of whom had by now received a French education. On the death of Moulay Yusuf (1927) the French choice as his successor was his younger son, Sidi Muhammad (Muhammad V). Chosen in part for his retiring disposition, this sultan was in time to reveal outstanding diplomatic skill and determination. Another significant event was the French attempt to utilize the differences between Arabs and Berbers to counterbalance the Arab character of the Sultan's government. This led to the issue of the Berber Decree of 1930. The Berbers had hitherto been brought to accept the Arab way of life by a gradual process rather than by any sort of compulsion. Now, however, the Berber areas were to be given a perpetual exemption from the Muslim law of the kingdom. This at once stimulated Arab nationalism and brought it widespread backing from the Arab and Muslim world in general. The resulting outcry forced the administration to give ground and to modify its proposals. In 1933 the nationalists initiated a new national day called the Fête du Trône to mark the anniversary of the Sultan's accession. When he visited Fès in the following year, he received a tumultuous welcome, accompanied by anti-French demonstrations that caused the authorities to terminate the visit abruptly. This episode was soon followed by the organization of political parties of nationalist sentiment. These events coincided with the completion of the occupation of southern Morocco, which brought with it the Spanish occupation of Ifni. In 1937 rioting occurred in Meknès, where French settlers were suspected of diverting part of the town water supply to irrigate their own lands at the expense of the Muslim cultivators. Thirteen of the rioters were killed and 100 were wounded. In the ensuing repression, Allal el-Fasi, the main nationalist leader, was banished to Gabon in Equatorial Africa, where he spent the following nine years. The tendency of French policy at this time is shown by a circular, issued to departmental chiefs, stating that insufficient attention was being given to "native policy," implying that the affairs of the French settlers, now numbering some 300,000, were the chief care of the administration, while the mass of the population was a matter of "native" policy.

The development of Casablanca

The Berber Decree of 1930

World War II and the attainment of independence. At the outbreak of World War II in 1939, the Sultan issued a call for wholehearted cooperation with the French; and a large Moroccan contingent, mainly Berbers, served with distinction in France. The collapse of the protecting power in 1940 and the installation of the Vichy regime naturally produced an entirely new situation. The Sultan marked his independence by refusing to approve anti-Jewish legislation. When in 1942 the Anglo-American landings took place, he refused to comply with the suggestion of Resident General Auguste Noguès that he retire to the interior. In 1943 the Sultan was much influenced by his meeting with the U.S. president Franklin D. Roosevelt, who came to Morocco for the Casablanca Conference and was unsympathetic to continued French presence there. The majority of the people were equally affected by contact with the American and British troops, who put them in touch with the outside world to an unprecedented degree. Among the people at large, the effect of the newly introduced broadcasts in Arabic with which the combatants, Allied and Nazi alike, sought to attract Arab listeners to their side, was very considerable. In these circumstances, the nationalist movement took

The Allied invasion

the new title of Hizb al-Istiqlal (Istiqlal Party)—the "Party of Independence." In January 1944 they submitted to the Sultan and the Allied (including the French) authorities a memorandum asking for independence under a constitutional regime. The nationalist leaders, including Ahmad Balafrej, secretary general of the Istiqlal Party, were immediately arrested on an improbable charge of collaboration with the Nazis. This caused rioting in Fès and elsewhere in which some 30 or more demonstrators were killed. In the situation thus created, the initiative passed to the Sultan, who, in 1947, persuaded a new and reforming resident general, Eirik Labonne, to gain the French government's permission for him to make an official state visit to Tangier, passing through the Spanish Zone on the way. The journey became a triumphal procession; when the Sultan finally made his speech in Tangier, it was under the emotion of the stirring reception in the Spanish Zone and Tangier. While emphasizing the links of Morocco with the Arab world of the East, he omitted the flattering reference to the French protectorate that had been anticipated. The result was the replacement of Labonne by Gen. (later Marshal) Alphonse Juin, of Algerian settler origin. With long experience in North African affairs, Juin expressed sympathy for the patriotic nationalist sentiments of young Moroccans and promised to comply with their wish for the creation of elected municipalities in the large cities. At the same time he aroused opposition by proposing to introduce French citizens as members of these bodies, thus taking a step toward a state of co-sovereignty. The Sultan thereupon used his one remaining prerogative and refused to countersign the resident general's decrees, lacking which they had no legal validity. A state visit to France in October 1950 and a very flattering reception did nothing to modify the Sultan's views, and, on returning to Morocco, he received a wildly enthusiastic welcome from the Moroccan crowds. On December 12 General Juin dismissed a nationalist member from a meeting of the Council of Government, which discussed the budget proposals; the ten remaining nationalist members walked out in protest. This event was hailed by the local French press as a victory for their cause, though a very different view of the matter was widely held in France. For his part, General Juin now began to think of the possibility of utilizing the Berber feudal notables, such as Tihami al-Glawi, to counter the nationalists. Thus, at a palace reception on December 21, 1950, al-Glawi told the Sultan to his face that he was not the Sultan of the Moroccans but of the Istiqlal and that he was leading the country to catastrophe. The Sultan forbade him to visit the palace again, while the local French press hailed al-Glawi's action as displaying the true Muslim spirit and suggested that the nationalists were Communists.

The regime of Gen. Juin

In the face of Sidi Muhammad's continued refusal to cooperate, General Juin surrounded the palace with tribesmen, after having provided a guard of French troops supposedly to protect the Sultan against his outraged people. Faced with this threat, Sidi Muhammad was constrained to disown "a certain political party," without specifically naming it, but still withheld his signature from many decrees, including that admitting French citizens as municipal councillors. General Juin's action was widely criticized in France, and on August 28, 1951, he was replaced by Gen. Augustin Guillaume. The Sultan, on the anniversary of his accession (November 18), declared that he was hoping for an agreement "guaranteeing full sovereignty to Morocco" but (as he added in a subsequent letter addressed to the president of the French Republic) "with the continuation of Franco-Moroccan cooperation." This troubled situation continued until December 1952 when the trade unions in Casablanca decided to organize a protest meeting over the assassination of the Tunisian trade union leader, Ferhat Hashad (Hached), supposedly by French terrorists. A clash with the police followed; hundreds of nationalists were arrested and were held in detention for two years without being tried.

In April 1953, the head of a religious confraternity, Abd al-Hai al-Kittani, together with a number of Berber notables, headed by al-Glawi, with the connivance of a number of French officials and settlers, began to work for the deposition of the Sultan, whom they accused of un-Islāmic conduct. The government in Paris, preoccupied with internal affairs, finally demanded that the Sultan should transfer his legislative powers to a council, composed jointly of Moroccan ministers and French directors, and append his signature to all the blocked legislation. The Sultan yielded but this was not sufficient for his enemies. On August 18, 1953, al-Glawi delivered what may be called an ultimatum to the French government. The latter thereupon deported the Sultan and his family to Madagascar, appointing in his place the more subservient Moulay ben Arafa. This did nothing to improve the situation. Sidi Muhammad became at once a national hero, and an attempt was made on the life of the substitute sultan. The authorities in the Spanish Zone, who had not been consulted about the measure, did not conceal their disapproval. The Spanish Zone thus became a refuge for Moroccan nationalists. In November 1954 the French position was further complicated by the outbreak of the Algerian rising. In June 1955 the Paris government decided on a complete change of policy and appointed Gilbert Grandval as resident general. His efforts at conciliation were obstructed by the tacit opposition of many officials and the outspoken hostility of the majority of the French settlers. Before he could get his proposals accepted by the French government, a massacre of French settlers at Oued Zem (August 20, 1955) led to his recall. A conference of representative Moroccans was then summoned to meet in France and at this it was agreed to replace the substitute sultan by a crown council. Sidi Muhammad gave his approval to the proposal, but it still took weeks to persuade the puppet sultan to withdraw to Tangier. Meanwhile, a Liberation Army began to operate against French posts near the Spanish Zone. On October 26 al-Glawi, wishing to secure the future of his family, declared publicly that only the restoration of Sidi Muhammad could restore harmony. The French government agreed to allow the Sultan to form a constitutional government for Morocco. In November Sidi Muhammad returned to Rabat, where he was greeted by huge welcoming crowds. On March 2, 1956, independence was proclaimed, and the Sultan formed a government representative of the various elements of the indigenous population, while the departments formerly headed by Frenchmen became ministries headed by Moroccans.

The Sultan's exile

Independent Morocco (1956–70). The protectorate had been highly successful in developing communications, in adding modern quarters to the cities, and in creating a flourishing agriculture and a modern industry of a colonial type. Most of these activities, however, were managed by Europeans who, by the end of the protectorate, numbered 360,000 in French Morocco; in the constitutional field there had been virtually no development. Though the government was in practice in the hands of the French, the Sultan remained in theory an autocrat; and when independence was finally conceded he became so again in fact, subject only to his undertaking to the nationalists to introduce a constitution. By French insistence, the first cabinet was composed of ministers representing the various tendencies in Moroccan society, and it included a Jew. Mubarak Bekkai, an army officer who had shown his loyalty to the Sultan during the independence struggle but was not a party man, was selected as premier. The Sultan (who officially adopted the style of king in August 1957) selected the ministers personally and himself retained control of the army and the police; he did, however, nominate a Consultative Assembly of 60 members. His eldest son, Moulay Hassan, became chief of staff and by degrees successfully integrated the irregular Liberation Forces, after they had promoted a rising against the Spanish in Ifni (1957) and against the French in Mauritania. In general, the change to Moroccan control, assisted by French advisers, took place smoothly. Relations with France, however, were badly strained because of the continuing Algerian problem, but independent Morocco remained greatly dependent on French technology and finance. There was a shortage of trained professionals; for example, out of a total of 900 physi-

The national government

cians in the country not more than 70 were Moroccans. In the sphere of education, Morocco had to rely on the continued service of thousands of French teachers. Another problem was the lack of confidence in the future felt by French industrialists and the consequent difficulty of maintaining capital investment. This lack was mitigated by international aid. European farms in Morocco remained under French control until 1963. Progress was made in some areas. The maintenance and extension of the fine roads in the French Zone and the bringing of the roads in the Spanish Zone to the same standard was successfully carried out. The mining of phosphates, of which Morocco is one of the world's leading producers, was extended to some 10,000,000 tons annually; and a vast plant was created at Safi for conversion into superphosphates. Assembly plants were established for cars and trucks. A major political change occurred when the Istiqlal Party split into two sections in 1958. The main portion remained under the leadership of Allal el-Fasi, while a smaller section, headed by Mehdi ben Barka, Abdullah Ibrahim, and Abd er-Rahim Bouabid, formed the National Union of Popular Forces (UNFP). Of these groupings the original Istiqlal represented the more traditional elements, while the UNFP was formed from the intelligentsia of the new industrial cities and favoured socialism with republican leanings. Muhammad V made use of these dissensions to assume the position of an arbiter above party strife. He nevertheless continued the preparations for the creation of a parliament until his unexpected death from minor surgery in 1961. Moulay Hassan succeeded him and carried on his policies.

Division in the Istiqlal

In 1963, when parliamentary elections were finally held, the two halves of the former Istiqlal Party formed an opposition, while a party supporting the King's government was created out of miscellaneous elements, known as the FDIC (Front for the Defense of Constitutional Institutions). This included a new, predominantly Berber, rural group opposed to the Istiqlal. The ensuing near deadlock caused the King to dissolve Parliament after only one year, resuming personal government with himself or his nominee as premier. In 1970 a new constitution was promulgated that provided for a one-house legislature. In foreign affairs, Morocco sided with the western European powers rather than with the eastern block, while sharing the general outlook of Arab states on such questions as Palestine. Its relative moderation did not prevent a large-scale Jewish emigration. Claiming Mauritania as a sphere of influence, the Moroccan government refused to recognize that country's independence before 1970. Nor would it recognize the western frontier where territory had been attached to French Algeria at Morocco's expense. This was of importance because of the vast iron ore deposits 80 miles southeast of Tindouf; the dispute resulted in an armed clash in autumn 1964. Peace was restored with the aid of the Organization of African Unity, and in 1970 it was agreed that the iron ore would be exploited by both countries under the Algerian flag but exported through a Moroccan port. At the same time, a rapprochement with Mauritania signified an alignment of the three Maghrib states against a Spanish plan to develop the phosphate-rich Spanish Sahara into an independent state associated with Spain. In 1964 the exiled Moroccan socialist leader Mehdi ben Barka attacked the Moroccan government over Algiers radio at the time of the Moroccan–Algerian frontier fighting; for this he was sentenced to death *in absentia*. A year later, while on a visit to Paris, he was kidnapped with the aid of two French police inspectors and was never seen again. After investigations, the French president, Charles de Gaulle, accused the Moroccan minister of the interior, General Oufkir, of responsibility. As General Oufkir was unwilling to go to Paris to answer the charge, the French ambassador was withdrawn from Rabat and French economic aid cut off. Relations were not fully restored until de Gaulle's abandonment of the presidency in 1969.

Border conflict with Algeria

The Spanish Zone (1912–56). The Spanish intervention in Morocco had roots in remote history. In a sense, Spain and Morocco form an intermediate unit between Europe and black Africa to the south and between Europe and the Arab world in the east. These historical and geographical facts forced the French to share their protectorate with the Spaniards and provided a very different background to the two regimes. The outward form given to the protectorate was similar since the Spaniards appointed a *khalīfah* or viceroy, chosen from the royal family, as nominal head of state, and provided him with a puppet Moroccan government created for the purpose. This enabled them to conduct affairs independently of the French Zone while nominally preserving the unity of the whole country. In 1904 the French had been prepared to leave Spain the whole northern half of the kingdom, including Fès and the Atlantic coast as far as Salé. The Spanish government of that time, however, did not feel sure of its ability to deal with such a large area and asked only for about one-tenth of the kingdom, from Larache on the Atlantic to 30 miles beyond Melilla (already a Spanish possession) on the Mediterranean. This was in the main a mountainous, Berber-speaking area that had often escaped the sultan's control. In addition, the Spanish received a strip of desert land in the south, adjoining the Spanish Sahara and known as Tarfaya or Tekna. Finally, in 1934, when the French occupied southern Morocco, the Spanish took possession of the strip of territory on the Atlantic coast known as Ifni. Tangier, though it had a Spanish-speaking population of 40,000, received a special international administration, under a *mendub* or resident, theoretically appointed by the sultan, actually by the French. In 1940, after the defeat of France, Spanish troops occupied Tangier but withdrew again in 1945 after the Allied victory. The Spanish Zone surrounded the ports of Ceuta and Melilla, which Spain had possessed for centuries, and included the iron mines of the Rif from which some 3,000,000 tons of ore were exported by Spanish companies annually through the latter port. As capital, the Spanish selected Tétouan, which they had occupied during the war of 1860–61. As in the French Zone, European-staffed departments were created; the country districts were administered by *interventores*, corresponding to the French *contrôleurs civils*. The first area to be occupied was that on the plain, facing the Atlantic, with the towns of Larache, Alcazarquivir (now Ksar el-Kebir), and Asilah. That area was the stronghold of the former Moroccan governor Raisuni, half patriot and half brigand. The Spanish government found it difficult to tolerate his independence, and in March 1913 he retired into a refuge in the mountains, where he held out until captured by another Moroccan leader, Abd el-Krim, 12 years later. Muhammad Abd el-Krim was a Berber, a good Arabic scholar, and had a knowledge of both the Arabic and the Spanish languages and ways of life. Imprisoned after World War I, probably because of his outspoken hostility toward the French, he was later transferred to Ajdir in the Rif, the mountainous district overlooking the Mediterranean near al-Hoceima. There he made full use of his knowledge of Spanish affairs to plan a rising. In July 1921 he destroyed a Spanish force sent against him and then established a Republic of the Rif, which endured for five years. It was only when his followers resisted French expansion in the north that French and Spanish forces combined, the latter landing at al-Hoceima and the French advancing from Fès, which had been in peril; Abd el-Krim was then finally overwhelmed by a total force of over 250,000 fully equipped men. In May 1926 he surrendered to the French and was exiled to the island of Réunion in the Indian Ocean. The remainder of the period of the Spanish protectorate was relatively calm. Thus, in 1936 Gen. Francisco Franco was able to launch his attack on the Spanish Republic from Morocco and to enroll a large number of Moroccan volunteers, who served him loyally in the Spanish Civil War. Though the Spanish had fewer resources than the French, their subsequent regime was in some respects more liberal and less subject to racial discrimination; instruction in schools was in Arabic, not Spanish, and Moroccan students were encouraged to go to Egypt for a Muslim education. There was no attempt to set Berber against Arab as in the French Zone, though this may have been prevented by the introduction of Muslim law by

The Rif Republic

Abd el-Krim himself. After the end of the Rif Republic, there was little cooperation between the two protecting powers. Their disagreement reached a new height when in 1953 the French, without consulting the Spanish, deposed the Sultan. The Spanish high commissioner did not recognize this action, and Muhammad V was still regarded in the Spanish Zone as the legitimate sovereign. Nationalists were, moreover, for the first time given departmental office, and the zone served as a refuge for those who had had to leave the French area. In 1956, however, when the French decided to grant independence to Morocco, the Spanish authorities were taken by surprise and hesitated before following suit. A corresponding agreement was nevertheless reached on April 7, 1956, and was marked by a visit of the Sultan to Spain. The Spanish protectorate was thus brought to an end without the troubles that marked the termination of foreign control in the French Zone. With the end of the Spanish protectorate and the withdrawal of the Spanish high commissioner, the Moroccan *khalīfah*, and other officials from Tétouan, the city became again a quiet, provincial capital. The introduction of the Moroccan franc instead of the peseta as currency, however, caused a great rise in the cost of living, and difficulties were also caused by the introduction of French-speaking Moroccan officials. In 1958–59 these various changes gave rise to disorders in the Rif. Tangier, too, lost much of the superficial brilliance that it had developed as a separate zone, as by degrees its special status was modified and its privileges were withdrawn. As in the former French Zone, there ensued a great drop in the number both of European and of Jewish inhabitants, while the fact that Ceuta (76,000 in 1966) and Melilla (80,000 in 1965) remained Spanish possessions, with an overwhelmingly Spanish population, provoked periodic Moroccan discontent. The southern protectorate area of Tarfaya was handed back to Morocco in 1958, while Ifni, in return for which the Spaniards had vainly hoped to gain a recognition of their right to Melilla and Ceuta, was finally unconditionally handed over in 1970.

Independence for Morocco

LIBYA

In 1834 the Pasha Yusuf abdicated and was succeeded by his son, Ali II. In 1835 a dispute over this succession provided the pretext for resuming direct rule of Tripolitania and Cyrenaica, which the Ottoman government thought to be the only sure means of averting a European occupation. For 77 years the area was administered by officials from Istanbul and shared in the limited modernization common to the rest of the empire. In Libya the most significant event of the period was the creation of the Sanūsīyah, an Islāmic order, or fraternity, that preached a puritanical form of Islām to the Bedouins, giving them instruction and material assistance and so creating in them an added sense of unity. The first Sanūsī *zāwiyah*, or lodge, was established in 1843 at al-Bayḍā', near the ruins of Cyrene in eastern Cyrenaica. The order spread principally in that province but also found adherents beyond its borders, particularly in the south. In order to avoid the cosmopolitan influences of the coastal region, the Grand Sanūsī, as the founder came to be called, moved his headquarters to the oasis of Jaghbūb near the Egyptian frontier, and in 1895 his son and successor, Sayyid Muḥammad al-Mahdī, transferred it farther south into the Sahara at Kufrah. Though the Turks welcomed the order's opposition to the spread of French influence northward from Chad and Tibesti, they regarded the political influence it exerted within Cyrenaica with suspicion. In 1908 the Young Turk revolution gave a new impulse to reform; in 1911, however, the Italians, with banking and other interests in the country, launched an invasion.

The Sanūsīyah

The Turks were soon forced out of the country, but the Italians found it more difficult to subdue the native population. By 1914 they had occupied most of the country but during the course of World War I were reduced to a few coastal towns—Zuwārah, Tripoli, and Homs in Tripolitania and Benghazi and Derna in Cyrenaica. At the end of the war, the liberal Italian government tried to cooperate with a rudimentary republic, formed during the fighting, by Tripolitanians advised by Abdurrahman Azzam, later first secretary general of the Arab League. In 1921, however, a more vigorous policy was inaugurated by the Italian governor, Giuseppe Volpi, and much facilitated when the Fascists came into power in Rome the following year. The coastal areas of Tripolitania were subdued by 1923, but in Cyrenaica the Sanūsī resistance, led by Omar al-Mukhtar, was maintained until 1932 and only overcome when the population had been herded into concentration camps with much loss of life.

The Italian government displayed great constructive activity in building fine new towns, with harbours at Tripoli and Benghazi, and admirable roads. Some 30,000 Italian peasant families were then settled on the Gefara Plain of Tripolitania and on the Jabal al-Akhḍar heights of Cyrenaica. The local Muslim nomads were moved nearer the desert and were subject to racial discrimination in such matters as the use of public means of transport. This new order, meticulously organized, lasted until 1940, when Italy entered World War II on the side of the Nazis. In 1942 Cyrenaica was totally evacuated by Axis troops and by the Italian settlers and officials; and in 1943 the last Axis troops also were driven from Tripolitania.

After the war, the future of Libya gave rise to long discussions. In view of the contribution to the fighting made by a volunteer Sanūsī force, the British foreign minister pledged in 1942 that the Sanūsīs would not again be subjected to Italian rule. During discussions, lasting four years, suggestions included an Italian trusteeship, a United Nations trusteeship, a Soviet mandate for Tripolitania, and various compromises. Finally, in November 1949 the United Nations General Assembly voted that Libya should become a united and independent kingdom not later than January 1, 1952.

Libyan independence

A constitution creating a federal state with a separate parliament for each province was drawn up, and the pro-British Sanūsī chieftain Muhammad Idris al-Mahdi was chosen king by a national assembly in 1950. On December 24, 1951, King Idris I as-Sanoussi declared the country independent. Political parties were banned and the King's authority was fundamental. Though not themselves Sanūsīs, the Tripolitanians accepted the monarchy largely in order to profit by the British promise that the Sanūsīs would not again be subjected to Italian rule. King Idris, retiring by disposition, showed a marked preference for living in Cyrenaica, where he built a new capital on the site of the original Sanūsī *zawiyah* at al-Bayḍā'. Though Libya joined the Arab League in 1953 and in 1956 refused British troops permission to land for the Suez Canal expedition, the government in general adopted a pro-Western point of view in international affairs. After being entirely dependent, for seven years, on international aid and on the rent for an American air base and a British air training base, the discovery of very rich oil deposits in 1959 both in Tripolitania and Cyrenaica, within easy range of the seacoast, assured Libya of resources on a vast scale. There followed an enormous expansion of all government services and of building and a corresponding rise in the economic standard and the cost of living.

What would happen on the King's death had always been a matter for speculation since his heir was without his uncle's outstanding qualities. The eventuality was anticipated early in 1969 when a group of young army officers led by Col. Muammar Qadafi seized power while the King was undergoing a cure in Turkey. The new regime was passionately Arab and puritanically Muslim, forbidding the sale of alcoholic liquor. Breaking the former close relations with Britain, the new government formed an alliance with the United Arab Republic and The Sudan and cancelled a very extensive order for arms that the previous government had placed in Britain, acquiring arms instead from France as well as from the Soviet Union. It also won withdrawal of U.S. and British bases and adopted a stronger Arab line toward Israel.

MAURITANIA AND THE SPANISH SAHARA

When the French completed the occupation of Morocco in 1934, they sent a force south to link up with another

expedition advancing from their West African colonies. The intervening country was constituted as a territory of French West Africa under the name of Mauritania. Its population was Arabic-speaking in the north; the Negro inhabitants of the south spoke various African languages. In 1946 it became an overseas territory of France.

Independent Mauritania (1960–70). Attaining independence in 1960, Mauritania was admitted to the United Nations in 1961. It lies between Morocco on the north, the Spanish Sahara and the Atlantic on the west, Senegal on the south, Mali on the east, and Algeria on the northeast. Formerly Berber-speaking, it was the homeland of the medieval Almoravid dynasty. It was converted to Arabic speech in the 14th century with the coming of an Arab tribe, the Banū Maʿqīl. Because of the longer period of French rule over them, the Negro population of the south gained a certain advantage in modern education over the Arabic-speaking north, which, however, had a high tradition of Islāmic learning. The form of Arabic used is known as Ḥassāniyah. The political supremacy of the Arabs has resulted in an unusual class structure: warriors of Arab origin; a scholarly and religious class known as *zwaya* of Berber origin; and *harratin* or workers of mainly Negro, and originally slave, origin.

The Arabic speakers, who form 80 percent of the population, tend to insist that Arabic should be an essential qualification for government service; this has led to periodical disturbances. The country's chief commercial asset is the iron ore at Fort-Gouraud. Before the coming of the French, the Arabic-speaking population looked to the sultan of Morocco for defense against foreign rule and as their religious head. For this reason Morocco refused to recognize Mauritanian independence until 1969. In that year, King Hassan invited the Mauritanian president, Mokhtar Ould Daddah, to attend an Islāmic summit meeting in Rabat, foreshadowing a united front of Morocco, Algeria, and Mauritania against the continuation of Spanish rule in the Sahara. Though not a member of the Arab League, Mauritania shares the Arab attitude with regard to the Palestine issue.

The Spanish Sahara (1884–1970). Spanish connections with this area date back to the occupation of the Canary Islands in the 15th century. The present Spanish claim stems from the interest aroused in the area at the time of the European "scramble for Africa," between 1870 and 1880. In 1884 the Spanish government laid claim to a protectorate over land between Cape Blanco and Cape Bojador and occupied the site of Villa Cisneros in the Río de Oro. Güera, adjoining Ouadibou (Fort Saint-Étienne), was occupied in 1920, and Semara in the Saquiat el-Hamra in 1934. The total area is about 100,000 square miles but has an indigenous population of only some 60,000, half of whom are nomads who live there only in the rainy season. Considerable phosphate deposits make it a valuable territory, which Spain retained after 1960 in spite of the claims of Morocco and Mauritania. In 1969 the Spanish government sought to come to an agreement with Morocco for the joint exploitation of the deposits, while themselves retaining sovereignty, to be based on a plebiscite of the very few and impoverished inhabitants. Anti-Spanish disturbances in el-Aïoun in July 1970 made it seem unlikely that Spain would be able to present a convincing case for retention of the territory to the United Nations, with which the question of the future of the area is periodically raised.

(N.B.)

BIBLIOGRAPHY

Ancient North Africa: S. GSELL, *Histoire ancienne de l'Afrique du Nord*, 8 vol. (1913–28), a magisterial account in full detail up to the beginning of the Roman Empire (although the archaeological sections are now outdated, Gsell's intuition has frequently proved correct); C.A. JULIEN, *Histoire de l'Afrique du Nord*, vol. 1, *Des origines à la conquête Arabe (647 ap. J.C.)*, rev. by C. COURTOIS, 2nd ed. (1951–52, reprinted 1966), much briefer than Gsell, but with an extensive modern bibliography. (*Prehistory*): C.B.M. MCBURNEY, *The Stone Age of Northern Africa* (1960), *The Haua Fteah (Cyrenaica) and the Stone Age in the South-East Mediterranean* (1967), two works by one of the foremost excavators and specialists in the subject; J.M. COLES and E.S. HIGGS, *The Archaeology of Early Man* (1969), the most recent survey comparing the prehistory of North Africa with that of the rest of Africa. (*Carthaginian*): B.H. WARMINGTON, *Carthage*, rev. ed. (1969), a detailed historical study of Phoenician Carthage; S. MOSCATI, *Il mondo dei Fenici* (1966; Eng. trans., *The World of the Phoenicians*, 1968), largely concerned with the archaeology, art, and religion of the Phoenicians; P. CINTAS, *Manuel d'archéologie punique* (1970–); G. and C. CHARLES-PICARD, *La Vie quotidienne à Carthage au temps d'Hannibal* (1958; Eng. trans., *Daily Life in Carthage at the Time of Hannibal*, 1961), an imaginative interpretation, but well based; G. CAMPS, *Aux origines de la Berbérie: Massinissa, ou les débuts de l'histoire, Libyca*, vol. 8 (1960), a detailed if sometimes speculative account of the Libyan peoples prior to and during the reign of Masinissa; J. DESANGES, *Catalogue des tribus africaines de l'antiquité classique à l'ouest du Nil* (1962). (*Roman*): T.R.S. BROUGHTON, *The Romanization of Africa Proconsularis* (1929), the basic study of the impact of Roman immigration and culture in the first two centuries AD; P. ROMANELLI, *Storia delle province Romane dell'Africa* (1959), the most recent complete account of Roman North Africa to the Vandal conquest; G. CHARLES-PICARD, *La Civilisation de l'Afrique romaine* (1959), an excellent interpretive study; P. SALAMA, *Les Voies romaines de l'Afrique du Nord* (1951); R.M. HAYWOOD, *Roman Africa*, in T. FRANK (ed.), *An Economic Survey of Ancient Rome*, vol. 4 (1938), a review of all the evidence for the economy of Roman Africa up to AD 284. (*Late Roman, Vandal, and Byzantine*): W.H.C. FREND, *The Donatist Church: A Movement of Protest in Roman North Africa* (1952), an original and controversial treatment of the economic and social significance of Donatism; C. COURTOIS, *Les Vandals et l'Afrique* (1955), the only major account of the subject; C. DIEHL, *L'Afrique byzantine* (1896), in spite of its age, still the standard account of the period; E.F. GAUTIER, *L'Islamisation de l'Afrique du Nord: Les Siècles obscurs du Maghreb* (1927), an outstanding if controversial account of the end of Byzantine rule and the Arab conquest. (*Cyrenaica*): F. CHAMOUX, *Cyrène sous la monarchie des Battiades* (1953); O. BATES, *The Eastern Libyans* (1914); R.G. GOODCHILD, "The Roman and Byzantine Limes in Cyrenaica," in *Journal of Roman Studies*, 43:65–76 (1953); C. LACOMBRADE, *Synésios de Cyrène, hellène et chrétien* (1951).

North Africa from the beginning of the Islāmic period until 1830: C.A. JULIEN, *Histoire de l'Afrique du Nord*, vol. 2, *De la Conquête arabe à 1830*, rev. by R. LE TOURNEAU, 2nd ed. (1951–52, reprinted 1966); C. DIEHL and G. MARCAIS, "Le Monde muselman au XI^e^ et au XII^e^ siècle," in vol. 3 of *Histoire du Moyen âge*, in the series "Histoire Générale," publ. under the direction of G. GLOTZ (1936). (*On the history of Morocco*): H. TERRASSE, *Histoire du Maroc des origines à l'établissement du Protectorat français*, 2 vol. (1949–50; abridged Eng. ed., *History of Morocco*, 1952). (*On the history of Algeria*): S. GSELL, G. MARCAIS, and G. YVER, *Histoire de l'Algérie* (1929). (*On the history of Tunisia*): A. BASSET *et al.*, *Initiation à la Tunisie* (1950), see esp. the chapters by R. BRUNSCHVIG, "La Tunisie du moyen âge," and J. PIGNON, "La Tunisie Turque et Husséinite." (*Studies*): A. BEL, *La Religion Musulmane en Berbérie* (1938); R. BRUNSCHVIG, *La Berbérie Orientale sous les Hafsides des Origines à la fin du XV^e^ siècle*, 2 vol. (1940, 1947); G. MARCAIS, *La Berbérie Musulmane et l'Orient au moyen âge* (1946). Numerous detailed studies may be found in the *Encyclopedia of Islam* (1908–38; new ed. 1954–); and in the journals *Hespéris* and *Revue Algérienne des sciences juridiques, économiques et politiques*.

North Africa since 1830: C.A. JULIEN, *Histoire de l'Afrique du Nord*, 2 vol., 2nd ed. (1951–52, reprinted 1966), with an extensive bibliography, *L'Afrique du Nord en marche* (1952); NEVILL BARBOUR (ed.), *A Survey of North West Africa (the Maghrib)*, 2nd ed. (1962), an authoritative account of modern conditions; ROGER LE TOURNEAU, *Évolution politique de l'Afrique du Nord Musulmane, 1920–1961* (1962), and C.F. GALLAGHER, *The United States and North Africa: Morocco, Algeria, and Tunisia* (1963), clear and informative accounts; *The Middle East and North Africa*, and *Annuaire de l'Afrique du Nord*, two annuals with current information; ALAL AL-FASI, *The Independence Movements in Arab North Africa*, trans. by H.Z. NUSEIBEH from the Arabic (1954), a work representing the viewpoint of a Moroccan nationalist leader; N. EPTON, *Journey Under the Crescent Moon* (1949), a travel book including conversations with French and Muslim leaders at a time when the latter were still little known outside their own countries. (*Algeria*): P.J.L. AZAN, *L'Émir Abd el Kader* (1925), an admirable account with a good bibliography; AUGUSTIN BERNARD, *L'Algérie* (1931), a clear presentation of the view of a liberal Frenchman of the period; C.A. JULIEN, *Histoire de l'Algérie contemporaine* (1964–), a

detailed and scholarly account, to be completed in three volumes; P. BOURDIEU, *Sociologie de l'Algérie* (1958; Eng. trans., *The Algerians*, 1962); and GERMAINE TILLION, *L'Algérie en 1957* (1957; Eng. trans. *Algeria: The Realities*, 1958), two well-informed sociological works; M.K. CLARK, *Algeria in Turmoil: A History of the Rebellion* (1959), a detailed account of the early stages of the independence struggle by an American journalist in sympathy with the colonists; MOULOUD FERAOUN, *Journal, 1955–1962* (1962), the diary of a French-educated Muslim who was assassinated by French terrorists in March 1962; M. LACHERAF, *L'Algérie: Nation et société* (1965), a work representing a Muslim viewpoint; D.C. GORDON, *The Passing of French Algeria* (1966), a well-informed work, anticipating a continued French influence in a noncolonial form; ZOUBEIDA BITTARI, *O mes soeurs musulmanes, pleurez!* (1964), a moving story of a rebel against feminine subjection. (*Tunisia*): A.M. BROADLEY, *The Last Punic War: Tunis Past and Present*, 2 vol. (1882), an account of the French occupation in 1881; ANDRE RAYMOND, *La Tunisie* (1961); and W. KNAPP, *Tunisia* (1970), two scholarly accounts dealing principally with recent times; SALAH-EDDINE TLATLI, *Tunisie Nouvelle* (1957), an informative and readable account by a Tunisian supporter of the new regime; N.A. ZIADEH, *The Origins of Nationalism in Tunisia* (1962); C.A. MICAUD, *Tunisia: The Politics of Modernization* (1964); HABIB BOURGUIBA, *La Tunisie et la France* (1954), and *Propos et Entretiens* (1960), letters and talks by the president of Tunisia. (*Morocco*): EUGENE AUBIN, *Le Maroc d'aujourd'hui* (1904; Eng. trans., *Morocco of Today*, 1906), a lucid description of the country and government before the protectorate; F.R. FLOURNOY, *British Policy Towards Morocco in the Age of Palmerston (1830–1865)* (1935); H. TERRASSE, *Histoire du Maroc des origines à l'établissement du Protectorat français*, 2 vol. (1949–50); J. BRIGNON *et al.*, *Histoire du Maroc* (1967); GEORGES CATROUX, *Lyautey, le Marocain* (1952); ROM LANDAU, *Moroccan Drama, 1900–1955* (1956), a good journalistic account; J. and S. LACOUTURE, *Le Maroc à l'épreuve* (1958), a study of the early days of independence; ROGER LE TOURNEAU, *Fès avant le Protectorat: étude économique et sociale d'une ville de l'Occident musulman* (1949); W.B. HARRIS, *France, Spain and the Rif* (1927), an account of the Rif War by a correspondent of *The Times*; R. REZETTE, *Les Partis politiques marocains* (1955); T. GARCIA FIGUERAS, *Marruecos: La acción de España en el Norte de África* (1939), a general account, mostly political; J. BECKER, *España y Marruecos: Sus relaciones diplomáticas durante el siglo XIX* (1903); E. ARQUES, *Tres Sultanes a la porfía de un reino* (1952), an account of the pretender, the Rogui, by a captive Spaniard. (*Tangier*): G.H. STUART, *The International City of Tangier*, 2nd ed. (1955). (*Libya*): G.F. ABBOTT, *Holy War in Tripoli* (1912), an account, by a war correspondent, of the Italian attack on the Turks in Tripolitania; R. GRAZIANI, *Pace Romana in Libia* (1937), an account of the defeat of the Libyans after World War I by the Italian commander in chief, Marshal Graziani; E. ROSSI, *Storia de Tripoli y della Tripolitania* (1968); J. DESPOIS, *La Colonisation italienne en Libye: Problèmes et méthodes* (1935); M. KHADDURI, *Modern Libya: A Study in Political Development* (1963); R. OWEN, *Libya: A Brief Political and Economic Survey*, rev. ed. (1961). (*Mauritania*): O. DU PUIGANDEAU, *Le Passé maghrébin de la Mauritanie* (1962), evidence for the past connections of Mauritania with Morocco; G.M. DESIRE-VUILLEMIN, *Contribution à l'histoire de la Mauritanie de 1900 à 1934* (1962), a definitive work, with bibliography; A.G. GERTEINY, *Mauritania* (1967), contains much information and an extensive bibliography, but hurriedly composed, with some gross errors; H.T. NORRIS, *Shinqītī Folk Literature and Song* (1968), a study of folk literature, but valuable historical and other background matter is included. (*Spanish Sahara*): J. CARO BAROJA, *Estudios Saharianos* (1955); T. GARCIA FIGUERAS, *Santa Cruz del Mar Pequeña–Ifni–Sahara* (1941), a historical account; JOHN LODWICK, *The Forbidden Coast: The Story of a Journey to Río de Oro, a Spanish Possession in North-West Africa* (1956), the only traveller's account of this area of Africa.

(B.H.W./H.-L.-É.T./N.B.)

North America

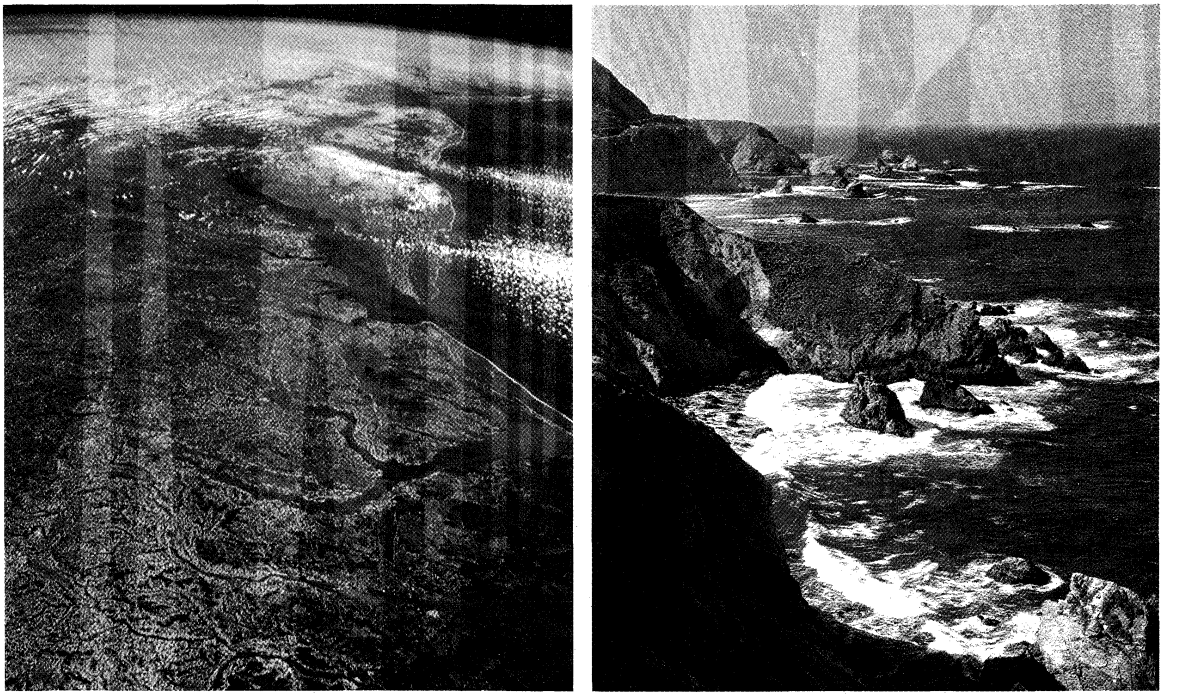
One of the most developed regions of the world, the North America of the 1970s enjoys an average income per person nearly twice as high as that of Europe; a food intake nearly a third greater than the average for Asia; and a per capita consumption of energy five times as great as the average of all the other continents. With an area of some 9,420,000 square miles (24,400,000 square kilometres) shaped like a huge inverted triangle, it is third in size among the world's continents. The home of over 320,000,000 people, or only about 8 percent of the world's population, it nevertheless produces over a third of all the factory-made goods of the globe. Yet, before the coming of European settlers and their African slaves, North America was not nearly as fully developed as the then contemporary Europe or Asia or parts of Africa. The subsequent extraordinary efflorescence can be attributed, in the opinion of many North Americans, to the free and dynamic nature of their societies, led by that of the United States. European colonizers did indeed set sail at a time of great scientific and technical innovation and were able to apply and develop new skills in lands untrammelled by the conventions and practices of the past, but it was North America's great natural wealth that accounted for much of the expansion, offering as it did extensive and, at the time, seemingly inexhaustible deposits of metals and fuels, vast forests, ample water resources, and a wide and stimulating range of climate and soils.

The name America is derived from that of the Italian merchant and navigator Amerigo Vespucci, one of the earliest explorers of what is now North America. He is thought by some authorities to have discovered and reached the mainland of the continent in 1497. There is no doubt that this claim was widely accepted after the publication, in 1507, of an account of his travels. The newfound lands that came to bear his name extended from Labrador in the north to Patagonia in the south, and it became convenient to designate those portions that widened out north of the Isthmus of Panama as North America, reserving South America for those that broadened to the south. It should be noted that some authorities take North America to begin, not at the Isthmus of Panama, but at the narrows of Tehuantepec, calling the intervening region Central America. Under such a definition, part of Mexico must be included in Central America, although that country lies mainly in North America proper. To overcome this anomaly, the whole of Mexico, together with Central and South American countries, may also be grouped under the name Latin America, with the United States and Canada referred to as Anglo-America. This cultural division is a very real one; yet Mexico and Central America (including the Caribbean) are bound to the rest of North America by strong ties of physical geography. Largely Danish-speaking Greenland is also divided culturally from, but connected physically with, North America. Finally, all the above areas are sometimes included under the general terms New World, the Western Hemisphere, or the Americas.

The naming of the continent

The first inhabitants of North America came from Asia, mostly by way of the Bering Strait out of eastern Siberia. Over the centuries, they had moved north and east away from the main centres of Old World civilization. They were comparatively primitive when they arrived and remained Stone Age hunters and fishermen for a very long time; indeed, many were still in this state when the first Europeans saw them. Their societies were not without interest in themselves, but most of them were ignorant of, or saw no use for, the wealth of metals and fuels at their disposal; they failed to use the vast waterpower resources and hardly tapped the timber reserves of the continent. Agriculture was only developed south of the Ottawa River and east of the Missouri Valley. Cities eventually appeared on the plateaus of Guatemala, Yucatán, and Mexico. Despite the great strides of the civilizations developed by the Mayans, Toltecs, and Aztecs, those peoples showed relatively little mechanical aptitude or commercial enterprise. They were no match for the Spaniards, who were scarcely lacking either attribute and who thus soon took over, conquering Mexico in 1519. The Spaniards were followed by the French, who set up a colony at Port Royal, in what was to become Nova Scotia, Canada, in 1605; and by the British, whose first Virginia settlement, at Jamestown, in 1607, was closely followed by that at Plymouth, Massachusetts, in 1620.

The Europeans soon began to piece together their own sense of the geography of the continent. In the south, the



Coastal Contrasts.
(Left) Low-lying eastern coastline of the United States, looking north from Virginia. Photographed from Apollo 9 spacecraft. (Right) Rugged coastline of the western United States at Big Sur, California.
By courtesy of (left) NASA; photograph, (right) Esther Henderson—Rapho Guillumette

Spaniards discovered a mountain arc swinging through the West Indies into Central America, together with a low, coastal plain around the Gulf of Mexico reaching up to lofty mountains bordering the high Mexican Plateau. Explorations revealed that the Mexican Sierras continued north in the Rocky Mountain and Cascade ranges. The British, meanwhile, had crossed the Atlantic Coastal Plain, a continuation of the Gulf Coastal Plain, penetrated the barrier of the Appalachian Highlands, and pushed into the vast Interior Basin centred on the Great Lakes. The French bypassed the Appalachians by sailing up the Gulf of St. Lawrence and, skirting the huge and inhospitable Laurentian Plateau to the north, moved deep into the interior, pressing on to discover the Great Plains skirting the Rockies. In the Far North, the British ventured into the Canadian Archipelago and explored the Mackenzie Basin down to the Arctic Ocean. Sailing up the Pacific coast, they circumnavigated Vancouver Island and penetrated the great Coast Range. In the space of 200 to 300 years, the European explorers had thus traced out the main limits of the North American continent.

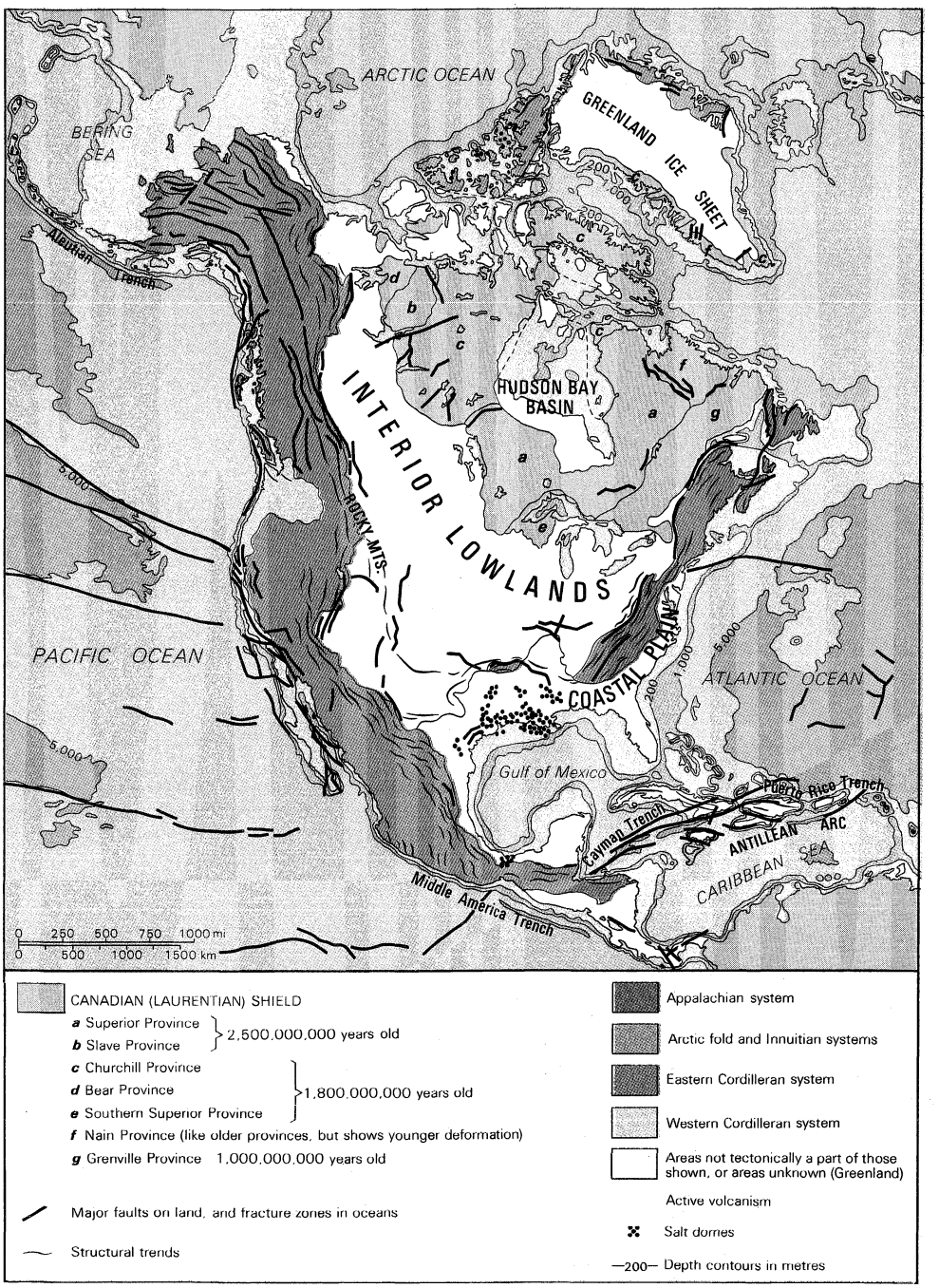
In due course, the Europeans also discovered a number of things about the continent: that the coastal and interior lowlands offered great scope for agriculture; that extensive reserves of timber occurred on the Appalachians, the Laurentian uplands, the northwestern Pacific ranges, the Southern Sierras, and the Central American mountains; that huge coalfields flanked the western Appalachians and underlay the Ohio and Central Mississippi basins; that rich deposits of iron lay on the edges of the Laurentian Plateau; that nonferrous ores were abundant in the Rockies and Western Sierras; and that oil and natural gas lay in pools under the Great Plains and the Gulf of Mexico lowlands.

These great resources were at first developed by the European empires: Spanish colonies in the southern parts of the continent; British colonies from Georgia to Newfoundland between the Atlantic shore and the Appalachians; and French colonies in the St. Lawrence–Great Lakes Basin and down the Mississippi in the continental interior. Later, the Russian Empire was to take over Alaska. North American societies, in essence, became European outposts in the New World, and they retain very strong ties with European cultures. Subsequently, first the United States, then Mexico and the Central American countries won their independence; the power of France gave way before that of British North America, and Canada became a self-governing country within the British Commonwealth. Swift progress was then made in developing North American styles of life, which became strongly Latin American south of the Rio Grande and Anglo-American to the north; slavery and its heritage added a significant minority culture of African origin. With this varied and often turbulent legacy, the continent became a powerful centre of world influence, instead of being on the margins of the civilized world, as in Indian times. The position of this new major power base between the two most densely populated regions of the Earth—western Europe and eastern Asia—has, for good or ill, given North America and its peoples immense global significance.

Evolution of North American societies from colonial to independent status

This article covers, in detail, the geological history and physical environment of the continent and offers an overview of its resource base, including human resources, and the patterns of economic and social development. Although this article covers the physical evolution of North America in terms of that larger definition of the continent which includes Greenland and the Caribbean islands, the sections covering human resources offer a conspectus primarily of mainland North America. There are separate articles on the major physical features of the continent; on its nation states and on their history; on the Canadian provinces and territories and the individual states of the United States; as well as articles on the major cities of the region; see also CONTINENTS, DEVELOPMENT OF.

This article is divided into the following sections:



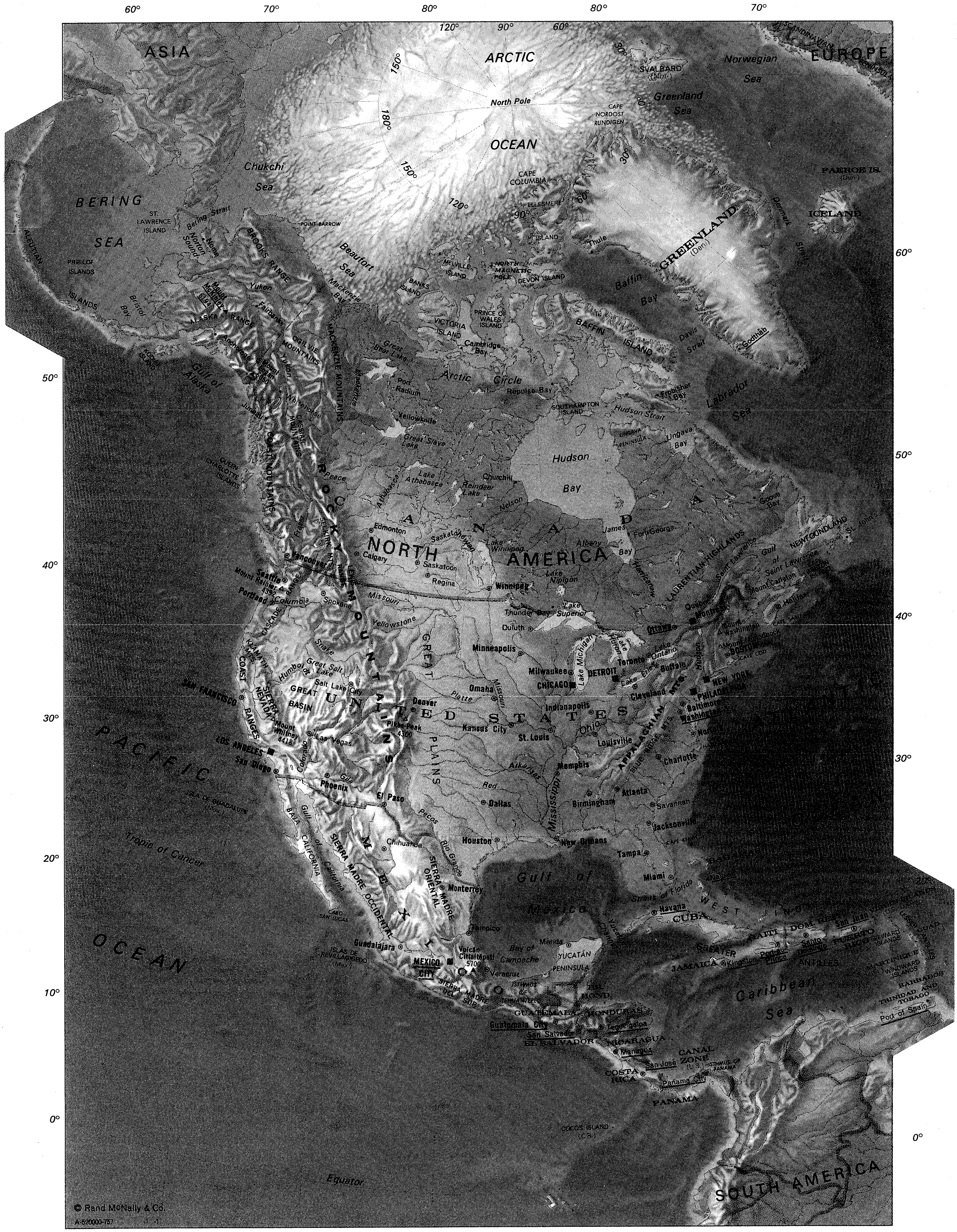
Structural features of North America.

I. Geological history

North America has a latitudinal span of over 76°, stretching from southern Panama to Cape Columbia, in northern Canada, and extends over 175° of longitude from Cape Nordost Rundigen, in eastern Greenland, to Attu Island, in Alaska. This enormous reach gives it every climate from the virtually equatorial to the Arctic, but the shape of the continent, very narrow in the south and very wide in the north, reduces the amount within the tropics and expands the area in the temperate and polar zones.

ELEMENTS AND PROCESSES IN THE MAKING OF THE CONTINENT

The shape of North America may have resulted from the fact that, over a vast period of geological time, the continental landmass drifted away from Europe and North Africa. A glance at the globe shows that the vast northern edge of North America would fit into the correspondingly receding edge of Europe and that its narrowness in the



NORTH AMERICA

Size of symbol indicates relative size of town • ◎ ■

Elevations in metres

0 400 800 1200km

0 200 400 600 800 1000 1200mi

south could correspond to the huge bulge of western Africa. Geologists have worked out a theory of continental drift based on rigid areas in the crust, like plates, with zones of weakness in between. These are affected by currents in the molten rock material, or magma, beneath the Earth's crust. Immense pressures on the so-called plates led them to move into new positions, squeezing out the weak zones, marked by deep depressions, or geosynclines, on their margins. These became filled in with sediments and were then squeezed out and upward into fold mountains. Faulting—a vast shearing process along lines of structural weakness in the Earth's crust—occurred subsequently between continents as they were torn apart from each other. Faulting was followed by renewed and extensive folding on the outward advancing continental fronts. In the formation of North America, it is thought that a very stable area called the Canadian or Laurentian Shield parted from the Baltic and Iberian shields of Europe in Jurassic times, about 170,000,000 years ago. The ensuing gradual widening of the Atlantic Ocean led to extensive faulting along the east coast of North America; and the push of the continent against the "plate" underlying the Pacific area produced widespread folding in the west.

Other geologists have sought to explain the evolution and growth of North America in terms of its present position, first by a hardening of the Earth's crust through immense granitic intrusions, or upswellings of molten rock, which formed the Laurentian Shield, and then by the rise of great geosynclines where sediments, stripped from the shield over millennia by the forces of erosion, depressed the margins of the continent. From a stable core, at the heart of the Laurentian Plateau, great downwarps are then thought to have developed in contact with the Atlantic, Arctic, and Pacific oceans. These became so burdened by the debris of the eroded shield as to form lines of crustal weakness, which, in turn, grew into zones of volcanic activity and fold-mountain systems. The continent thus evolved by the outward expansion and final immobilization of marginal geosynclines. Indeed, there is an outward regression in age from a very old centre, where the rocks are nearly 4,000,000,000 years old, to younger margins, usually less than 600,000,000 years.

Conflicting theories of continental origin

Both of these theories have attractive features: each accounts for a central shield flanked by belts of marginal mountains. The extensive faulting in the north and east, separating Greenland from Canada, and the breaking off of the apparent connections between the Appalachians of North America and the Caledonian and Hercynean mountain-fold systems of Europe might well correspond to a shore torn from the Old World by the widening of the Atlantic and Arctic oceans. Similarly, the lofty region of mountain building in the west would correspond to an advance on the Pacific basin. On the other hand, the outward progression of geosynclines, formed by the weight of waste material washed from a central core and turned into lines of marginal mountains, would also explain existing North American structural patterns adequately.

Recent work on the Atlantic, emphasizing active widening by periodic vulcanism and faulting, strongly supports the drift theory. Ocean-floor spreading—a movement away from the centre line of the ocean basin, leaving a great midocean ridge, flanked by deeper parts, and often ending in major depressions toward the continents—is seen as the mechanism for the breaking up of the continents. It has been suggested that the North Atlantic began to open up in the Upper Triassic and the Lower Jurassic, between 187,000,000 and 172,000,000 years ago, while the South Atlantic started to widen in the Upper Jurassic, about 138,000,000 years ago. The major rifting of Europe and North America was not accomplished until the Lower Cretaceous, about 93,000,000 years ago. Greenland and Canada started to split up in the Upper Cretaceous, while Greenland and Europe moved apart somewhat later, in the Cretaceous–Paleocene, say 63,000,000 years ago. From that time on, North America became a truly separate continent, still drifting west, it is claimed, at a rate of just over half an inch, or about one and a quarter centimetres, a year. The "plate" flooring the Pacific basin, meanwhile, may have been thrust under the advancing continent.

The four chief structures of North America are a central shield; marginal mountains; interior lowlands between the shield and the mountains; and coastal plains. All have taken a long time to evolve. The shield itself, made up of the oldest rocks in North America, took longer to form than the rest of the continent put together. After its main structures had been built, it became the core area against which the marginal mountains were thrown up. These grew from immense geosynclines, which served as sites for the accumulation of tens of thousands of feet of material, sorted by the seas and later lifted up and squeezed into mountain ranges. The geosynclines divided themselves into (1) a shallower belt, the miogeosynclines, along the edge of the shield, and (2) a very profound belt, the eugeosynclines, farther out. Mountains evolving from the miogeosyncline were less intensively folded and less subject to volcanic activity than those growing out of the eugeosyncline. The Appalachians were the first mountain system to so evolve, followed by the Arctic fold belt and, in turn, by the comparatively young mountains of the west. Ancient seas, meanwhile, lying between the rising mountains and the shield, spread out the debris eroded from surrounding heights. Horizontally bedded sheets of sedimentary rock thus came to form the interior and coastal plains. Faulting, leading to immense fault-bounded blocks or to sea-drowned depressions, frequently interrupted the building up of these structures. Thus the geological history of North America has, right down to this day, provided the framework for its evolving physical geography.

North America's four chief physical structures

THE EVOLUTION OF THE CENTRAL SHIELD

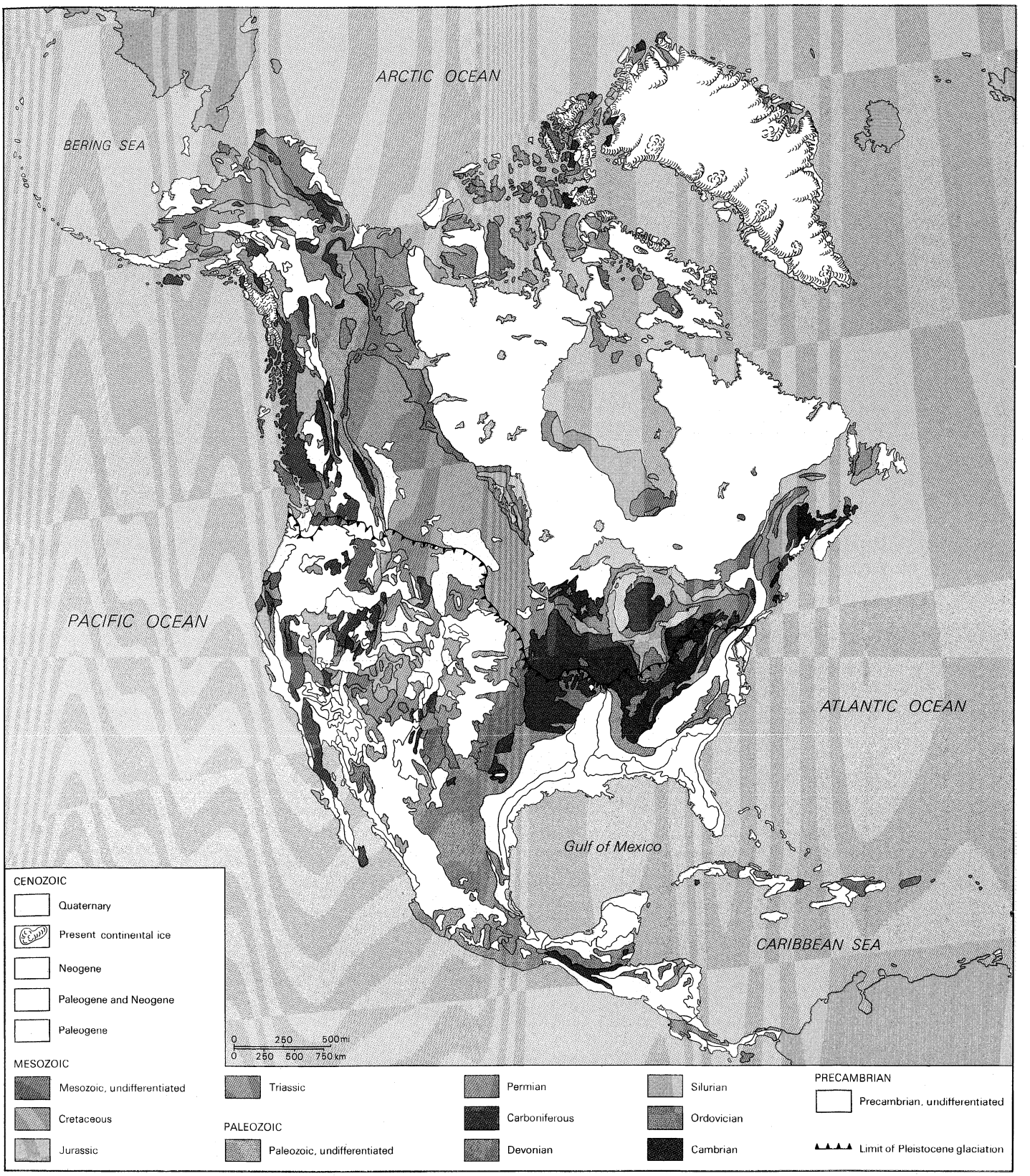
Basic structure. The central shield, variously called Laurentia, the Laurentian Shield, or the Canadian Shield, is the key structure of the continent, extending from the Adirondacks and the Superior Upland just south of Lake Superior, northward through Canada, both east and west of the Hudson Bay, to Greenland. It is, in fact, a group of shields that gradually coalesced, covering a total of 2,550,000 square miles, including 700,000 square miles in Greenland, and is largely composed of Archean and Proterozoic rocks formed in Precambrian times, more than 570,000,000 years ago.

The main trends of the shield sprang from an initial Y-shaped structure identified as beginning west and east of the Hudson Bay depression and then extending, as a subsurface welt, south of the Superior Upland to the Colorado Plateau. This great structure was subsequently tilted up in the east, in Greenland and in Labrador; it was faulted along the south and west to give way to the basins of the St. Lawrence and Great Lakes, of the Red River of the North and Lake Winnipeg, and the Mackenzie and its attendant lakes; and it sank beneath the vast hollow of what is now Hudson Bay to the north. Various accretions, joined to the initial structure, formed the rest of the shield.

The rocks of the shield fall into two main groups, the Archean and the Proterozoic, separated by a long interval of erosion. The Archean formations are subdivided into the Keewatin and Timiskaming, from the Canadian type areas where they were identified. The Proterozoic was accompanied by four great shield-building times, again named from their type areas, the Lower and Middle Huronian, the Animikie (or Upper Huronian), and the Keweenawan.

Development of Archean rocks. The original shield, going back at least 3,900,000,000 years, consisted of a series of volcanoes, with lava tablelands that spread out and filled up innumerable lakes in between. In early Archean, or Keewatin, time (there is great controversy over the dating of this process, and precise figures are meaningless in this context) the initial Y-shaped structure developed from great thicknesses of lava. These flowed out at intervals, were attacked by erosion and coated by weathered material, and then flowed out again, producing sheets of volcanic rock interbedded

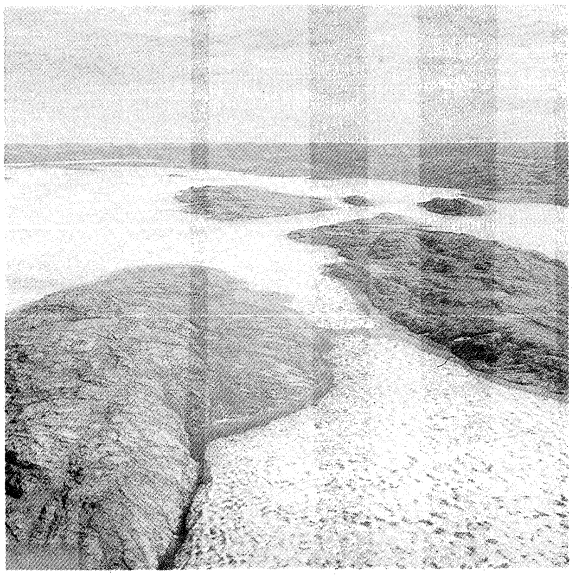
Initial Y-shaped structure of shield



Rock strata of North America.

with sedimentaries. In the late Keewatin, prolonged erosion led to wide deposits of conglomerates (rocks made up of different materials) and sandstones on the western and southern edges of the shield. At a later time, these rocks were caught up in the Laurentian mountain-building movements, and, as a result, they were heavily intruded from below by granitic domes.

The shield then became faulted along its southeastern edge, forming a long trough where thick layers of limestone interbedded with shale (typified by deposits at Grenville, near Ottawa) were laid down. More changes took place in Timiskaming time, with the growth of the iron-bearing southern part of the shield, in the Superior Upland. During the structural upheaval known as the Algoman revolution, these rocks west and south of Lake Superior and along the north shore of Lake Huron were



Frozen sea inlet in the crystalline rocks of the Laurentian Shield, Boothia Peninsula, Northwest Territories, Canada.
George Hunter—Shostal

widely metamorphosed (or heat altered). Mountains also arose in the vicinity of Great Slave and Great Bear lakes, and the dislocations threw lava beds and sedimentary strata into nearly vertical folds, accompanied by large granitic domes.

Development of Proterozoic rocks. In the Eparchean interlude—a long stretch of time in which the mountains were worn almost completely down—the shield became virtually a plain. But, as the Proterozoic dawned possibly 1,500,000,000 years ago, lava flows occurred, indicating further crustal disturbances. The Lower Huronian series of lavas, conglomerates, and limestones were deposited in the northern part of the Great Lakes depression. Continued activity in the Middle Huronian raised these deposits about 1,700 feet, tilted the shield to the north, and laid down new deposits, preserved in the Cobalt and Chibougamau basins and in the Gulf of Richmond, on its inner edge. Further enlargements occurred in the Lake Athabasca–Darnley Bay area, in the northwest, and the Labrador–Baffin Island–Greenland region, to the east. Upper Huronian, or Animikie, time began with strong erosion that spread iron-associated deposits into flanking depressions from Lake Athabasca around to the Ungava trough. As erosion continued, the burden of material affecting the crust beneath led to volcanic outbreaks and the uplift and squeezing out of the geosynclines. The end of Animikie time saw new injections of igneous masses and the rise of fold mountains in the Great Slave Lake, the Belcher Islands, and the Ungava–Baffin regions.

Final growth of the shield

The final growth of the shield occurred in Keweenawan times. Its southern extension was enlarged by more iron-rich sedimentaries; the great Sudbury norite (granular rock) intrusion was formed; Lake Superior became a distinct elongated trough; and, in the Far North, the great Victoria trough saw a downwarping that gathered the Coppermine series of sandstones and lavas. Further shield extensions into the Arctic islands and Greenland continued. Keweenawan times ended in the massive Killarnean intrusions, east of Sudbury, and in renewed mountain building, especially just north of Lake Athabasca.

By the end of the Proterozoic, at about 570,000,000 years ago, when long periods of erosion had worn down the heights and built out extensive plains, the shield was a virtual lowland stretching from Greenland through to the Great Plains.

Later evolution of the shield. The shield never grew any bigger, but its chronicle of changes was by no means over. With the rise of the Appalachians against its southeastern edge in Ordovician times, about 480,000,000 years ago, the shield sank down in the centre to produce even wider and deeper waters than now occupy Hudson Bay. Further uplift in Middle Silurian times, however, raised the shield so high that Hudson Bay disappeared. Sinking followed; but in Devonian times, renewed strain stemmed from the rise of the Acadian mountains. As a result, the eastern rim was tilted up, and extensive faulting occurred. A long quiescence followed, wearing the whole shield down until Cretaceous times. Renewed disturbances, associated with the rise of the Rockies, then lifted up and broke the western shield edge; while between the Late Cretaceous and the Paleocene, about 65,000,000 to 54,000,000 years ago, the Labrador Sea began to open up between Baffin Island and Greenland. In mid-Pliocene, there was a resumption of ocean spreading in the North Atlantic, a general uplift of the shield, and, especially in the east and north, widespread faulting. The shield became depressed in the Pleistocene, over 1,000,000 years ago, under the enormous weight of the ice cap, and in postglacial times subsequently rebounded to its present level, size, and shape, as the massive core of North America.

Postglacial rebounding to present conditions

THE EVOLUTION OF THE MARGINAL MOUNTAINS

At the end of the Proterozoic, about 570,000,000 years ago, the shield was so massive that it then became the most stable area in North America, and subsequent continental deformation shifted to the great fold-mountain systems. The first of these to develop was the Appalachian fold belt, from the Late Cambrian (about 500,000,000 years ago) through to the Pennsylvanian (about 280,000,000 years ago). This first belt was followed by the Greenland and Arctic fold belts from the Late Paleozoic to the Early Cenozoic (from about 345,000,000 to 65,000,000 years ago), overlapping the Appalachians but continuing beyond them both in time and space; these were succeeded in turn by the Western Cordilleras, which began forming in the Jurassic Period (about 180,000,000 years ago) and continued right on into the Miocene (about 10,000,000 years ago), with a few volcanic outbursts in the Pliocene as late as 7,000,000 years ago.

Growth of the Appalachians. Growing out of a complex and varied system of geosynclines along the southeastern margin of the shield (or its buried extension), the Appalachians exhibited striking early contrasts between their western and eastern portions. The western ridges were lifted up from a comparatively shallow depression, or miogeosyncline, while the eastern ranges rose from a much deeper eugeosynclinal trough beyond the continental edge. The western division is made up of more regularly folded rocks with few igneous intrusions; the eastern parts are intensively folded and faulted and have been heavily intruded by crystalline bodies.

In Cambrian times, new disturbances in North America and Europe developed long depressions on the forefront of the central shield into which sands and clays and vegetable and animal remains were carried outward from the old and stable interior and laid down in such quantity as to create persistent, if slow, crustal subsidence. This pressure, in turn, generated a counterthrust that created island arcs of spouting volcanoes or molten igneous domes, which lifted earlier deposits and transformed them into great land swells. Subsequent folding produced the Appalachian Mountains proper, which depressed the edge of the shield into a sunken platform, buried under strata laid down by invading seas.

The geosynclines were squeezed out in a whole series of mountain-building movements, from Cambrian to Pennsylvanian times, or perhaps for as long as 200,000,000 years. The older northern movements were separated from the younger southern activities by a line running approximately through present-day New Jersey. The northern activity ran through the Maritimes and Newfoundland Island to Greenland, with a possible continuation in the Caledonian and Hercynian mountains of western Europe.

Northern Appalachians. In Cambrian times the northern Appalachians formed themselves into three distinct belts, in which were to mold the whole evolution and eventual relief of New England and of Canada's Atlantic

provinces. The first belt was an inner Laurentian geosynclinal trough along the present St. Lawrence River; the second, a central plateau-like upwelt, running through central New England to Newfoundland; and the third, the outer depression of the Acadian Geosyncline. The inner trough may be referred to as the Taconic Geosyncline from the Taconic orogeny, or mountain-building process, during Ordovician times, of old rugged fold mountains running through eastern New York. Further folding, mainly during the Silurian Period, saw the rise of the parallel ridges in western New England, swinging on to western Newfoundland and perhaps to Greenland and beyond. The central portions, however, remained a relatively stable mass, although they were extensively injected with volcanic rocks. During Devonian times, the outer Acadian trough saw the rise of short, fold ridges in eastern New England and the Maritimes, strengthening themselves in eastern Newfoundland before stretching on to possible European continuations. The zone includes a notable series of Triassic red sandstone beds, cut by ridges of volcanic origin.

The basic tripartite division of the northern Appalachians

Southern Appalachians. The younger southern Appalachians began southward of what is now the Hudson Valley, were interrupted by the Mississippi Basin, became active again in the Ouachitas (in Arkansas and Oklahoma), and may have had a Central American continuation to Venezuela. They arose in Mississippian to Pennsylvanian times, from 345,000,000 to 280,000,000 years ago. Like their northern counterparts, they evolved from inner and outer troughs but lacked the intervening geanticline, or huge upwelt. The inner trough was filled by silts and sands from great deltas pushed out between the rising Appalachians and the shield, together with limestones in deeper gulfs; these layers came to typify the Catskill, Allegheny, and Cumberland plateaus. Distinctive short folding produced the undulating topography now characterizing the central belt known as the Ridge and Valley region. The outer trough saw much volcanic activity, which, together with intense folding, helped to mold the hard granites, slates, and quartzites of the Blue Ridge and the Piedmont. Farther south, the Ozarks continued the sweep of the inner plateaus, while the wavelike surface of the Ouachitas pushed against older Texan rocks to form the great Ouachita Geosyncline.

Arctic fold belts. Structural disturbances then moved north, and, in Greenland and the Canadian Archipelago, the Arctic fold-belt mountains began to emerge as the Appalachian orogeny died down. This process perhaps reflected a distinct torque in continental drift, with a northern twist in North America's main move west. All along the east coast of Greenland, Paleozoic rocks, especially those of the Silurian (from 425,000,000 to 400,000,000 years old), were subjected to folding—possible equivalents of the European Caledonian orogeny. A northeastward trend leads, via a submarine ridge, to Svalbard (Spitsbergen). Greenland's northern shore is crossed by mountains with a west–east axis, considered an extension of ranges sweeping across northern Canada. In the Arctic as in the Atlantic, ocean spreading was putting pressure on North America. An Arctic geosyncline resulted, deepening in late Cambrian and Ordovician times (about 500,000,000 years ago) to attract clays and sands eroded from the Canadian and Greenland shields. Massive Cambrian sandstones overlie Precambrian rocks in eastern central Ellesmere Island; similar rocks and sheets of Ordovician shale and limestone slope northwestward off the high Precambrian spine of Baffin Island; and, far to the west, Ordovician strata overlap the shield in Victoria Island. Silurian rocks lie farther north, typically developed in western Devon Island and the Parry Islands and sloping beneath Carboniferous and Permian beds.

The Innuitian mountain system

The Innuitian mountain system, a major fold belt, rose up in this area, made up of rocks of Mesozoic age, from Prince Patrick Island to Ellesmere; folding may well have continued in Ellesmere until the early Miocene, some 25,000,000 years ago, or well into the mountain-building period typical of the Western Cordilleras.

The Innuitians, nevertheless, resemble the Appalachians more than the Cordilleras. Through Eglinton Island, in the west, and across northern Melville and Bathurst islands, they developed along an undulating series of parallel anticlines and synclines, with a distinct ridge and valley effect. The rocks consist of whitish sandstones with bands of black coal overlain by limestone. Pitching anticlines, with their cigar-shaped hills, and synclines, with their trough-topped ridges, became prominent and gave rise to the zigzag drainage patterns so striking in the inner Appalachians. The intensified folding and intrusions of volcanic rocks, however, are more reminiscent of the outer Appalachians.

The Cordilleras. The last, but certainly the greatest, system of mountains to form in North America, the magnificent Cordilleras rose mainly in post-Jurassic times, less than 136,000,000 years ago, when Atlantic ocean spreading had reached a critical point with the accelerated drift of the continent westward against the Pacific. The spreading of the Pacific floor, meanwhile, thrust eastward against the Pacific borderlands. The two forces must have created tremendous pressures all along the west rim of the central shield and so helped to inaugurate the squeeze-out of the giant geosynclines of the west that gave rise to the Rockies, Cascades, and Coast ranges, from Alaska to Mexico.

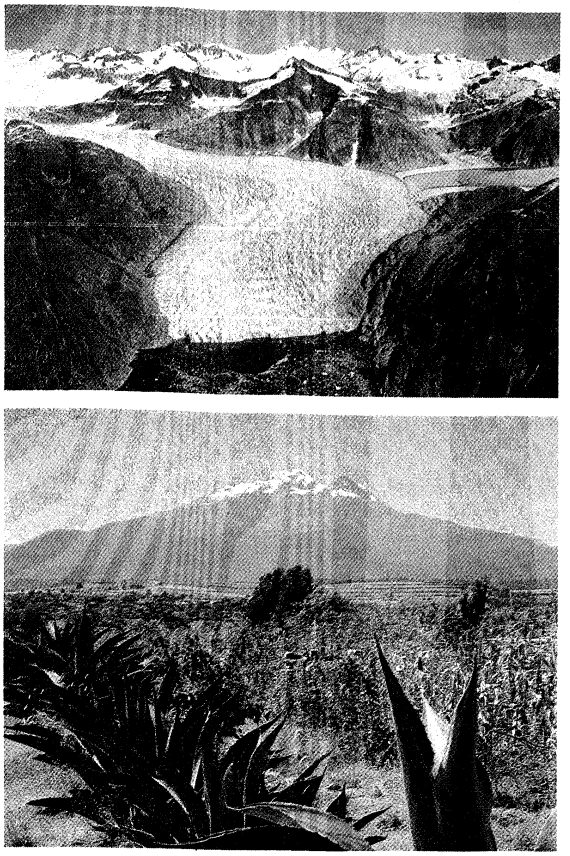
Cordilleran geosyncline. Rocks going back to Proterozoic times accumulated in a deep trough at what must then have been the outermost shield edge. Much erosion continued until in Cambrian times, over 500,000,000 years ago, the immense Cordilleran geosyncline developed from Alaska to Central America. It was not deep and may be regarded as a miogeosyncline, or a belt of troughs, often called the Millard zone of crustal weakness. Westward, the land swelled up into a tremendous geanticline of very old rocks underlying the Yukon, Colorado, and Mexican plateaus. Since this upwelling lay between the main lines of mountain building, it came to be known as the Mesocordilleran Geanticline. Its erosion deposited sediments into the Pacific that put pressure on the earth's crust and generated a deep geosyncline, a very complex belt with an inner and an outer division. This eugeosynclinal trough is known as the Fraser zone.

The tripartite division of the Cordilleras

These basic structures dominated the rise of the Cordilleras, dividing them essentially into three great belts: the Eastern Cordilleras, or Rocky Mountain system, along the inner edge of the old continent; a central string of high intermontane plateaus; and the Western Cordilleras, divided into the Cascadian and the Coast Range systems, on the continent's outer edge. A major difference occurred between the outer and inner mountains. Like those of the Appalachians, the inner mountains arose from a shallow miogeosyncline and consist of long parallel folds or blocks of relatively unfolded strata lifted up on high. The outer mountains, reminiscent of the emergence of the outer Appalachians, originated in a profound eugeosyncline intruded by enormous igneous masses, studded with tall volcanoes—some of them active until modern times—and intensively folded and faulted.

Eastern Cordilleras. In the late Jurassic, over 136,000,000 years ago, the Eastern Cordilleras emerged from the deeper Millard zone and spread inward over the buried shield edge. In the ensuing Upper Cretaceous, the Laramide structural revolution brought forth the western Rockies, spreading debris over what are now the prairies. Later still, in Eocene and Oligocene times, as late as 25,000,000 years ago, the eastern ranges of the Rockies started to rise, marked by slow uplift of great blocks of strata.

Western Cordilleras. The Fraser zone had also become disturbed. The Western Cordilleras exhibited Mesozoic conglomerates and coarse sands that were laid down unconformably, or disjointedly, on the older rocks, thus showing that upwells within the troughs had already been lifted up and then eroded. Movement became intensified during the Nevadan mountain-building revolution of the late Jurassic, raising and folding mountains from Central America to Alaska, including the Sierra Nevadas and the Cascades. Enormous igneous bodies were intruded into the rock layers: the largest was the Coast Batholith, or dome-shaped structure, underlying coastal British Columbia. The line of deformation then shifted outward



Three views of the Western Cordilleras.
(Top left) Mountainous region near Petersburg, Alaska, with glacier-filled valley. (Bottom left) Mexican landscape showing Ixtacihuatl, an extinct volcano on the uplands east of Mexico City. (Above) Sierra Nevada Range of California, seen from Glacier Point in Yosemite National Park. Half Dome, a glaciated bedrock knob, is at left.
(Top left) Ray Manley—Shostal, (bottom left and right) Josef Muench

The Coast Batholith and other massive igneous intrusions

to a trough overlying the present coast ranges of California and Oregon and along the Columbian Trough, represented by the islands off British Columbia and Alaska. Deposition, from inland, overloaded the geosyncline and caused violent volcanic activity. Intense folding, uplift, and faulting followed, with thrusts toward the continental interior mostly in Late Cretaceous and Early Tertiary times, roughly between 100,000,000 and 50,000,000 years ago. The Late Tertiary Cascadian orogeny raised the western ranges yet again, led to renewed volcanic activity from southern Mexico through California to southern Alaska, and severely faulted the Sierras.

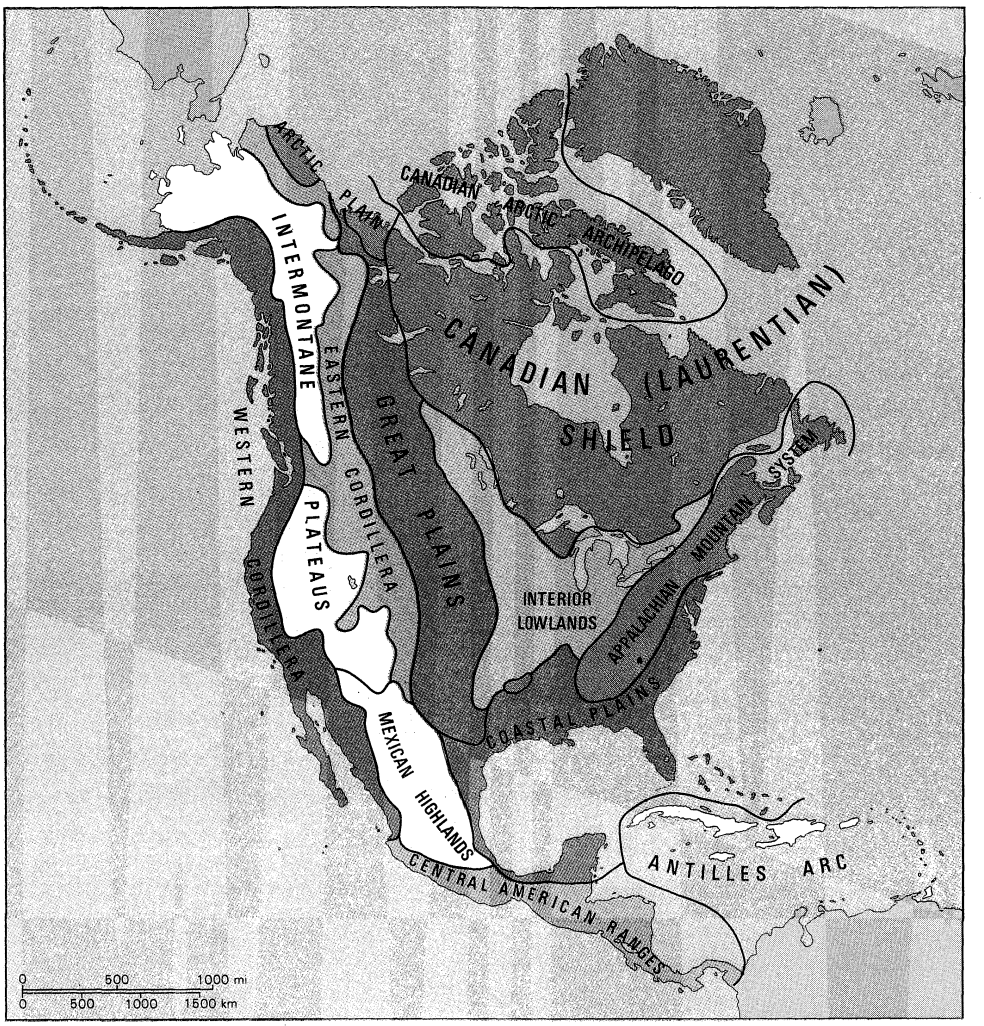
The intermontane plateaus. Lying between the eastern and western fold belts of the Cordilleras, and developed from several very old Precambrian and early Cambrian landmasses, the intermontane plateaus emerged as a broad median mass, or geanticline, that somehow resisted the downsinking, uplift, and folding of the mountain zones on either side. The Grand Canyon of the Colorado has exposed some of the oldest Precambrian rocks in North America; thus, the Colorado Plateau may well have been part of the initial structure of the continent, forming the base of the Y-shaped stable area that expanded in the north into the Canadian Shield. Although shallow seas covered it during the 250,000,000 years or so from Mississippian to Cretaceous times, this base was never depressed, and it divided the western geosynclines from the beginning. It was, however, subject to uplift at each of the mountain-building periods, and it was intruded by igneous rocks. To the north and south of it lie the intermontane basin-and-range provinces of the United States and Mexico, where uplift and faulting have produced a series of sharp-edged ridges between downthrown depressions. Farther north the Columbia Basin lava tablelands and the great igneous intrusions of the Caribou, Omenica, and Yukon plateaus dominate the structure.

By courtesy of NASA

Mountain ranges and basins of the arid continental interior of the southwestern United States. Photographed from Apollo 9 spacecraft.

THE EVOLUTION OF LOWLANDS AND COASTAL PLAINS

The Interior Lowlands. While the marginal mountains slowly emerged, the Interior Lowlands of the continent



Physiographic regions of North America.

had been forming, growing between them and the central shield as well as over the shield's platform-like buried prolongations. The lowlands expanded by the infilling of great basins that sank between broad upwarps, or arches, in the buried platform. Seas spread in from the Mexican, St. Lawrence, and Beaufort gulfs and worked over and redistributed the debris being washed down from the surrounding mountains. Central marine and peripheral deltaic deposits built up the Interior Lowlands in the form of all-but-flat strata, a great contrast to the contorted structures of both shield and marginal-mountain belts. A mantle of Paleozoic rocks is preserved in the river basins around the western and southern margins of the shield. The covering Mesozoic rocks have since been worn back to the higher levels of the Great Plains and Peace River plains, often protruding back of such long, low escarpments as the Missouri Coteau. Cap rock of Tertiary deposits, swept out from the Rockies, formed plateau-like sections, or the High Plains, above the main stretches of the Great Plains. Finally, Tertiary marine deposits invaded the Lower Mississippi and Gulf plains to the south and the Arctic lowlands in the northwest.

Role of the buried shield

Pressures from the marginal mountains warped the buried platforms beneath the central plains, occasionally forming low plateaus above the general level of the lowlands; *e.g.*, the Nashville Dome, part of the Interior Highlands reflecting the upwarp of the Cincinnati Anticline. These upwellings have been eroded for so long that the tops of the rises have been worn away to form basins with infacing scarps. In the Ozark Dome, erosion has worn back the overlying deposits to expose a crystalline core represented by the St. Francis Mountains.

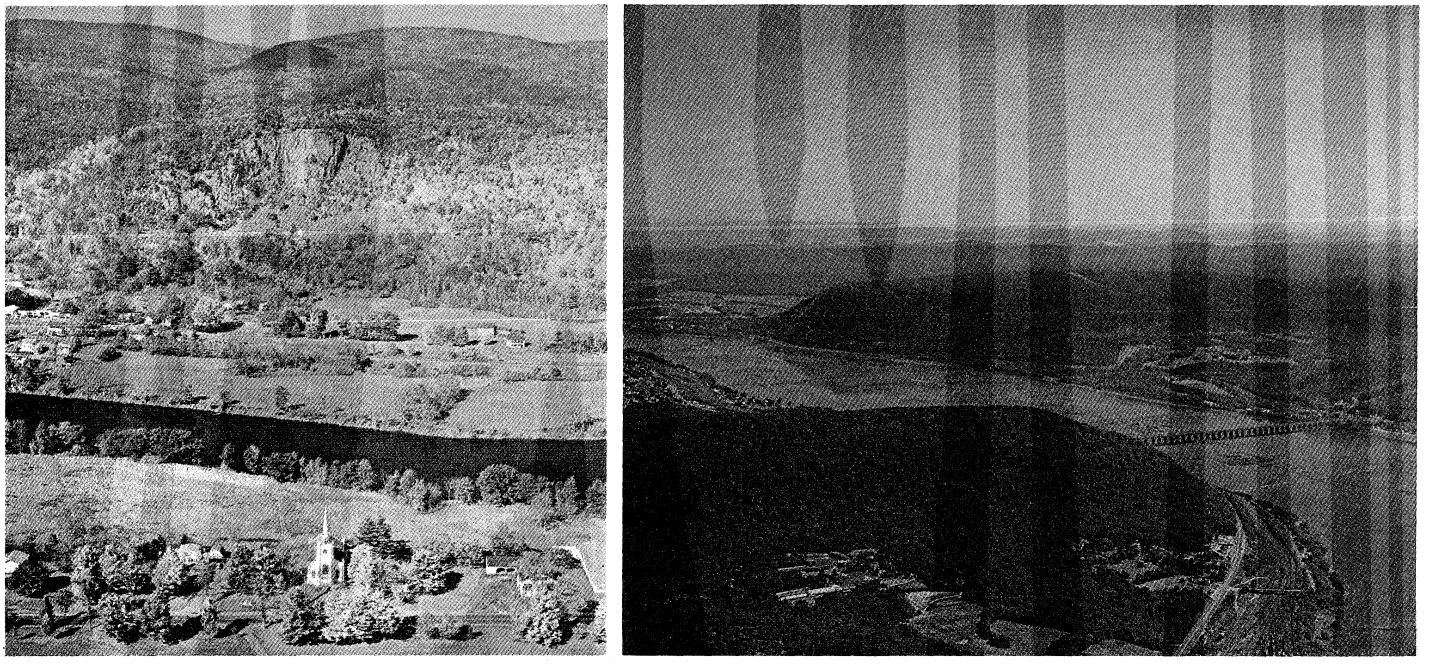
The coastal plains. Coastal plains are poorly developed in North America, mainly because of the extensive faulting that has let much of the land be drowned by the sea or—especially along the Pacific—because the marginal mountains step right down to the ocean. The whole of the west coast is virtually without a plain, except for such small and isolated deltas as those of the Fraser and the Sacramento. Folding and faulting along the northeast and northern coasts have practically eliminated lowland from New England to Ellesmere Island. Only in the southeast and the extreme northwest, along the gently declining shores of the Gulf of Mexico and of the southern Atlantic states or the slowly shelving levels around the Beaufort Sea have any extensive coast plains emerged. The Gulf and the Arctic plains are of interest in that they seem to be sinking into new geosynclines.

II. Physical geography

Although geology has been so important to North America that the 19th-century United States historian Frederick Jackson Turner once contended that the life of America flowed down the arteries of its geology, the continent has nevertheless been altered very considerably by climate and drainage and, to some extent, by soils and vegetation. The resultant physiographic regions dominate the contemporary geography of the continent.

THE CENTRAL SHIELD

Averaging 1,400 feet in height, with a very nubbly surface where old worn mountains and domes rise above ancient basins, the central shield has been left as a low plateau, tilted at its edges, and sinking down to Hudson Bay at its centre. The southern edge has the mountainous Algomans and Laurentians (more than 2,000 feet) and rises to above 4,000 feet in the great dome of the Adirondacks. The eastern edge is much higher, lifting itself near-



Appalachian landscapes.
(Left) Northern New England, looking across the Connecticut River Valley toward the low mountains of Vermont. (Right) Southern Ridge and Valley Province, showing the Susquehanna River near Harrisburg, Pennsylvania.
(Left) Laurence R. Lowry—Rapho Guillumette, (right) Grant Heilman—EB Inc.

ly 6,000 feet in the Torngats and 10,000 feet in Baffin Island; in Greenland, too, it tilts up appreciably in excess of 6,000 feet. The western rim is much lower, reaching only about 600 feet in parts. The old Snare and Nonacho ranges west of Hudson Bay lift the edge of the plateau to nearly 2,000 feet. Faulting broke up the northern rim into a series of prongs, extending into southeast Ellesmere Island and across Victoria Island, with sea-drowned channels and low sedimentary basins in between forming the Canadian Arctic Archipelago. The whole shield was under ice in the Pleistocene, and its high eastern rim still contains relics of the ice sheet. Ice-cut valleys in the higher areas, ice-plucked basins everywhere, and the ice-deposited ridges known as eskers and drumlins point to a major centre of ice dispersion over central Labrador, still noted today for its very heavy snowfall. Greenland was also a main glacial centre, while Keewatin in the west was an important secondary focus. After most of the ice had melted, portions of the shield rose, leaving traces of former beaches all along Greenland, Baffin, Labrador, and the Gulf of St. Lawrence, thus providing narrow but vital benches for human settlement. Ice-cut rock basins and ice-dammed streams have left countless lakes that make parts of the shield almost more water than land.

Centres of glaciation over the shield

THE MARGINAL MOUNTAINS

The Appalachians. Erosion also profoundly altered the marginal mountains. The Appalachians have been planed down to such an extent that their crest lines are smooth-topped for hundreds of miles. Several levels at which summits accord with each other in height indicate a series of uplifts followed by planations. In Canada, the highest level lies at about 4,000 feet, in the flattops of the Shickshocks; another exists at 2,000 feet on Mt. Carleton; and lower ones lie at roughly 1,100 and 600 feet in the Acadian ranges. In New England very resistant mountains like Mt. Washington and Mt. Monadnock rise above a broad mass of ridges at just above the 2,000-foot level, which in turn rise above the 1,100-foot-high New England Upland. Glaciation deepened and straightened the valleys, strewed their sides and parts of the coast with debris. Portions of sea-buried end moraines, which mark the limit of the tonguing glaciers, form offshore banks east of Newfoundland, Nova Scotia, and New England. The unglaciated Appalachians, south of the Susquehanna River, have striking accordances of long flat-topped summits at about 2,500 feet and broad terraces at 500 to 600 feet. The Ridge and Valley section's pattern of drainage consists of short, deep gaps across the ridges and long parallel stretches in between. East of the Blue Ridge extends the Piedmont Upland, terminating abruptly in the Fall Line, where its rivers plunge down to the Atlantic Coastal Plain over rapids or falls. The Hudson–Mohawk gap represents a major break between the northern and the southern Appalachians and affords a natural entry to the interior of the continent.

The Cordilleras. Taking up about a third of North America, the Cordilleras completely dominate Alaska and Central America and swell out widely in the United States Rockies.

The Canadian mountains. In Canada the Cordilleras consist of six well-marked zones: the 10,000- to 12,000-foot-high Rocky Mountains continuing north into the Brooks Range of Alaska; the Rocky Mountain Trench, a profound fault feature carrying the headwaters of the Columbia, Fraser, Peace, and Yukon rivers; the interior uplands and old fold mountains from the Selkirks and Okanagans in the south beyond the Cassiar Massif, to the Yukon plateau in the north, mostly lying at about 2,400 feet but with ridges over 8,000 feet; the Coast Mountains extending north into the Alaska Range, and including lofty volcanoes in the north; the Inner Passage from Puget Sound to Alaska, possibly a downfaulted zone flooded by the sea; and a structurally complex outer Island Arc, running from Vancouver Island to the Aleutians. The magnificent scenery of the northern Rockies, including U-shaped valleys often extending westward into sea-drowned fjords, has been heavily glaciated, and some areas still nurse sizable glaciers.

The six zones of the Canadian Cordilleras

Cordilleras of the United States. In the United States the Rockies, typified by flat or gently folded rocks, sweep south from Canada into northern Montana as the Lewis Range. They then change to a group of domes or long anticlines with "parks" or broad basins between them. This park and dome area is characterized by the peeling back of younger rocks from the cores of much older, primary rocks at the heart of the anticlines. The southern Rockies have striking volcanic peaks. Westward is the vast region of the intermontane plateaus, extending from the Selkirks of Canada across the immense lava tablelands of the central Columbia–Snake River Basin. South lies the Basin and Range Province, where the broad central plateau has been split by a great number of fault

ranges, the slopes of which plunge under basins partly filled with debris worn from the ridges—a pattern often repeated in the central Mexican Plateau. The Colorado Plateau is a massive feature with a series of relatively flat bedded ridges, made steplike by faulting action and intruded by domes of igneous rocks. Its slow elevation was matched by the steady downcutting of the Colorado River, producing, in the Grand Canyon, one of the greatest gorges in the world. Westward rise the beautiful high Sierras, reaching to nearly 15,000 feet, intensively folded and faulted, and continued north in the Cascades, marked by some of America's most beautiful extinct volcanoes. Seaward of this lofty mountain zone is a line of depressions marked by Puget Sound, the Great Valley of California, and the Gulf of California. These are separated by knots of volcanoes, as in the Klamath Mountains, and enclosed by the Pacific coast ranges, including the volcanic peaks of the Olympus group. This whole area has been profoundly faulted. Along some of the faults, notably the San Andreas, earthquake shocks occur from time to time, occasionally with devastating results.

Mexican and Central American mountains. In Mexico the folded Sierras to west and east of the central plateau terminate in the grandeur of a mass of high volcanoes, of 15,000 to 17,000 feet, south of the fertile lake-filled basins of Guadalajara and Mexico City.

The structural break at the Sierra Madre del Sur

The Balsas Basin then makes a distinct break. To the south, the Sierra Madre del Sur and the mountains of Guatemala and Honduras exhibit a west–east trend. This structural region includes a sweep of fold mountains of from 4,000 to 6,000 feet, with Caribbean extensions in Jamaica, southern Cuba, the island of Hispaniola, and Puerto Rico. These mountains swing south through the West Indies, a chain of volcanic islands fringed with coral reefs or limestone plateaus. Another arc, of two lines of fold mountains on either side of the Nicaraguan trench, dominates Central America and links it with the folds of western Colombia, in South America.

THE LOWLANDS

Glaciated areas. The lowlands of North America show marked differences between glaciated and unglaciated areas. Long irregular lines of coarse morainic deposits mark the regions where the Cordilleran ice sheet, moving down from the Rockies, met the continental sheet from Keewatin. The huge, tonguing loops of end moraines also occur between the Red and Mississippi rivers and along the south side of the Great Lakes—some running for hundreds of miles and standing 300 or 400 feet above the plain. Four major glacial advances covered the lowlands with debris: the Nebraskan crossed the Missouri River; the Kansan came to the Lower Missouri and the Ohio rivers; the Illinoian overlay these older areas to southern Illinois; and the Wisconsin heaped its moraines over the southern Canadian Prairies and the Great Lakes–Ohio–Mohawk–Hudson plains (see also below *Drainage patterns*).

Unglaciated areas. In the ice-free areas, lowlands are river molded. Streams debouching from the Rockies have spread sands, occasionally whipped up into sand hills, well beyond their banks; those funnelling into the Mississippi have created a vast alluvial plain running out into the Mississippi Delta. The Coastal Plains are marked, in addition, by lines of sand hills, the relics of stranded beaches that eroded as the plains were lifted slowly out of the seas in postglacial times.

THE ROLE OF CLIMATE

Climate has made a great deal of difference to relief—particularly in the extremes of cold or hot and dry or humid landscapes; climate has also, in turn, been affected by relief. The enormous northern width of the continent has meant a great extension of Arctic and cool-temperate climates, while the tapering south has greatly reduced the area under the tropics. The Cordilleras have very wet, windward, Pacific-facing slopes and dry, leeward, interior-facing slopes. Water systems such as the Mississippi–Ohio and the Great Lakes–St. Lawrence tend to funnel the sweep of rainstorms across the central parts of the continent, and the Gulfs of Georgia, Mexico, and the St. Lawrence also concentrate wind streams.

Temperature. While the greater part of North America falls within the temperate zone, a fact that has made it attractive to European settlers in post-Columbian times, large cold areas lie in the north and extend as far south as the Ozarks in winter. The continent's northerly position means that Greenland, the Canadian Shield, the Mackenzie Lowlands, and the northern part of the Cordilleras have an unusually long and cold winter. Much of this land has permanently frozen subsoil and is under snow and ice most of the year. The frequently frozen seas interlacing the Canadian Archipelago, together with innumerable northern lakes, produce an enormous chilling effect on the air above, and the temperatures for this huge regions are 6° to 8° F (3° to 4° C) cooler than the average for their latitude. The North Pacific, warmed by an extension of the Kuroshio current, has a positive anomaly of 8° to 10° F (4° to 6° C) warmer than the average for its latitude. Related trends over the northern part of the North Atlantic affect Iceland and Europe rather than North America but still raise the temperatures off the northeast coasts by perhaps 2° F (1° C). The climate thus shows marked contrasts between the maritime and continental areas. A notable warm loop of temperatures extends up the west coast from Vancouver Island to Alaska, while a great cold loop extends down the Mackenzie plains and the Canadian Shield over the heart of the continent. The chilling effects of the great Greenland ice cap drag cold continental conditions over the northeast coast at least as far as Newfoundland. The average January temperatures of Annette Island in the Alaskan panhandle, 29.5° F (−1.4° C), of Fort Smith, Northwest Territories, −13.8° F (−25.5° C), and of Nain, Labrador, 1.3° F (−17.1° C) show the difference between coastal and continental conditions and also between the west and the east coasts—differences negligible in the tropical parts of North America.

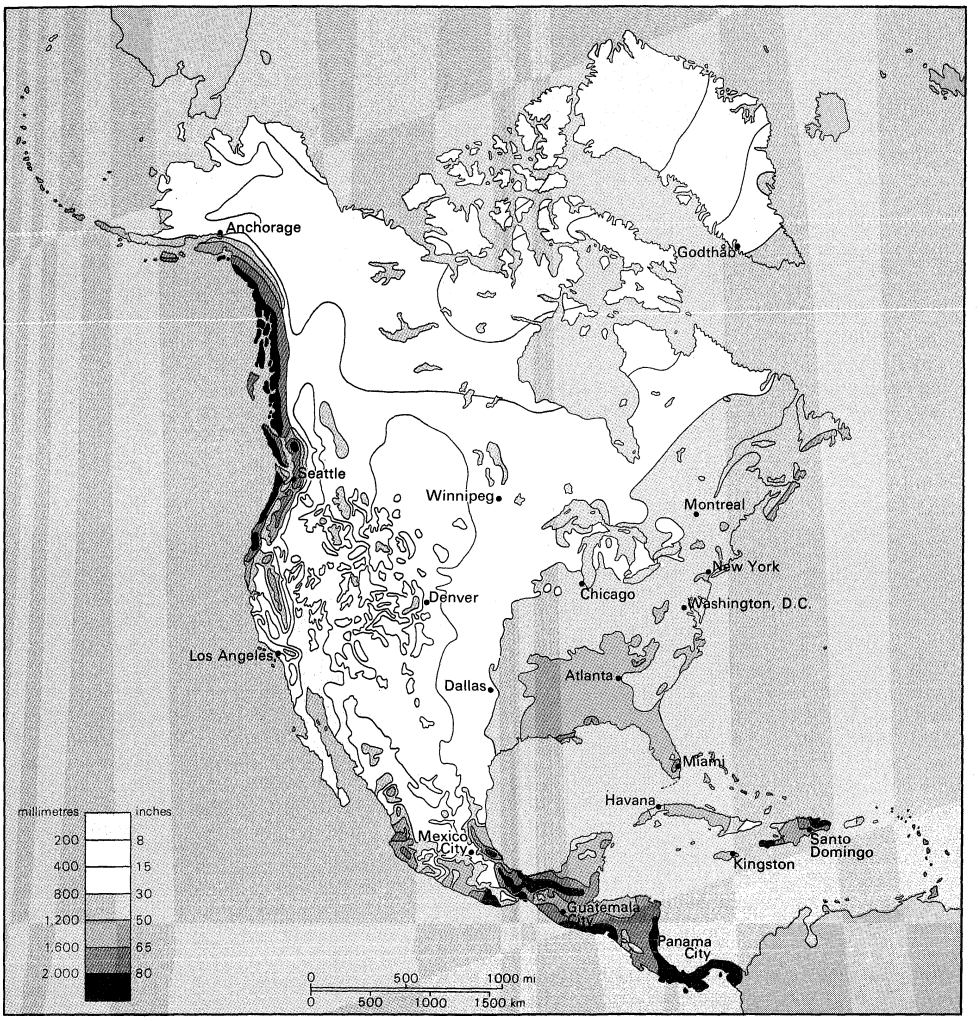
Chilling effect of frozen subsoils and seas

Precipitation. Most of the continent is humid and provides a good water supply for settlement and development. From mid-California north, the Pacific Coast is bathed with rain- or snow-laden Westerlies, giving from 40 to 200 inches (1,000 to 5,000 millimetres) of precipitation a year. Westerlies again reassert themselves east of the Rockies and, especially east of the Missouri River and Red River of the North, bring about moderately wet (20 inches) to wet (45 inches) conditions in the central and eastern regions. There are two main areas of drought: one in the extreme north and northeast, under the influence of very cold and relatively dry winds, with a thin dusting of winter snow and a meagre fall of summer rain in the Canadian Archipelago and Greenland; the other in the southwest, where the mid-latitude high-pressure system leads to dry offshore winds from mid-California to southern Mexico. Since these winds blow from the interior out to sea, they carry very little moisture, with precipitation in that area usually less than ten inches a year.

Two main areas of drought

Air masses. The continent's air masses reflect very different conditions of temperature and precipitation; they include northern and southern components, subdivided into continental and maritime types. In the north are found: the Arctic air mass, over Greenland and the Canadian Arctic Archipelago; the polar continental, over northern central Canada; the polar Pacific, over Alaska and the northern Pacific shores; and the polar Atlantic, off the Atlantic provinces of Canada and New England. The tropical continental air mass, over the intermontane basins of the Cordilleras from Utah south; the tropical Gulf, centred in the Gulf of Mexico and the Caribbean; and the tropical Atlantic off the southeastern states characterize the south-southeast.

The polar continental, the tropical gulf, and the polar Pacific are the most significant air masses. The polar continental reflects the spread of a negative temperature anomaly over much of the continent. It is a cool to cold mass of stable air forming an immense dome of high pressure above the Canadian Shield, with winds blowing outward to sweep over Labrador and New England or south-



Average annual precipitation for North America.

ward across the Great Lakes and the Great Plains. At its maximum, it extends from the Canadian Arctic Archipelago to the Ozarks. In winter it joins with the Arctic air mass over Greenland to make a formidable body of cold heavy air that carries subzero weather as far as the Ohio Valley and may overflow the Appalachians and seep through the Rockies. Exceptionally, it can carry killing frosts into the Great Valley of California, the coast of Texas, and the neck of Florida.

In the spring it shrinks north before the swift advance of the tropical Gulf air, which is drawn northward by low pressures developed in the Mississippi Basin as the heart of the continent heats up. The air mass is warm, wet, and unstable; at its height in late July, it extends two enormous loops of warm air, one northwestward up the Mississippi and down the Mackenzie and the other northeastward up the Ohio and down the St. Lawrence. The July average of 60° F (16° C) is then carried north of Edmonton, in the west, and to Quebec, in the east. The storm-generating polar Pacific air mass is very active from northern California to Alaska, especially in the winter, when its mild, wet air reflects the North Pacific temperature anomalies.

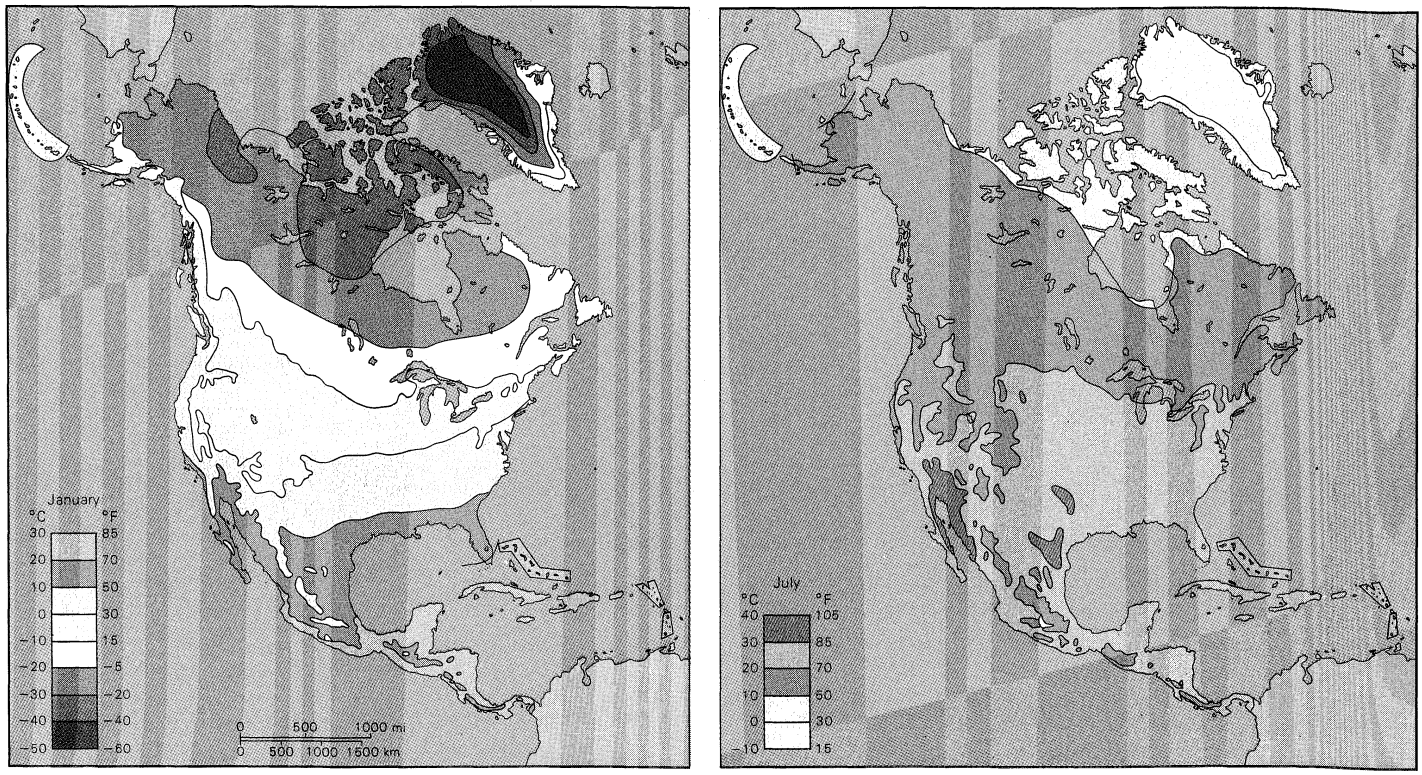
Storm tracks. Where cyclones develop persistently along the advancing air-mass edges, strong storm tracks occur. Pacific storm tracks thread the Gulf of Georgia, Puget Sound, and the Inner Passage to Alaska. In summer they shift north of Prince Rupert; in the depth of winter they migrate south to San Francisco. Moving up the great Columbia and Fraser river valleys, they may pass through the Rocky Mountain passes. An upper stream of Pacific air overtops the mountain barrier and, on descent, starts off lee storms that then traverse the continent. These draw in air from the polar continental air mass on their advancing cold sectors and from tropical Gulf air in their warm sectors. As the polar continental air mass begins to expand in September, a Mackenzie–James Bay line of storms develops, migrating progressively southward to reach a Texas–Ohio track in January. As the tropical Gulf air mass expands north, the successive tracks become activated again until, in August, the Gulf air brings a swirl of storms to the Mackenzie. Most of these storm tracks begin in the western plains, converge on the Great Lakes–Ohio area, and then bunch together in the climatically stimulating St. Lawrence–Hudson–Mohawk zone. The Atlantic Coastal Plain becomes a storm track in winter as tropical maritime air contests the advance of the continental air from the north.

Pacific-coast storm tracks

Climatic regions. Differing continental climatic regions reflect the considerable amount of Arctic land, the great spread of temperate conditions, and the small but significant tropical area; dry climates also stand out in strong contrast to the prevailingly humid ones.

The Arctic zone. Including the northern parts of the Canadian Shield and Alaska, the Canadian Arctic Archipelago, and Greenland, the Arctic zone is dominated by Arctic and polar continental air masses and is perennially cold or cool. Subzero weather lasts for five to seven months, and subfreezing weather lasts for eight to ten months. It is only June to September that temperatures rise above 32° F (0° C). The frost-free season is not more than 60 days. Precipitation is low, with two to four inches of summer rain, plus 30 to 60 inches of winter snow. Most of the Arctic is a cold desert.

The cool-temperate zone. The cool-temperate zone extends from Newfoundland to Alaska and from Hudson Bay to the Ohio. It is dominated by the polar continental



Average temperatures for January and July for North America.

Zone of long, severe winters

air mass. Winters are long and severe. After an "Indian summer" that continues into October, temperatures fall quickly and do not rise substantially until April or early May. In January and February they drop to below 32° F (0° C) on the Ohio and below 0° F (−18° C) north of the Great Lakes, with minima as low as −20° to −80° F (−29° to −62° C). Winter killing of crops and spring and autumn frosts are a hazard in the Canadian parts of the region, where the frost-free season is from 90 to 165 days. A swift transition occurs with spring; tropical Gulf air raises monthly mean temperatures to over 50° F (10° C) in June and from 60° to 78° F (16° to 26° C) in July. Precipitation is moderate, with 15 to 35 inches; as evaporation is low, however, most of it is effective for growth. The maximum precipitation occurs in summer to fall, when the James Bay, Alberta, and Wyoming storm tracks are activated.

Warm-temperate zone. On the southeast coasts of the United States, the warm-temperate zone extends to the Mississippi and over the Gulf plain; the zone is strongly influenced by the tropical Gulf air mass. The long frost-free season lasts more than 200 days. The tropical airs spread north in February and dominate the region until November, when polar continental air makes itself felt. Winters are mild, with January means of from 40° to 54° F (4° to 12° C). July averages are tropical, being as high as 81° F (27° C). This warmth and the long growing season allow for subtropical crops. Rainfall is ample, ranging from 40 to 60 inches, and benefits from the presence of the Colorado and Texas low-pressure systems and from the strong summer movement of tropical maritime air. By then the landmass is intensely heated, and this, combined with the air movement, produces frequent thunderstorms, especially in early summer. Hurricanes are an annual hazard along the Gulf and up the Lower Mississippi.

In the United States Southwest a different regime pertains, with a Mediterranean type of climate; summers are dry, since the tropical continental air is dominant. July means of 70° to 80° F (21° to 27° C) are found, with bright, sunny skies. Winters are mild (45° to 50° F [7° to 10° C]) and wet, with polar Pacific airs swinging south and bringing heavy rain. Frost is rare but may occur when polar continental air thrusts through to the coast. Los Angeles has a record minimum of only 23° F (−5° C). The annual rainfall of from 15 to 30 inches, along with very high evaporation, often is insufficient unless supplemented by irrigation. Drought is a frequent hazard.

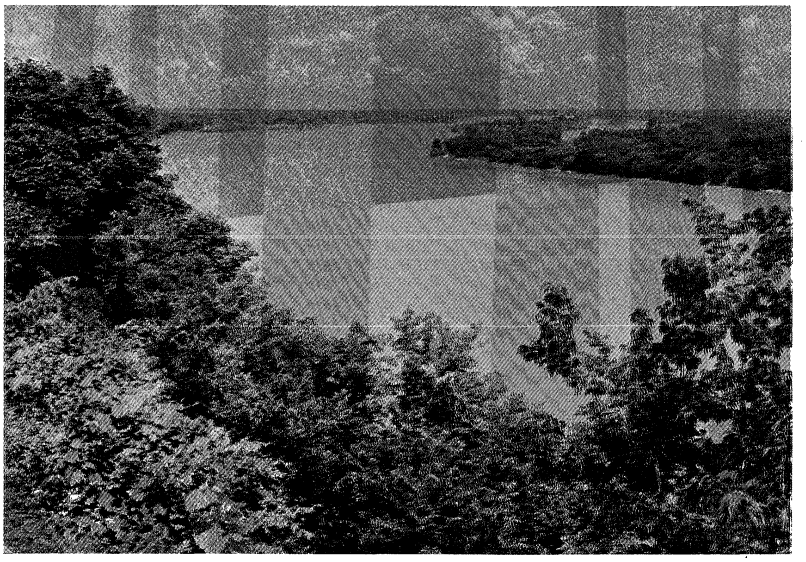
The tropical humid climate. Central America, with its tropical humid climate, knows no winter; even the coldest month averages above 64° F (18° C). With summers of 80° to 82° F (27° to 28° C), the mean annual temperature range is low—markedly different from most of North America. Rain is ample and regular, with 45 to 80 inches where the trade winds blow onshore. Lee valleys are, however, often quite dry. Summer hurricanes frequently recur, causing much damage.

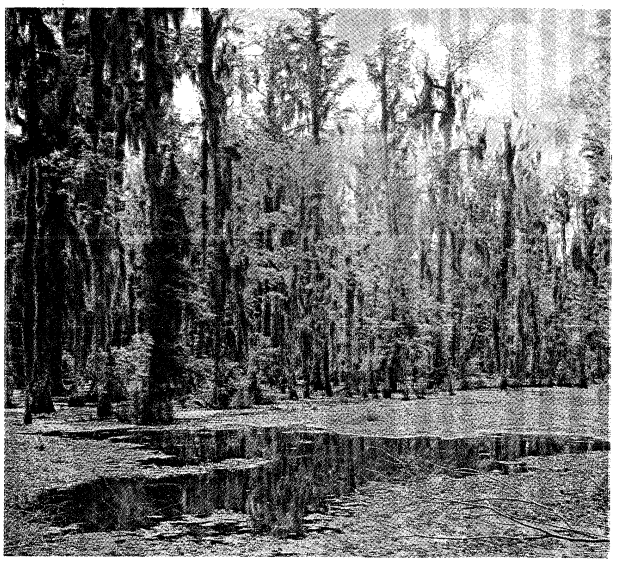
Dry climates. About a third of North America, including Arctic areas, has a dry climate. Chief dry areas lie in the Southwest, where a combination of the mid-latitude high-pressure belt, the tropical continental air mass, and the rain-shadow effects behind the high Western Sierras has led to lack of rainfall. Winds blow from the continent outward, discounting the effect of the Pacific. As the winds move down from high interior plateaus, they become drier as they descend. The intermontane region of the United States and Mexico, from the Columbia Basin to Guadalajara, and the Pacific Coast from San Diego south to Mazatlán are therefore arid, with less than ten inches of rain a year. Some years have no rain. The Great Plains, from the South Saskatchewan River to Mexico, are semi-arid, with from ten to 15 inches of rainfall; the high midcontinental airflow known as the jet stream is depressed over them, strengthening downmoving dry wind from across the Rockies and tending to fend off cyclones from tropical Gulf or polar continental air masses.

Factors in the Southwest's aridity

DRAINAGE PATTERNS

Drainage conditions and water supply are affected markedly by climate, though they also reflect relief. North America has one of the longest rivers in the world (the Mississippi) and also a drainage system with one of the greatest water capacities (the St. Lawrence–Great Lakes system). It is a continent of immense rivers—very largely because of their vast drainage area in the long and broad plains between the central shield and the marginal





Mississippi River Basin.
(Left) The Mississippi River at Hannibal, Missouri. (Right) Coastal swamplands of southern Louisiana, United States, showing moss-draped cypress trees.
(Left) Fred Bond—Publix, (right) Josef Muench

mountains. Rivers rising in the shield, the Appalachians, or the Cordilleras that flow into the Interior Lowlands have a long way to go to reach the sea. The Great Lakes–St. Lawrence on the east and the Mackenzie on the west drain the outer edges of the shield. The Nelson takes advantage of the low saddle in the shield to carry the gathered waters of the Saskatchewan and the Red out by way of Lake Winnipeg to Hudson Bay, which also has fairly long rivers draining into it from the uptilted edges of the shield. The vast Missouri–Mississippi–Ohio system draws from the Cordilleras, the shield, and the Appalachians to unite the Central and Gulf Lowlands, in the heartland of the continent. From the Rocky Mountains, long rivers like the Colorado, Columbia, Fraser, and Yukon continued to flow west to the Pacific even across the Cascade-Nevadan and the Coast ranges as these were lifted up. Fed from eternal snows, they are particularly valuable in the arid Southwest.

There is a marked asymmetry to the continent's drainage: the chief continental divide, along the Rockies, is well to the west, thus shedding the longest rivers to the east. The longest tributaries—the Peace into the Mackenzie, the Saskatchewan into the Nelson, and the Missouri into the Mississippi—coming from the west, tend to displace the mainstreams to the east. The chief gulfs—Hudson Bay, the Gulf of St. Lawrence, and the Gulf of Mexico—taking the discharge of many of the rivers, are also on the east. All these factors helped the European settlers to move from their Atlantic bases into the heart of North America.

Lakes. Lakes abound in North America. Most of them are products of glaciation, which had a vast effect on the continental drainage pattern, notably by the widening of passes through the northern Appalachians and the Cordilleras and the creation of big lakes in ice-deepened basins. The Great Lakes proper have a fascinating history, as Lakes Superior and Huron were vast synclinal depressions even in Precambrian times. In the place of the present Lower Great Lakes, a scarp-and-vale topography existed, with the high front of the Niagara limestone scarp separating vales of shale to the west and east. The glaciers picked out the vales and synclines and deepened them into ice-cut basins, where water gathered; as the ice melted away, the Great Lakes formed. While the ice front of the glaciers still blocked the St. Lawrence outlet, the early lakes drained southward into the Mississippi–Ohio, the Susquehanna, and the Mohawk–Hudson. When the ice retreated from the Gulf of St. Lawrence, the lakes sought the lowest outlet through the St. Lawrence River, lowering the level of the Great Lakes and leaving beaches around them that stand out as raised beaches, or strandlines. The land, depressed under the enormous weight of the ice, has been rising since, lifting the old beaches above the present diminished bodies of water. Similar strand lines follow the Gulf of St. Lawrence, once under glacial Lake Champlain; Lake Winnipeg, once part of the immense glacial Lake Agassiz; and Athabasca, Great Slave, and Great Bear lakes, which are the relics of once deeper and larger glacial lakes. The western lakes were formed by ice blocking the free drainage of water to Hudson Bay or the Beaufort Sea. Much farther south, in the intermontane basins, a pluvial period of climate, matching the ice age in the north, led to the enormous lakes Lahontan and Bonneville. The Great Salt Lake is a relic of Lake Bonneville, the ancient beaches of which are stranded 1,000 feet above the present ones. Similarly, present-day Lake Mexico represents only a small part of the large body of water that accumulated in the Mexican Basin and whose level fluctuated during several pluvial periods. The contribution of all these—and many more—lakes to the drainage of North America has been outstanding. Much of the Canadian Shield is so ridden with lakes as to form an amphibious landscape.

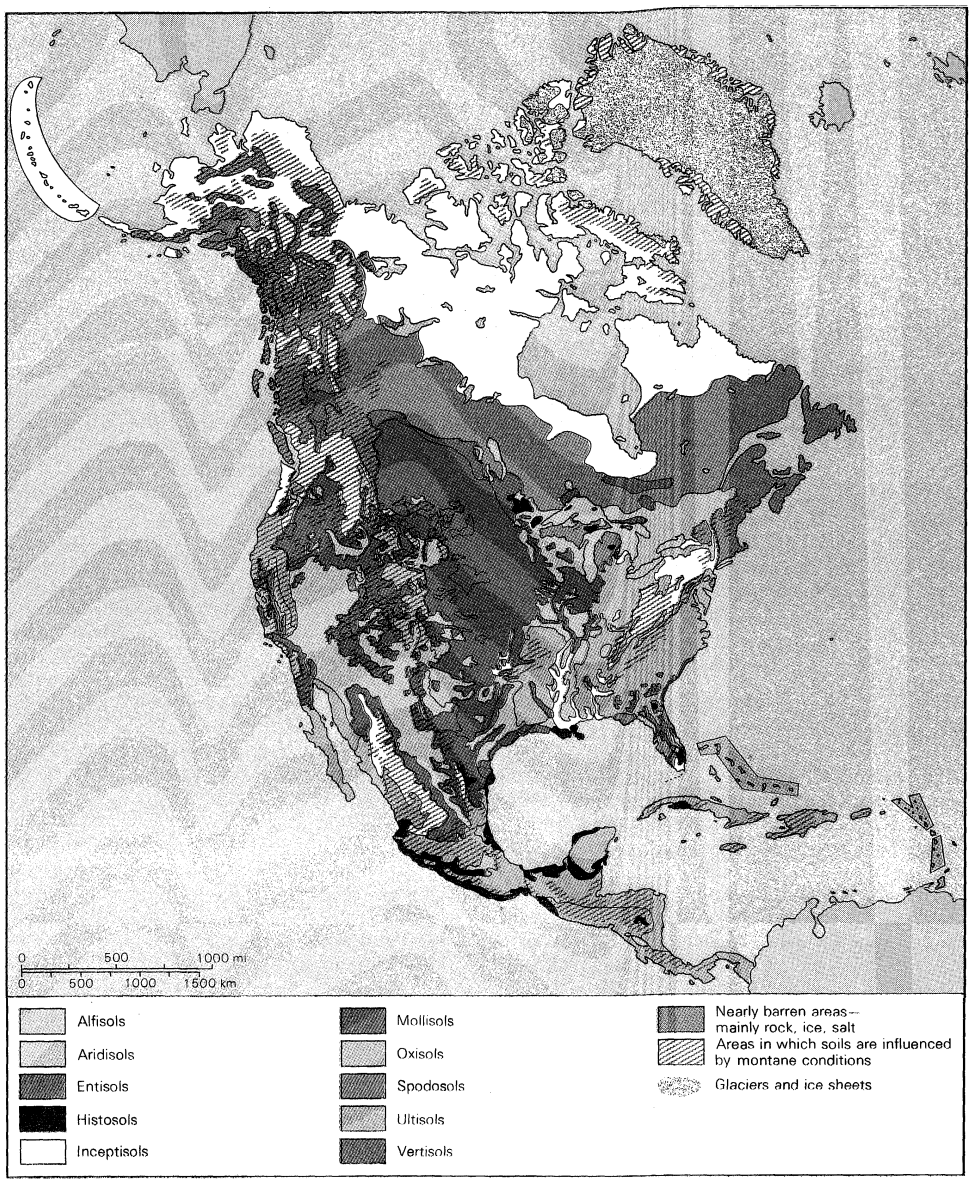
The Great Lakes

River regimes. The river regimes of North America exhibit great variety. Northward-flowing rivers like the Yukon, Mackenzie, Red, Nelson, and the rivers of Labrador freeze up in winter. Because their upper parts then melt before the lower sections are free of ice, they frequently flood, especially if the thaw is late enough to coincide with early summer rains. The Mississippi system is also swollen by spring meltwaters as the winter snows give way in April; flooding can then become a major hazard. River water is kept high by the rains that tropical Gulf air and local convection storms bring until midsummer. A marked falloff then occurs, giving way to full flow in late autumn and winter as polar continental air reactivates midcontinental storm tracks.

The St. Lawrence runs high in spring and early summer; the winter precipitation falls on a frozen surface and serves mainly to heighten the spring thaw. Most other eastern rivers have a double maximum, occuring in early summer and late winter. In the United States Southwest the winter is the main floodtime; rivers dwindle appreciably in the summer. The northern Pacific region, by contrast, has rain at all seasons, though with a winter maximum. In the southern tropical regions, rivers have a much more regular regime, running full all year, except in the dry rain-shadow areas leeward of the mountains.

III. The living environment

The physical factors in the North American environment —geology, relief, climate, and drainage—have a great



Soils of North America.

effect upon the soils, vegetation, and wildlife of the continent. Nevertheless, biotic factors, which govern the multitudinous life-forms of the continent, are also important and help to account for the biogeography of the continent.

NORTH AMERICAN SOILS

Soils reflect the rocks from which they were derived, the way in which they were laid down, conditions of temperature and moisture, and the plants and animals living in or on them. Climate has been used as the main basis for categorizing soils, with a division first into humid and arid groups, and then into subgroups according to the way in which temperature and moisture acted together to produce different horizons, or layers, in the soil. More recently, attention has been focussed on the horizons themselves and on their unique characteristics. In the 1960s the United States Soil Survey used this horizon-based classification to propose a new comprehensive system, which nevertheless still reflects the strongly zonal distribution of soils, following the dominance of climate in the continent.

Horizons as a basis for classifying soils

Humid soils. The humid soils have the widest range in North America and include a number of subdivisions.

Inceptisols. The west maritime temperate-equable climate—swept by frequent polar Pacific storms and with a growing season of up to 300 days and ample rainfall of 40 to 200 inches—has favoured the creation of a deep, acidic, brown-coloured soil, with an upper horizon rich in humus formed by partly decayed organic matter from the region's dense forests. These soils, known as inceptisols, are the fertile soils of the Pacific Northwest and of the British Columbia and Alaska coasts.

Spodosols. The cool-temperate climatic zone is characterized by spodosols, soils with a moderate humus layer at the surface succeeded beneath by a gray, leached, or washed-out, horizon. Rather infertile, acidic, grayish soils, known as podzols, result from this leaching process. They extend under the boreal forests from Alaska to Newfoundland.

Alfisols. Found in the warm-summer subregion of the cool-temperate zone, where mixed forests of conifers and deciduous trees cover the Great Lakes–St. Lawrence to the Ohio area, alfisols are characterized by a deep layer of humus, succeeded by a shallow leached layer, which, in turn, goes down to a wide horizon of both plant and mineral nutrients including oxides of iron and aluminum. These are the moderately fertile gray-brown soils so typical of the northeastern United States.

Ultisols. Farther south, ultisols occur, roughly extending from the lower Ohio and Chesapeake Bay southward to the Gulf Plains. They correspond to the area of the southeastern mesothermal climate, which has more than 200 days free of frost and a rainfall of up to 60 inches a

year. Here a rich deciduous forest supplies a deep layer of humus, some of which is leached down and accumulates with red oxides of iron or yellow hydroxides of aluminum to produce podzolized red-yellow earths of considerable fertility.

Oxisols. The tropical climates of southern coastal Mexico and of Central America, with constantly high temperatures of 65° to 82° F (18° to 28° C) and perennial rainfall of from 80 to 120 inches, have caused very active weathering of rock. The resulting soils, called oxisols, have developed; hydroxide and sesquioxide compounds have produced a thick layer of red laterite, the residual product of rock decay, sometimes mottled with yellow, beneath the surface horizon of rotted humus. A deep, strongly acidic, lateritic soil has developed, which, though fertile when protected from erosion, can be very infertile if the bricklike laterite is exposed.

Semi-arid and arid soils. Semi-arid and arid soils cover an extensive area of North America, including the prairies, deserts, and tundras.

Mollisols. Marking the transition between humid and arid soils, mollisols are found in the open parkland or the tall-grass prairie of the outer Great Plains or in the humid prairies of the western Central Lowlands. They have a moderately deep surface horizon, which is black or chocolate brown in colour from the humus of the closely matted roots set in the dense sod under the thick-growing grasses. With a short rainy period from April to mid-July followed by great evaporation in a dry, sunny summer, whatever leaching occurs is short-lived and not pronounced. The leached layer is very shallow and passes down to a horizon in which the upward movement of water to offset the high evaporation at the surface has brought up basic salts, especially lime in solution. The lime neutralizes the acidity of the surface humus. A very fertile neutral soil—called chernozem, or black earth, by the Russians—has thus developed, seen at its best in the Dakotas and the fertile belt of the mid-Canadian prairies.

Development of black earth soils

Aridisols. Characterizing the dry climates of the United States intermontane basins, of most of the Mexican Plateau, and of the Southwest coast, aridisols are found where vegetation is sparse and where, accordingly, little humus has formed at the surface. Leaching has virtually ceased, and very strong evaporation has led to the upward movement of basic salts through capillary attraction, often working up to ground level in a skin of white crystals. The soils are too rich in lime and potassium to be fertile unless extensively irrigated. They are known as brown desert soils in the more temperate north and desert yellow soils in the subtropical south.

Permafrost soils. In the Arctic tundras, permafrost soils also accumulate very little humus and are strongly leached by the melt of winter snows. The water cannot pass down more than a foot or so before it strikes a permanently frozen horizon, and thus a thin, very acidic, waterlogged, infertile soil has developed. It is of practically no use to man.

VEGETATION AND WILDLIFE

North American vegetation and wildlife are obviously closely allied to soil, as their habitats, too, reflect the powerful influences of climate. Forests dominate the humid regions and once covered about two-thirds of North America; grassland, scrub, or lichen typify the dry third of the continent.

The forests and their inhabitants. *The Pacific coniferous forest.* Offering one of the great spectacles of the continent, the Pacific coniferous forest is made up of immense redwoods and firs forming vast cathedral-like groves, where the tall trunks rise hundreds of feet like great pillars to support a fane of evergreen branches overhead. A long growing season and heavy, constant supply of moisture together have fostered the densest and tallest of North American forests, with redwoods and western cedar along the north coast of California giving way to Douglas fir and western hemlock from Oregon to British Columbia, and Sitka spruce in Alaska. In the south redstemmed arbutuses lend a Mediterranean touch; giant-leaved maple, oak, and ash are common in the middle sector; birch and aspen are subdominants in the north. This coastal forest is still one of the continent's chief sources of construction timber. It is also a major source of pulp and paper and is still a home for a significant number of red deer and mountain elk and also of black bear, lynx, and beaver. Fish-eating hawks and eagles abound. Up the rivers, North America's greatest runs of salmon are seen here, as the fish swarm upstream to spawn in mountain lakes; while off the Queen Charlotte islands lies one of the continent's chief halibut fisheries. The meeting of the Kuroshio, a warm current from across the Pacific with the cool water along the western offshore deeps provides ideal conditions for fish life in great numbers. The faulted and glaciated coasts with great fjords and the perennial rivers attract fish inland and thus make fishing more easy and profitable.

Fish-breeding conditions in the offshore deeps

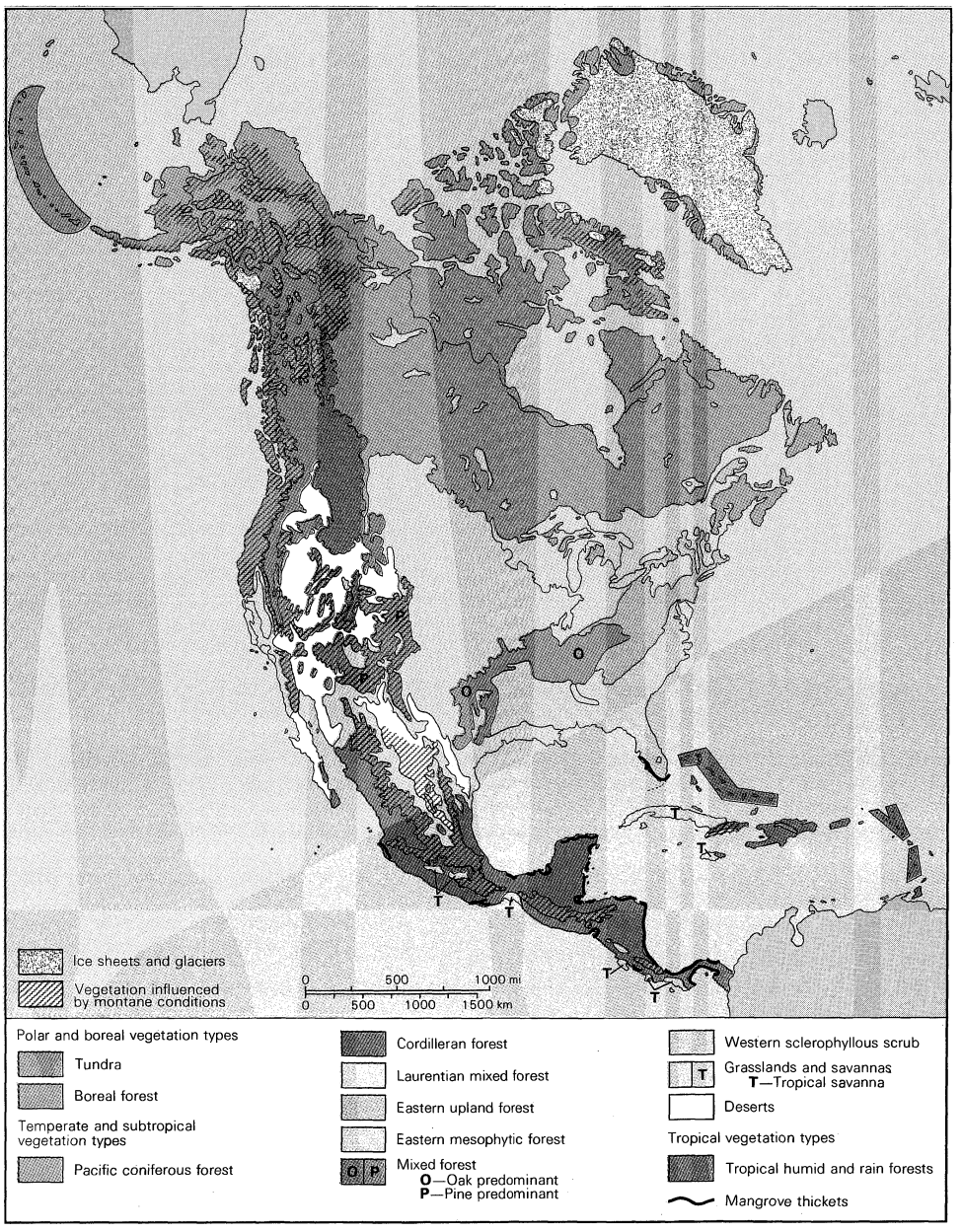
The boreal forest. One of the greatest sweeps of forest in the world, the boreal forest extends in a vast and virtually unbroken sheet of green eastward from the Aleutians through Alaska and northern Canada to the island of Newfoundland. Its conifers are shorter than those of the Pacific Coast but grow in denser stands. The boreal forest is essentially the domain of the spruce, with, however, the contorted pine becoming significant in the west, and the jack pine and tamarak in the east. From Alaska through the Mackenzie plains to Keewatin, the white spruce dominates, while, through eastern Canada and upland New England, the black spruce is common. The whole region is the Western Hemisphere's greatest source of pulpwood.

Great herds of caribou shelter in the northern fringes of this forest in the winter. They are preyed upon by packs of timber wolves. Farther south, deer, elk, and moose are still common, though thinned out appreciably by man. Both the black and the brown bear are frequently seen, especially in berry patches. Many fur-bearing animals, including the marten, squirrel, mink, and beaver, occur; muskrat abound in the marshes. In the spring, pickerel run up the rivers to spawn, and lake trout and whiting like the cool, deep waters of the innumerable northern lakes. Whitefish are caught in great numbers in Great Slave Lake but, because of man-made pollution, are much less prevalent than they used to be in the Great Lakes. Cod and haddock are found in vast numbers on the banks off Newfoundland southward to New England, where the cold Labrador Current mixes with part of the warm Gulf Stream, thus stimulating conditions for fish life.

The Cordilleran forest. The Cordilleran forest lies between the Pacific coniferous forest and the interior boreal forest. On the west, it is made up of cedar and Doulas fir, with Sitka spruce and Englemann spruce at the higher levels; while, in the east, it has more pine and spruce, with lodgepole pine and white spruce making close straight-limbed stands. On the intermontane plateaus and ridges, western hemlock and yellow or sugar pine form groves with parkland between. Altitude and aspect dominate tree distribution, with tall and dense fir woods occurring on the wetter faces at lower levels, spruce clothing the higher slopes, and pine abundant mainly on the drier exposures. Animal life is still rich: elk and deer in the fir–spruce forests, antelope in the open parkland, goat high up on the Alpine pastures. Preying on these are the mountain lion and the occasional pack of wolves. The grizzly bear keeps to higher and less accessible haunts, but the black bear is common in the lower forests. Mountain trout are abundant, and, outside the Far North, this region is the continent's chief game preserve.

Coniferous Cordilleran cover

The Laurentian mixed forest. Lying in the warm-summer region of the cool-temperate zone, the Laurentian mixed forest occurs in the Great Lakes–St. Lawrence, the Upper Mississippi–Ohio, and the New England lowland regions. It is mainly deciduous hardwood—beech, maple, elm, oak, and birch—but has a good deal of coniferous softwood, including pine and the eastern hemlock. White pine, and white and red oak used to be abundant but have largely been cut out for timber. The elm is being killed out everywhere by the Dutch elm disease. The long, hot, wet days of summer, when tropical Gulf air predominates, lead to huge-crowned, large-leaved trees, which



Vegetation zones of North America.

shed their cover with the return of keen winter under polar continental airs. Rain or snow fall most of the year, averaging 35 to 40 inches, and thus provide ample moisture for dense growth. Deer are still common, although the moose is seen less often. Wolves, too, have largely been hunted out, a fate that the bear is also suffering, and beavers have been reduced to a small population. Squirrels still abound, but wild mink and marten are rare, and the carrier pigeon that once made this forest its home has become extinct.

The Eastern Upland Forest. Also known as the Acadian Forest in Canada, the Eastern Upland Forest clothes much of the central and southern Appalachians: here, polar continental airs are pronounced, while altitude modifies the tropical maritime winds. The growing season ranges from 90 to 120 days, and winter cold brings subzero temperatures. The forest, therefore, consists of fast-growing evergreen softwood species such as black spruce and balsam fir, along with alder and birch. Deer are still quite plentiful, as are such small fur-bearing animals as muskrat and squirrel.

The Eastern mesophytic forest. Extending from the mid-Atlantic states to northern Florida, the Eastern mesophytic forest is a mixture of hardwoods and softwoods. On the clays of river bottoms and the coast plain, great-crowned oaks form a tall, dense forest, mixed with hickory, walnut, and yellow poplar on the lower slopes of rivers, and ash and elm on higher slopes. Chestnut at one time was widespread but was virtually eliminated by disease in the 1920s. With summer temperature means of from 75° to 85° F (24° to 29° C), and a rainfall of 45 to 60 inches, many subtropical trees and bushes, such as pawpaws, crape myrtles, magnolias, laburnums, and mimosas flourish. Live oaks and gum trees are also distinctive of the area. On the sandy soils left by old stranded shorelines, magnificent stands of loblolly, longleaf, and slash pines form the Southern Pineries, now one of America's major sources of timber. The Virginia deer, black bear, raccoon and opossum are typical animals. Wild turkey, once very plentiful, but then rare through overhunting, are making a comeback.

Subtropical trees and bushes

Mangrove thickets. Fringing southern Florida and the Mexican lowlands facing the Caribbean, the mangrove thickets are backed by quick-oak and palms. Ibis fleck the woods with their gleaming white feathers. Moccasin snakes are still common in swamps, and alligators occur —not all in "farms."

The Western sclerophyllous scrub forest. In southern California and much of the American Southwest the Western sclerophyllous forest occurs. There trees have to

adapt themselves both to dry, hot summers when the tropical continental air is dominant and to wet, mild winters when polar Pacific air sweeps down from the north. A thin, short, open scrub of chaparral, or stunted evergreen oak, occurs, mixed with yellow pine and sagebrush. Pronghorned antelope, wild rabbit, the mountain puma, coyotes, land turtles, and snakes are common. The hawk is typical, preying on small desert rodents.

The tropical rain forest. The dense covering of all windward slopes in southern Mexico and Central America is provided by the tropical rain forest. The forest consists of such very tall, hardwood, broad-leaved evergreen trees as mahogany, ironwood, and palm, which form a spreading canopy over a lower tier of tree ferns, grape bays, gum trees, rattans, and mangrove, laced with lianas. Numerous species of plants are widely scattered in the forest. Wildlife is also varied with a great number of parrots, cockatoos and nutcrackers, with troops of monkeys, and many snakes. Panthers are still quite common. Ants, beetles, and flies feed on the decaying vegetation, and bacterial activity is high. This is an environment in which such tropical diseases as yellow fever, malaria, and blackwater fever historically have taken a heavy toll. A dry, tropical scrub of thorn trees, cactus, and sagebrush often takes over from forest in a remarkably short distance on leeward slopes and in rain-shadow basins.

Grassland and desert. Covering about a third of North America, grassland and desert are found in the drier and colder regions.

Tropical savannas. Located in patches in subhumid parts of Central America, tropical savannas usually occur at the intermediate levels of lee slopes and on plateaus. They are significant in Guatemala and the Yucatán Peninsula of Mexico. Heavy, though short-lived, summer rains bring on a thick rapid growth of tall grasses: cyclones associated with the northeast trades bring enough rain over the rest of the year to maintain a thin cover.

Temperate grasslands. The temperate grasslands, or prairies, form a belt between forest and desert, mainly on the Great Plains but also on the midslopes of the intermontane basins, above the desert flats. At the "break of the plains" on the eastern subhumid margin, invaded by rain-bearing tropical Gulf airs in spring and early summer, the grasslands are made up of a dense growth of tall grasses, such as the blue-stem sod grass and Indian grass, along with small berry bushes, wild roses, and stunted aspen trees. These are the tall-grass prairies that once were home to most of America's buffalo, before hunters virtually exterminated the breed. "Wet" prairie exists in the Middle Mississippi Basin, where a rainfall in excess of 30 inches occurs; their origin has been attributed to constant fire setting by the Indians, either to stampede buffalo in mass hunts or to burn off trees and thicken grass to provide buffalo pasture. Where rainfall drops below 17 inches (in Texas) or 15 inches (in Alberta), summer evaporation is very high, and tropical Gulf air provides less moisture, only enough for short-grass prairie. There, the vegetation is made up of blue-stem bunchgrass, thin needlegrass, tough grama and wire grasses, along with patches of cactus, sagebrush, or, farther south, mesquite grass. These regions still pasture many small antelopes, although they have been turned mainly into rangeland for raising cattle or sheep.

The "break of the plains" and the tall-grass prairies

Deserts. Creosote bush, mesquite, and cactus offer a very thin, open, stunted plant cover for the Southwest's dry intermontane basins and arid coasts, where rainfall drops below ten inches a year.

IV. The material-resource base

North America has rich and varied resources. Although it contains only 8 percent of the world's population, it has an extraordinarily high proportion of the world's wealth. By the 1970s, it produced about 24 percent of the world's oil, 18 percent of the iron ore, 21 percent of the steel and ferro-alloys, 38 percent of the copper, and 36 percent of the lead and zinc. With about 25 percent of the world's coal output and 37 percent of the world's electrical power production, it possesses the main sinews of modern world industry.

MINERAL RESOURCES

The shield. With a large shield area and with mountains strongly intruded by igneous rocks, the continent is unusually well endowed with metals. Its vast interior lowlands and some long stretches of coast plain are marked by major fuel formation. Metal-bearing regions include shield structures affected by mountain building or trough development, together with the intensely folded ridges that rose from the eugeosynclines on the periphery. The shield has four main metal-bearing areas: the iron of the Adirondack and of the Superior Uplands in the United States; the iron–nickel–copper and gold belt of Ontario and Quebec, along the old fold zones north of Lake Superior and Lake Huron; the iron of the Ungava trough; and the copper and nickel and the gold and uranium of the fault and fold systems of the shield's western rim. Iron in Baffin Land and cryolite in Greenland are additional outliers of wealth. Three areas are of special importance: the taconite hematite ores of Mesabi and Ungava, mined with relative ease by opencast, or open-pit, methods; the world's largest body of nickel in the norite intrusion at Sudbury, Ontario; and the large copper and gold bodies associated with the greenstone intrusions in northern Ontario and Quebec.

The marginal mountains. The Appalachians have also contributed significantly in metallic deposits, especially in the median mass between the Caledonian and Acadian folds, where lead and zinc are important in Newfoundland and New Brunswick; and also in the eastern or outer intensively folded rocks, with iron deposits in Belle Isle (Newfoundland) and the Trenton prong and in the Birmingham (Alabama) Basin.

The Cordilleras are rich in ores, mainly because of the immense igneous intrusions that underly many of the structures. The median mass between the Laramie–Rockies and the Nevada–Cascade systems has major gold, silver, copper, and iron ores in the old plateau of Colorado–Utah. Large lead, zinc, and copper ores occur in the Selkirks and adjacent ranges of British Columbia. Famous silver, lead, zinc, and gold ores dot the Cassiar Massif and Yukon Plateau, in the north; and far to the south, the Mexican Plateau holds iron, lead, and silver ores. To the east of this long, north–south line of plateaus lie the Rockies, which were not intensively folded and hence are not very rich in metals. But the vast intruded mass of the Idaho Batholith and the igneous bodies in the "dome and park" region were associated with copper, silver, and lead ores of great importance. The west Cordilleran ore deposits are widespread. They are found linked with the intensively folded and intruded rocks of the Nevadan–Cascade–Coast systems, notably the copper–gold ores of south and west Alaska; they also consist of the copper, lead, zinc, and iron of the enormous Coast Batholith of British Columbia; of the gold, copper, and iron deposits of Arizona and of the Nevadas in California, where the discovery of gold touched off the famous Gold Rush of 1849; and of the copper, gold, and silver of the Western Sierras of Mexico.

THE FUEL BASE

The fuel wealth of the continent is also great: coal, natural gas, and oil are found in large quantities. These fuels accumulated as vegetative, or carbon, deposits in the lakes and shallow seas of the great lowlands stretching between the shield and the marginal mountains. They also underlay the coast plains and the continental shelf, particularly in the Atlantic and Arctic.

Coal deposits. Coal deposits were preserved in basins between gentle upwarps in the buried extensions of the shield beneath the Interior Lowlands and also in mildly folded rocks in the miogeosynclines of the inner, less disturbed parts of the Appalachians and Cordilleras. Below the Mississippi–Ohio lowlands and the Great Plains, the outer edge of the shield became depressed and buried and then buckled into basins and warps. The Cincinnati Anticline created a vast elongated basin between mid-Ohio and the Appalachians, in which the western Pennsylvania, West Virginia, and Kentucky coalfields were preserved—probably the single largest coal reserve in the

Coal reserves of the Cincinnati Anticline

world—together with the Lima (Ohio) oil field. The Kankakee rise, south of the Great Lakes, has preserved coal and oil in the Michigan (Saginaw) Basin, to the north of it, and the Indiana and Illinois basins, to the south. The last named, also known as the Eastern and Western Interior fields, are separated from each other but kept close to the surface by the La Salle Anticline. The Llano uplift has similarly helped to preserve the Southwest Interior field in Texas.

An enormous midcontinental arch, the stem of the ancient Y-shaped structure connecting the Canadian Shield with the Colorado Plateau, separates the interior from the western coalfields lying in basins in front of the Rocky Mountains. Tolerably good bituminous coal occurs in the Raton Basin, which is cut off from the Denver Basin by the Las Animas uplift. The huge Williston Basin extends farther north, though it contains rather low-grade coal. Beyond this lies the vast Alberta Basin, with coal exposed in the foothills of the Rockies. Here, too, is one of the largest coal deposits in the world.

Oil and natural-gas deposits. The same coal-containing "rises" and basins in the buried shield also have controlled the distribution of oil and natural gas. The Appalachian oil and gas basin, on the west flanks of the Appalachians, was the first to be developed. The Illinois, Kansas, and Oklahoma basins lie in the huge quadrilateral formed by the Cincinnati Anticline and midcontinental arch to east and west and by the Kankakee rise and Ozark Dome to north and south. Between the Ozarks and the Eastern Sierras of Mexico are the tremendously wealthy fields of west and east Texas and of the Gulf Coast. Northward, between the midcontinental arch and the Rockies, are found a number of important fields including the Denver, Big Horn, and west Alberta fields, close to the Rockies, and the Williston, east Alberta, and Mackenzie Valley areas, halfway toward the shield. Small fields of oil and gas lie on the flanks of folded mountains within the intermontane zone, as at Paradox, Utah, and San Juan, New Mexico. The western basins, bordering the Coast Ranges of California, are of moderate size but very rich. In the extreme north, the Prudhoe Bay Basin of Alaska and Mackenzie delta oil have proved that the potentialities of the Arctic shore are real; domes —very much like the salt and sulfur domes of the Gulf Coastal Plain, associated with Louisiana's oil and gas— go with oil on the plains sloping away from the Innuitian fold mountains in the Canadian Archipelago.

WATER AND WATERPOWER

Availability and use. Considered as a resource, water and waterpower are also abundant, although rather unevenly distributed. The average rainfall in North America is 30 inches a year, which produces some 51,000,000,000 acre-feet of water. About a half of this is lost through evapotranspiration, which is direct evaporation plus transpiration from plants. A further sixth is lost through rapid runoff, while yet another sixth percolates down into the groundwater. The amount available from rivers and lakes is thus relatively limited, a fact of growing concern as the demand for water grows. It takes 16,000 gallons of water, for example, to produce one ton of steel; while an average steam-generated power plant requires 4,000 cubic feet of water per second. In 1960, the United States was using a little less than half its potential available water supply of 3.5×10^{14} gallons per year, but, by AD 2000, it will be needing over two-thirds of that supply. The gap between use and availability is narrowing, though through the pumping up of artesian water, and through the desalinization of seawater, the supply may be considerably increased.

The narrowing gap between water use and availability

Regional distribution. The water resources of the continent vary with regions. In northern Alaska, the Canadian north, and Greenland, they are low, mainly because they are locked up in ice most of the year; when the summer melt comes, runoff is high. Central Alaska and Canada midnorth have a moderate precipitation of from 12 to 15 inches per year, but again much of this is tied up in winter ice. The spring melt leads to extensive flooding, which makes the control of water difficult. Late summers are dry, but evaporation is low. Though rivers dwindle, they carry enough water for present needs.

Water problems in the Great Plains

The Great Plains area also has marked high- and low-water periods, the latter posing serious problems. Most rivers rise in mountains with an extensive snow supply. Meltwater gives an early spring flush; high flow is continued into early summer through storms generated by the tropical Gulf air. In late summer and the fall, however, dry air descends from the tropical continental air mass behind the mountains, the storms cease, and the rivers drop. Streamflow originating within the region may dry up; the bigger rivers sink into braided channels between bars of sand. Evapotranspiration also exceeds precipitation, resulting in a net loss of water—not made up again till early winter with the return of polar-front storms. Surface water therefore is scarce for four to five months and often needs to be supplemented from groundwater.

The intermontane basins stretching from southern British Columbia to central Mexico exhibit a strikingly unequal pattern, with areas of water surplus in the mountains lying adjacent to areas of a marked water deficit in the basins. Major rivers like the Columbia, Colorado, Grande, and the Guadalajara rise in snowy or rainy mountains and supply enough water, especially where their waters are trapped by dams, to serve the basins through which they flow. Lesser rivers, however, often peter out and are intermittent. Groundwater supply in areas with artesian wells alleviates the situation.

The eastern parts of southern Canada and the United States are well watered, with rainfall in most months, as the southern movement of polar continental air and the northern expansion of tropical Gulf air draw storms regularly across the area. Rainfall is from 30 to 60 inches a year, and evapotranspiration is not in excess of precipitation except in late summer. Streamflow is perennial, averaging over ten inches in depth per stream per year.

Finally, the tropical areas in the trade-wind belts in Central America are well supplied with water, yet less is available than might be expected due to swift runoff after heavy rains and high evaporation. Rivers are relatively short and steep and are likely to flood quickly.

FOREST RESOURCES

Forests are another of North America's magnificent natural resources. By no means all the forested land, however, may be regarded as a source for pulp or timber. Probably two-thirds of the boreal forest, for example, is not usable because of the thinness of the cover, the shortness of the trees, the very slow rate of natural recovery after cutting, and the inaccessibility of the northern parts of the region. The southern third provides a major base for pulp and paper industries. Most of the northern mixed forest and eastern hardwood forest likewise has become of little service. Good timber was cut out long ago, with much of the forest completely cleared for agriculture. What remains, though it looks plentiful, is commercially unattractive second- or third-growth bush. This woodland, however, is gaining value for the buildup of wildlife and for recreational purposes. The Pacific coast forest and some of the Cordilleran forest still make excellent stands for timber, and, though cutting has sometimes exceeded natural replacement, yet, with controlled use, growth is generally rapid enough to produce a continual yield. The Southern Pineries, long left untouched by agriculture because of their sandy soils, now have become the main source of timber and pulpwood in the United States. The tropical hardwood forests are also important for timber, but they have been cut down to some extent and also have been replaced by banana plantations or by poor second growth. Oak–pine forests above 2,000 feet, in the *tierra templada* ("temperate land") remain a useful resource.

THE FUTURE OF THE NATURAL ENVIRONMENT

The early attitude of the Europeans in North America favoured the clearing of the forests and the killing off of the wildlife, with the aim of making room for crops and domesticated animals. In a continent that was so vast and at the same time so empty, they also developed the idea that environmental resources were unlimited

and only awaited the coming of the white man to be tapped. It should also be remembered that many of these immigrants were to come from a Europe in which, during the Agricultural and Industrial revolutions, there had been an increasing attack on natural resources, particularly associated with the rise of industrial cities. When the United States and Canada became industrialized, they used coal, oil, iron, other metals, and wood with extravagance, a process involving great waste. The waste products of the factories of these nations started to pollute air, land, and water; and, as the multimillion cities began to appear, the majority of people came to live in a man-desecrated as well as a man-improved environment. By the mid-20th century, the people of the United States had killed off about four-fifths of the nation's wildlife, cut over half its timber, and used up two-fifths of its high-grade iron ore; the nation was consuming its oil so fast that, even with its great resources, it began to import from Canada, Venezuela, and other areas. Conscious of the great drain on the resources of the nation, and suffering from the increasing ill effects of pollution, the United States began to conserve its valuable reserves of forest, soil, water, fuels, and minerals; and today the country leads the world in its conservation programs, particularly in renewing the forests and grasslands, repairing the soils, and effectively controlling the waters. Canada, too, has an active conservation program and was the first North American country to pass a clean-water act to help fight the pollution of its lakes and streams. Mexico likewise has an active, though small, conservation service. The nations of North America seem determined in the 1970s to undo the damage done and to pass on to future generations a revivified environment.

Conservation programs

V. Human resources

North America long remained a relatively empty and undeveloped land in global terms, but, with the coming of the Europeans and the Africans they introduced, it began to fill up rapidly with people of diverse traditions and skills. The section that follows covers primarily the peoples of mainland North America. The ethnohistory of Greenland, which, simply because of distance, is quite distinct from that of the rest of North America, is covered in the articles GREENLAND and WESTERN ARCTIC CULTURES. For the human resources of the Caribbean, see the articles on the states of the region and their history and the articles CARIBBEAN CULTURES; CENTRAL AMERICAN AND NORTHERN AMERICAN CULTURES. See also NORTH AMERICAN PEOPLES AND CULTURES; MIDDLE AMERICAN PEOPLES AND CULTURES; and articles on individual cultural areas of mainland North America (*e.g.*, AMERICAN SUB-ARCTIC CULTURES; CALIFORNIAN INDIANS; and NORTH MEXICAN INDIAN CULTURES).

THE NORTH AMERICAN INDIAN HERITAGE

The Indians themselves arrived late in North America. Having originated in Asia, they had a long way to travel —through Siberia and across Alaska—before getting into the heart of the continent. The Ice Age began shortly after man emerged and blocked any northern advance into Siberia and Canada throughout most of the Pleistocene. It was only during the interglacial periods between ice advances that man was tempted to move north. He may have come to North America before the Wisconsin advance, about 60,000 years ago. It is more probable he chose an interglacial period during a recession in Wisconsin time, probably before the Mankato readvance.

The Indians came in as Stone Age hunters leading a nomadic life, and many remained in this condition until the white man came. In moving down from the narrow neck at Alaska to the broad expanse of the continent between Florida and California, the tribes tended to separate out and hunt in comparative isolation. Until they came to the narrows of southern Mexico and the confined spaces of Central America, there was little of the fierce competition or the close cooperation between them that might have stimulated progress. Though they made great progress in those southern regions, they made fewer advances in the use of metals, the growth of industry, or the development of transport and trade than, for example, did the contemporaneous civilizations of Asia, Europe, and parts of Africa. City life arose first among the Olmecs, in the strategic narrows between Mexico and Central America, and the Mayans, in the plateaus of Guatemala and Yucatán. Subsequently the Toltecs and the Aztecs rose to power and developed notable cities in the high Mexican Basin. These people flourished on a rich agriculture based on maize, beans, and squash, along with manioc, potatoes, tomatoes, tobacco, and cacao. They also raised cotton and worked leather.

Some authorities contend, however, that these civilizations already were on the decline before the Europeans came, having been divided by wars and riddled by disease, with much of their land wasted by erosion. Louis de Velasco, a 16th-century Spanish governor, put the total population of the West Indies, Mexico, and Central America at probably 5,000,000 Indians. The population of the less developed Indians of what now is the area of the United States and Canada has been variously estimated at somewhere between 600,000 and 1,200,000. The Indians there had not developed intensive agriculture or a city way of life, though it is true that the raising of maize, beans, and squash supplemented hunting throughout the Mississippi–Ohio and the Lower Great Lakes–St. Lawrence river regions, as well as along the Gulf and Atlantic Coastal Plains. In those areas, semisedentary peoples had established villages, and among the Cherokees and the Iroquois quite powerful federations of tribes had arisen. Elsewhere, however, on the Great Plains, the Canadian Shield, the northern Appalachians, and the Cordilleras, fishing and hunting constituted the basic economic activity and required an extensive territory to support and feed a small population (see also NORTH AMERICAN PEOPLES AND CULTURES).

THE EUROPEAN HERITAGE

The white concept of the environment. When European colonizers arrived, they regarded much of the continent as empty and waiting to be developed. To William Bradford, the early historian of New England, the white races were moving into a virtual wilderness offering a wonderful opportunity for settlement and development. The indigenous agriculture was rudimentary and scarcely opened up the forest; there were no herds or flocks to make a better use of natural pasture other than those of moose or deer; the value of the forests for making homes, buildings, fences, roads, and tools seemed unknown; no mills exploited the waterfalls along which the whites were soon to establish the Fall Line of manufacturing cities; the deposits of iron and coal had not been developed. Seen through the world view of the Europeans, who had had a long tradition of iron and steel making, of maximizing use of the forest for fuel through charcoal, of coal mining and the mining of a great range of metals, of linen and woollen weaving, and of harnessing both water and wind for the operation of mills, the comparative backwardness of most of North America offered both a challenge and an opportunity. They saw no reason why they should not claim land that they could use to better advantage, and hence they bought out or pushed out the Indian and took over the country.

The dispossession of the Indians. The process of removing the Indians from their lands led to bitter disputes, which the British tried to end by setting up the Proclamation Line of 1763 along the Appalachian divide, allowing whites to take over what lay to the east but attempting to reserve what lay to the west as Indian territory. After independence the Americans continued to adopt this ideal of virtually a two-nation state, but in practice it soon collapsed as they pushed the Indian line to the Ohio, then to the Mississippi, and then to the Missouri. The Indians in the east were swiftly displaced to the west and pushed into Indian territories on the Great Plains and in the intermontane basins. Even there their land and their way of life were not respected, as ranchers, the railways, and the homesteaders opened up the West. The Indian "territories" soon were reduced to isolated "reservations." As a

result, most of the contemporary Indian population in the United States is found west of the Missouri—indeed, a great deal of it farther west between the Rockies and the Sierras in the dry climates unsuited to extensive agricultural settlement. The most recent policy of the United States government has been to encourage Indians to leave the reservations and mix with the whites in the great cities of the Midwest and the West, a process that some—including many younger and ethnically conscious Indians—would claim to pose as many dangers to the individual and social well-being of Indians as the communal misfortunes visited upon them by history, though others see it as a means of ending white–Indian differences.

Canadian policy and system of reservations

In Canada, the system of reservations was early adopted and protected Indian settlements throughout the east, even in rich agricultural areas like the Montreal plain or peninsular Ontario or next to great cities like Montreal. As the whites moved west, more care was taken to retain Indians in part of their lands, even on such fertile plains as the Red River Valley or the Fraser Delta. Intermarriage between whites and Indians was much more common than in the United States and led to the nation of the French-and-Indian Métis, which in the Riel Rebellion unsuccessfully attempted to set up a separate state in the Canadian prairies.

In Mexico, racial admixture has gone on much further, and the mestizo, of mixed Indian and white descent, accounts for fully 55 percent of the population. Pure-blooded Indians amount to only about 29 percent; whites make up 15 percent, and Mexicans of African descent one percent of the population. European immigration to Mexico and Central America since the original Spanish conquest has been negligible. When de Velasco made his 1574 count of 5,000,000 Indians under him, there were only 150,000 Spaniards in the New World.

The policies of the colonizers. *Spanish policy.* Colonial policies strongly affected the evolving human resources of North America. It is often stated that Spanish interests centred around God and gold—the Christianizing of the Indians and quick wealth from gold and silver. This is, of course, an overstatement. They were also interested in land, which they wished to develop as great ranches or plantations; but this land was carved up into big estates, to be worked by tenants or by direct labour in the form of serfs or peons. As a result, not many Spanish settlers were attracted. Spanish policy brought a highly competent entrepreneurial and professional group: mineowners, the owners of great estates, merchants, administrators, and priests to North America, but comparatively few from the middle and working classes. Little industry was set up, and, though towns were important, they served mainly as centres of trade and services. Even today the whites remain very much of an elite in the area.

French policy. The policy of France was much the same. The first Frenchmen on the continent were mainly entrepreneurs interested in the fur trade; they hired Indians to collect and carry furs from the hinterland to the French trading posts, and they opposed, sometimes violently, the idea of farm settlements. France, nevertheless, felt it necessary to have a stronger French-population base in its new colonies. To induce settlement, therefore, it gave large seigneuries, or grants, to landed proprietors who promised to bring in settlers, clear the forest, and develop the country. The seigneurs came with the French ideas of tenant farming, under which the seigneury was divided into many small holdings, each paying rent for the land. Once again, few Frenchmen were moved to go to the New World under Old World conditions. As a consequence, when the British took over from the French in Canada in 1763, there were only about 80,000 Frenchmen in Acadia and Quebec, whereas the British on the continent then numbered 3,000,000. The French, however, had been there for more than 150 years, and, since they were allowed by the British to retain their own language, religion, school system, and laws, they kept up their traditions in a remarkable way. Today the people of French origin in Canada total about 27 percent of the population and are firmly entrenched in the province of Quebec, with sizable communities in New Brunswick and Ontario. They form a distinctly Latin element within the Anglo-American realm, and, indeed, some of them would like to see an independent French-speaking country, based on a free Quebec, a situation that has generated some intercommunal strife since the 1960s.

The Quebec French community

British policy. British policy encouraged the large-scale settlement of its colonies by religious refugees, landlords anxious to get New World estates, freemen seeking land of their own to develop for themselves, businessmen trying to found new enterprises, millowners and workers engaged on necessary domestic industries, and craftsmen and professional men out to capitalize on their own skills and callings. The British colonies, therefore, attracted many white settlers with a wide range of competence. The British also opened their colonies to non-British white settlers, notably such religious minorities as the Huguenots, from France, and the Mennonites, from Germany. Under the Hanoverian regime, many German mercenaries were also settled. In Canada, the British opened up the Canadian west to Germans, Scandinavians, Ukrainians, and Poles on a large scale and later accepted Chinese and East Indian settlers.

United States policies. The newly independent United States took over many British policies but gave more scope for freehold land and especially—after the overthrow of the Southern plantation system of chattel slavery—for capitalist business and industry. It also adopted the Open Door Policy to refugees and others from Europe in the belief that America would become, at least for the white races, the "melting pot" of the world and thus in a sense develop what the historian Turner called a "new race of men." These views exerted a powerful attraction for British, Irish, and continental Europeans, and, perhaps to a lesser extent, for people from the Middle East, China, and Japan. Subsequently, as free land was used up, and as the United States became more interested in skilled workers and professional men, a quota system of immigration limited nationalities according to their proportion of the total population after World War I. Still later, a personal selection policy was adopted to choose those of any race or creed who, as individuals, were deemed best able to serve national needs.

THE AFRICAN HERITAGE

North Americans of African descent were originally brought to the continent involuntarily, as Negro slaves. In the United States alone, black citizens now number almost 23,000,000 (authorities concede there is some under-representation in official census data), or more than the entire population of Canada. The whole question of the transatlantic slave trade and its turbulent legacy is still fraught with deep emotions, a not unexpected development considering that this tragic episode set in motion forces and counterforces that molded much of European and African, as well as North American, society over the course of nearly four centuries. Slaves were not widely used in Mexico or Canada but were imported in great numbers to the Southern states of the United States, where they laboured on tobacco, cotton, and sugar plantations. The importation of slaves to the United States was officially abolished in 1808, but illicit traffic and natural increase kept the black population growing in conditions that took a high toll in terms of difficult physical existence, disrupted social and family life, and harm to the individual's psychological makeup. The abolition of slavery in 1865 gave Negroes, in theory if not in practice, freedom to work and live where they liked. Many saw no alternative to staying on plantations as impoverished tenants or sharecroppers. Economic forces in the 20th century—particularly the mechanization of farming—brought increasing numbers of black Americans to the great ghettos of the Northern cities, where new problems of racial discrimination had to be faced. The resulting social and racial polarization struck at the very roots of United States society and challenged the sincerity of its professed egalitarianism. Although the black citizens' struggle to win full equality made some progress, especially during the 1950s, 1960s, and beyond, their battle remained far from won, well over a century after the Emancipation Proclamation. Schooling, housing, social services, and,

The slave trade and its consequences

above all, access to employment opportunity on an equal basis remained key areas of concern. Perhaps because of this systematic exclusion from the mainstream of United States society, the African contribution to the life of the major nation of the continent has generated a rich and unique Afro-American culture.

Problems in race relations have also molded the attitudes and institutions of the white majority, not least in its dealings with the peoples and races of other continents. Hawaii has nevertheless been successfully incorporated as a state, and the United States as a whole has taken in numbers of Polynesians and many more Japanese. It has become quite a polyglot nation. The resulting rich ethnic mix, while posing immense problems, may perhaps ultimately prove an asset rather than a liability in the evolving world community.

THE NATIONS OF THE CONTINENT AND THEIR ALLIANCES

The United States and subsequently Mexico and the Central American countries cut off their ties with Europe and became independent republics. They continue to recognize each other's independence and also their need for each other in various regional groupings and alliances, including the Organization of American States (OAS). Cuba has nevertheless taken a different course of political, social, and economic development than the rest of North America and the Caribbean region. Canada has not joined the OAS because, unlike other American states, it has ties with the United Kingdom and other members of the Commonwealth of Nations. It has adopted the British parliamentary system of government rather than the American system of checks and balances. Both Canada and the United States are, nevertheless, members of the North Atlantic Treaty Organization (NATO), along with most other countries of western Europe. The United States in addition is a member of the Southeast Asia Treaty Organization (SEATO) and of the Australia, New Zealand, and United States agreement (ANZUS). North America thus has a strong hemispherical unity and in addition is firmly linked up with the British Commonwealth, with western Europe, Southeast Asia, and Australia and New Zealand. There are also vital trade links with South America, Africa, and Japan. North America's major ties are with other predominantly industrial nations, but all these connections are of immense importance for the continent's security and prosperity.

Position of Canada in ties with regional groups

VI. Resource development

AGRICULTURE

The various peoples that have developed North America have made it a world economic leader and, in general, a well-used and productive continent. Agriculture, though no longer the principal economic activity, except in the southern Latin nations, is still important.

Tropical areas. In tropical areas, the Spaniards made the most of the strong altitudinal zonation by raising sugar in rainy parts of the low *tierra caliente* ("hot land"), wheat and cattle in the middle levels of the *tierra templada,* and sheep on the upper slopes in the *tierra fria* ("cold land"). Later, orange groves, coffee, cocoa, and banana plantations utilized the coast plains and wet windward slopes of the tropical areas; cotton and hemp were grown in the warmer and drier basins of the intermediate zone. These remain important export crops for Central American countries and Mexico, being shipped mainly to the United States and Europe.

Subtropical and warm-temperate areas. An enormous extension of fruit, winter vegetables, cotton, and tobacco farming has occurred in the subtropical and warm-temperate areas of the United States. Citrus fruits do well in east Texas and Florida, where the Ozarks and Appalachians protect them from polar airs, and the Gulf invites warm tropical air with early rain but much late-summer sun. The Central Valley of California—guarded from frosts by the Sierras, with winter rain for growth and prolonged summer sun for ripening—also is a prime citrus-growing area. Drought is a challenge, however, and has been met only by extensive irrigation. Winter vegetables are widely grown on the sandy soils of the Gulf Coast Plain and the southeastern parts of the Atlantic Coast, with a long frost-free season (200–340 days) and ample rain. Cotton has proved a success in areas with less than 60 inches of rain and over 200 days free of frost. Tobacco has concentrated on the sandy soils of old shores and deltas, from Virginia to Kentucky. Many of the tobacco and cotton fields are now intertilled with rye, corn, and winter wheat grown as fodder for cattle or as additional cash crops. These help to maintain the fertility of the soil, which has long been threatened by the practice of monoculture.

The decline of monocultures

Cool-temperate, humid regions. In the continent's cool-temperate, humid regions, crops include hardy fruits grown on the valley sides of the Appalachians and the Piedmont from Georgia through Virginia; the Finger Lakes region of New York; the Niagara Peninsula and the east shore of Lake Michigan; and parts of the Columbia Basin in Washington and British Columbia. In all these areas, aspect, frost, and drainage are important factors.

The zone known as the Corn Belt evolved from a concentration on corn in the warm-summer region eastward from the Ohio River to the Lower Missouri shores, where winter snowmelt, spring rains from the northward surge of tropical Gulf air, and early summer convection showers bring on the plant, while strong late-summer sun, with July means of 70° to 80° F (21° to 27° C), ripen the cob. Most of the corn is fed to fatten pigs and cattle.

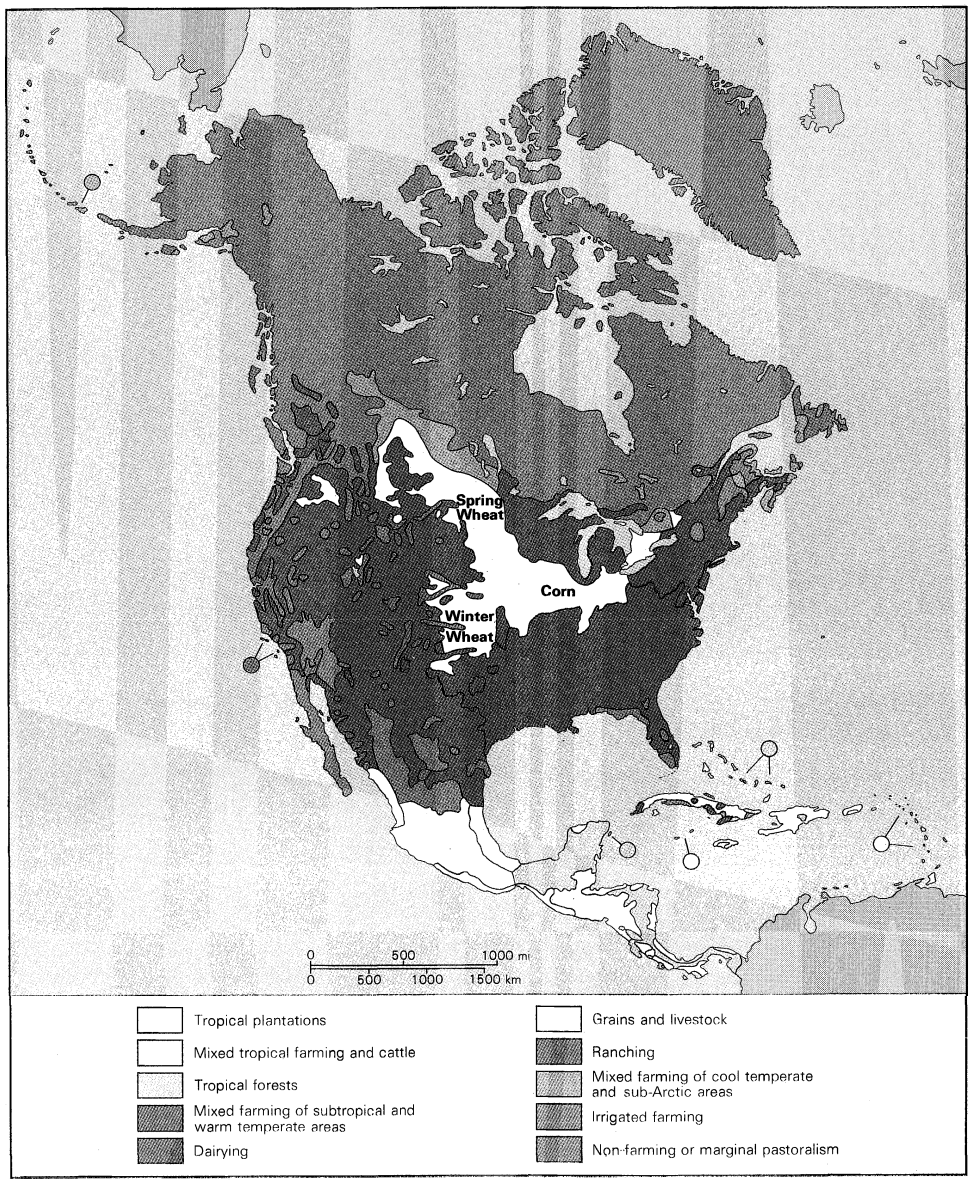
The Dairy Belt, another recognized division, makes use of a shorter growing season and cooler summers in New England and the Great Lakes–St. Lawrence region, where clover, timothy hay, and hardy small grains thrive. Dairying also exploits the lush pastures of the Pacific coast equable climate in Washington and British Columbia.

West of the Corn Belt, in subhumid regions, lie the continent's vast wheat areas. The Winter Wheat Belt, mainly in Nebraska, lies south of killing frosts. As the polar front retreats, the early sweep of rainstorms brings on the grain sown in the previous fall. The Spring Wheat Belt—in the Dakotas, the Canadian Prairies, and part of the Columbia Basin—has a severe winter that forces the postponement of sowing to spring. Then the warmth and wetness of the sudden surge north of tropical Gulf air bring on the new-sown wheat very quickly, which ripens in a usually dry, sunny fall. Wheat farming is being carried on on an ever bigger scale, using more machines and producing more per acre.

Dry regions. Dry areas in the Great Plains and intermontane basins were long left to ranching. The Hereford cattle brought in from England could feed on the short-grass prairies, which were not suitable for the homesteader. The sheep, raised in still drier parts or up in the mountains, are bred mainly for wool. Near rivers or in artesian areas, irrigation for supplementary fodder has greatly helped ranching. Irrigation, however, is being used increasingly for fruit and cotton farming, resulting in a great drain on water supplies.

DEVELOPMENT OF WATER AND HYDROELECTRIC RESOURCES

Water development. Water development is crucial both to circumvent drought and prevent flooding. Irrigation in the United States consumes 85,000,000 acre-feet of water per year as compared with industry, which needs 25,000,000. About 40,000,000 acres of irrigated land have been developed in the United States, with large schemes of dams and conduits in the Columbia and Snake valleys, the Central and Imperial valleys of California, the Salt and Gila tributaries of the Colorado, the Upper Rio Grande, and more recently, the Upper Missouri and the Upper Platte. In western Canada, a vast scheme is developing on the Bow–South Saskatchewan rivers, while in Mexico the Lower Rio Grande, shared with the United States, the Fuerte Basin on the dry west coast, and the Balsas in the south have actively promoted water development. Water transfer from surplus to deficit areas is already taking place, and interstate water transfer proposals include transfers from the Columbia Basin to both the Sacramento and the Colorado, and from the



Agricultural regions of North America.

head of the Missouri system to the Colorado, and thence to the Gila. Flood control is a problem in the Mississippi Basin. The Tennessee Valley and the Ozarks' schemes have involved building many dams to pond up and redistribute river water.

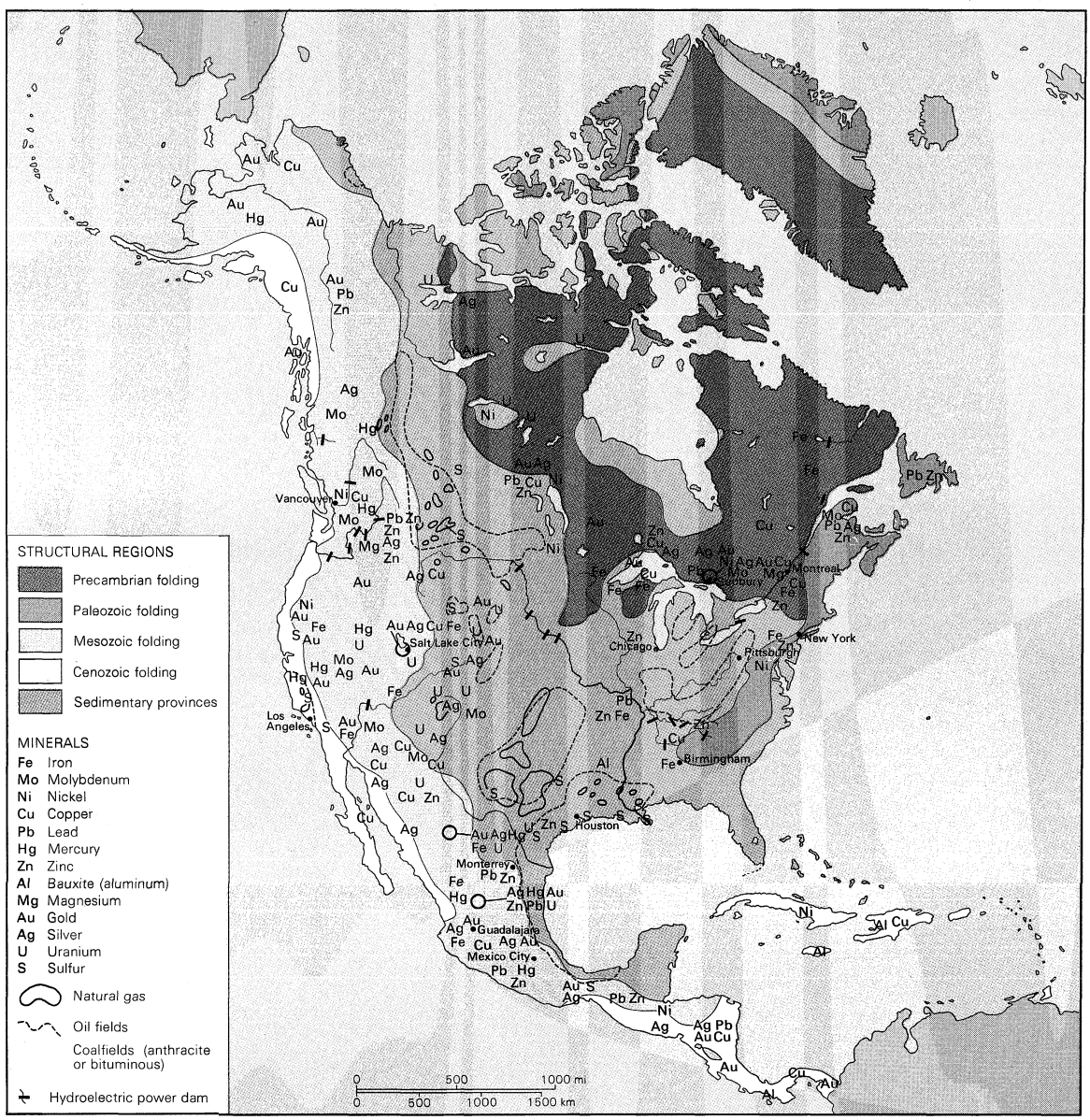
Hydroelectric development. Hydroelectric development has been immense in the United States and Canada, which rank first and second in the world in their installed generating capacity. The rivers of the Canadian Shield, fed from lakes and falling abruptly over the edge of the plateau, provide many sites, especially in Quebec and Ontario: these are linked to such Great Lakes–St. Lawrence sites as Niagara Falls and International Falls, which, in their turn, tie in with a power grid developed from Appalachian rivers. The north central and northeastern areas are thus well supplied.

The snow-fed rivers from the high Cordilleras, where dammed (as at Grand Coulee, Hoover, Glen Canyon, Fort Peck, and Garrison), also provide an immense amount of power. Yet in the U.S. hydroelectric power represents only one-fifth of all electricity generated, the rest coming from coal- or oil-fired plants and nuclear-power stations. In the north, steam plants based on coal are not subject to winter freeze-up, as are hydro plants.

FUEL DEVELOPMENT

Development of new fuel sources has seen the dramatic displacement of coal, from nine-tenths of the energy used to less than a fifth in the 1970s. Oil contributes about two-fifths; and natural gas, just over a third. Although coal production in the U.S. has risen from 510,000,000 tons in 1940 to about 610,000,000 tons in 1970, the industry's work force has been cut by more than two-thirds. The vast reserves of coal are thus being mined more efficiently, and workings are now concentrated mainly where strip mining is effective, in West Virginia and Kentucky, in the Warrior field (Alabama), and in eastern Illinois. Coal is sent to the big power plants and steel works of the mid-Atlantic region and the Lower Great Lakes. One reason why oil and natural gas have replaced coal is their ease of transport. Pipelines carry both fuels from their remote sources in the Gulf fields of Louisiana and Texas, the midcontinental fields of Oklahoma, and the Rocky-front fields to the shoreline cities of the Atlantic, Great Lakes, and Pacific. By the 1970s, there were more than 700,000 miles of gas pipeline as compared with 214,000 miles of railroad. Oil, too, is piped great distances, although much is sent by tanker from the ports close to the Gulf oil fields. Consumption has increased enormously; in spite of the richness of the oil fields in California, that state has become a net importer, piping in oil from Texas. Similarly, the Illinois, Ohio–Indiana, Michigan and west Pennsylvanian fields, though quite important before World War II, cannot possibly supply the Great Lakes and mid-Atlantic regions, which are fed by oil piped from Oklahoma and Kansas. The U.S. now im-

Changes in energy sources



Basic structural regions and principal mineral and hydroelectric sites of North America.

ports much oil. In Canada, the industrial regions at Vancouver and in the Lower Great Lakes–St. Lawrence area are fed gas and oil by pipe from Alberta; and, in Mexico, Mexico City and Monterrey, on the central plateau, are supplied by pipe from the Gulf Coast oil fields around Reynosa and Tampico–Tuxpan.

INDUSTRY

The role of ports in industry

Coastal sites. The industry of North America is its chief contemporary source of wealth. It first grew up at Atlantic Coast and Mississippi River ports, where raw materials transported from abroad or brought by coastwise trade from other colonies could be made into goods for distribution in the interior. Inland products also might be transformed before being exported from such ports, where immigrant labour was plentiful and capital brought in or developed locally was abundant. In many respects, the ports still perform these roles. New England ports bring in wool, leather, hardwoods, and metal from abroad and cotton from the Southern states; New York imports coffee, cocoa, sugar, timber, pulp, and oil; the Philadelphia region brings in great quantities of iron from Canada or Venezuela and oil and petroleum coke from the Gulf; together, these cities manufacture textiles, leather goods, petrochemical products, iron and steel, ships and machines, books, clothes, and foods not only for their own dense populations but for the interior United States as well. In spite of the enormous development of the interior, coastal sites have still remained paramount, especially if the definition includes the shores of the Great Lakes and the Puget Sound and California coast. From Buffalo through Detroit to Chicago, the movement of coal from the Appalachian and eastern interior fields up to the lakeshore, combined with the shipment of iron ore from Lake Superior and Ungava to the lake ports, has led to a vast and dynamic belt of iron and steel, transport facilities, and machine-making cities. The Pacific ports of Seattle, San Francisco, and Los Angeles have developed from an outpouring of forest, fish, farm, mine, and oil-well products, partly abroad and partly by the Panama Canal to eastern America.

Similarly, in Canada both the importing of oil, wool, cotton, leather, and food-based raw materials into the St. Lawrence–Great Lake ports such as Montreal and Toronto and the exporting of Canadian iron, nickel, copper and other metal-based goods, of wood products, and flour from the eastern ports and from Vancouver have concentrated population at the gateways in and out of the country. Mexico's gateway district, at Vera Cruz, is also industrialized. Most of North America's industry and population is thus concentrated at its seaward or Great Lakes' margins.

Inland sites. Sites in the interior are, however, not without importance. The first to develop were the fall-line

power centres, strung out from the falls of the Merrimac at the edge of the New England Upland, then southward along the east front of the Piedmont, to the Coosa south of the Appalachians. Later, with the advent of steam and electric power, these sites continued as major textile, pulp and paper, and engineering locations. A bigger shift inland occurred with the use of coal for power, in the eastern and western Pennsylvanian coalfields around Wilkes-Barre and Pittsburgh, in the Birmingham coal and iron fields, and the Saginaw Bay, Indiana, and Illinois coalfields. Pittsburgh soon used up its local iron ore but was sufficiently near the Great Lakes to bring in Mesabi Range iron ores, which, in combination with the vast amounts of high-quality coking coal at hand, formed the basis for a great iron and steel industry. Except where coking coal is used in the steel plants at Pueblo, Colorado, Geneva, Utah, and Fontana, California, Western coal has been too low grade to attract industry. Oil and natural gas, however, have become the base of active petrochemical industries in Alberta, Oklahoma, and Texas. Since oil and gas can be easily piped, they have not stimulated the development of industry on a big scale near their sources but have fuelled the northeastern and Pacific Coast industrial areas. Modern industry has become less tied to sites where fuel and raw materials are available and more oriented toward the market. Service industries especially have concentrated in the highly populous areas of Boston–New York–Philadelphia, Pittsburgh–Detroit–Chicago, and San Francisco–Los Angeles. Space-age developments have been supported by science-based industries from Texas through Louisiana to Florida. Industries to meet the immense demand for travel and recreation are springing up on the major motorways and in the tourist areas in the Appalachians, the Cordilleras, and along the sea coasts. Though industry is more free to disperse, it nevertheless continues to focus on areas of existing urban agglomeration. In the United States its growth is greatest in the New York–Washington, Cleveland–Chicago, and Los Angeles regions; in Canada, in the Montreal–Toronto and Vancouver districts; and in Mexico, in the Mexico City Basin. These major cities are also unhappily the focus of critical social and economic problems.

More generally, automation is everywhere creating a major problem of technological unemployment, met in part by reducing working hours and retiring people earlier. These trends, in turn, raise the problem of the use of leisure time—now the target for much of America's fastest developing industries.

TRANSPORTATION AND TRADE

Waterways. Industry has been strengthened by the ease of movement in North America. Waterways, widely used by the Indians and early Europeans, are still important. In spite of the barriers of the Laurentian Shield and the Appalachians, the routes up the Gulf of St. Lawrence, the Hudson Estuary, Chesapeake Bay, and the Gulf of Mexico permitted the swift development of coastal ports and allowed the interior to be opened up. The Mississippi–Ohio and the Great Lakes–St. Lawrence waterways drew navigation into the heartland. Connecting these two systems, the Chicago Sanitary and Ship Canal, linking the Illinois River with Lake Michigan, and various Ohio–Lake Erie canals provided a tremendous network, extended by the Erie Canal to the Mohawk–Hudson and by the Intracoastal Waterway to river ports of the Gulf of Mexico. The St. Lawrence Seaway, which overcame the Lachine and International rapids and Niagara Falls, has made ocean ports of Chicago and Toronto. No other continent has such a system of inland waterways.

Railways. *Development of the rail network.* Railways soon offered the challenge of more direct and speedy access than the waterways. Developed principally from bases in Baltimore, Philadelphia, New York, and Boston, they made the most of gaps through the Appalachians, debouched on the Great Lakes or Ohio at Buffalo and Chicago, Pittsburgh, and Cincinnati, and pushed on to the Mississippi at Memphis, St. Louis, and St. Paul–Minneapolis. Other lines were then thrown across the Great Plains and, making the most of Cordilleran passes, built terminals at San Francisco, Seattle, and Los Angeles. Most of the Western railways were given large land grants to encourge immigrants to settle along them, while low promotional rates on long-haul traffic developed transcontinental trade. In Canada, the transcontinental railways linked up the Maritime Provinces with the St. Lawrence–Great Lakes, and thence, from Montreal and Toronto, they crossed the shield to converge at Winnipeg; here, reinforced by large land grants, they fanned out across the prairies to be drawn together by the Fraser down to Vancouver.

Development of Canada's railway system

Mexico overcame difficult grades in building a railway from Vera Cruz to Mexico City and added extensions north and south along the Gulf Coast, with lines into Monterrey and to Mérida. Eventually lines were pushed through the Western Sierras at Guadalajara to the Pacific coast.

Railways and urbanism. Railroads had a tremendous impact on urban development. The major railroad cities are New York, Chicago, St. Louis, and Los Angeles, in the United States, and Montreal, Winnipeg, and Vancouver, in Canada. Mexico City dominates the network in Mexico. Railroads led to the rise of east–west over north–south lines and rapidly displaced most waterways, particularly the Mississippi. The main economic axis in the United States lies along the railway belt from New York to Chicago. Inadequate overall planning in major metropolitan regions resulted in crucial transportation problems by the 1970s, and inner-city rapid-transit systems often fared no better.

Motorways. North America's road network first began to offer serious competition to the railways after World War I. The United States government financed over 200,000 miles of transcontinental highways, and U.S. Highway 20 (from Boston to Portland, Oregon) became the axis of midcentury America. In Canada, the Trans-Canada Highway offers a coast-to-coast through route, while from Mexico, the Pan-American Highway links all the countries of Central America. These highways have enabled trucks to take over short-haul routes, and the financially plagued railroads have had to concentrate on long-haul, low-cost routes. Truck and train are, however, being integrated in "piggyback" containerized carriage. The car, meanwhile, has displaced commuter trains in many cities, and radial and ring routes are drawing the cities well out into the countryside. The attendant problems of congestion and pollution have approached the critical stage in many cities.

Air transport. Air transport has taken most of the long-distance passenger traffic from the trains, and air-freight has cut deep into truck-freighting trade. Intense overall competition is thus a recurrent feature of North American transportation systems. Airways have tended to centre on the larger cities and to magnify their importance, so that intervening intermediate and smaller cities have declined in importance. Links with Europe and Asia make North America the chief crossways of air routes in the world. The United States alone accounts for more than a third of all the world's air traffic. Of the ten chief airports in the world, nine are in the United States—two in New York and one each in Chicago, Los Angeles, San Francisco, Atlanta, Washington, Miami, and Dallas. Montreal in Canada and Mexico City in Mexico are also of major importance.

Trade patterns. North American trade patterns offer interesting contrasts. Canada, with a small population but with immense resources and high productivity, has a low home consumption and depends on foreign trade more than any other developed country on the continent. The United States, on the other hand, with a vast internal market and the highest per capita consumption of goods in the world, depends mainly on internal trade; only 8 percent of its total trade is with countries abroad. Mexico and Central America, by contrast, still have large areas where people live at a subsistence level and produce little more than goods for local trade. Production of certain metals, oil, and tropical crops, however, is expanding rapidly for sale in foreign markets.

The Canadian segment. Canada's internal trade is dominated by the provinces of Ontario and Quebec. Together they make 81 percent of all manufactured goods, which they ship across Canada in exchange for fish, lumber, and fruit from British Columbia, wheat and meat from the Prairies, and pulpwood, iron ore, and fish from the Atlantic provinces. Most of Canada's trade abroad consists of raw or semiprocessed materials, including pulp, paper, timber, iron ore, nickel, lead and zinc, uranium, and asbestos, sent to Britain, the United States, and Japan; and wheat, exported to Britain, the Soviet Union, the People's Republic of China, and Japan. Some oil and natural gas are sold to the United States.

Until World War II, Canada traded mainly with Britain; until that time the United States still produced a surplus of most of the things Canada raised and thus was not a major customer. Canada, in fact, bought far more from the United States than it sold to it. By the 1970s, however, the United States had become short of metals, wood, pulp and paper, power, and water and was importing these items from its neighbour on an increasing scale. It has thus replaced Britain as Canada's chief market, taking about 68 percent of Canada's export trade compared with 8 percent by Britain. The European Economic Community and Japan also are growing in importance as customers for Canada's metals, wood products, and wheat.

The United States segment. Internal trade in the United States is enormous, often surpassing that among sovereign states on other continents. It is dominated by New England's need for fuel, cotton and wool, leather, wood products, and metals; by the mid-Atlantic states' demand for coal, oil, natural gas, iron ore and other metals, and food products; by the Pittsburgh region's need for iron, copper, oil, and gas; by the Lower Great Lakes–Lake Michigan area's need for coal, oil, gas, iron, pulp and paper, and wood; and by the Los Angeles–San Francisco regional demand for steel, aluminum, cellulose products, oil, and chemicals.

Most of the other areas of the United States trade their raw materials or semifinished goods to these major manufacturing regions, though of course there are local industrial centres of importance. Trade is concentrated in servicing, or in being served, by such giant metropolitan centres as New York, Chicago, and Los Angeles. These also handle a great deal of America's foreign trade. Almost a third is cleared through the New York–New Jersey area. Southeastern ports send out cotton, tobacco, and wood products; and the mid-Atlantic coast ports send out wheat, corn, meat, and a wide range of manufactured products.

The giant metropolitan centres

With the development of the St. Lawrence Seaway, great cities like Chicago, Detroit, Cleveland, and Buffalo have been exporting directly the steel products, cars, planes, agricultural machinery, cereals, and meat for which the northern Midwest is famous. New Orleans still continues as a notable exporter of cotton, corn, and other agricultural products from the vast Mississippi hinterland; while Houston is the most rapidly expanding southern port, basing its trade on oil and chemical products. Los Angeles dominates the West, with its sales of aircraft, ships, films, and chemicals. Seattle is important for its trade in fish and forest products. American imports include a wide variety of products: tropical fruits, woods, fibres, rubber, and vegetable extracts, mainly from Latin America, West Africa, and Southeast Asia; oil from the Caribbean, Canada, Mexico, Venezuela, and Africa; tin from Bolivia and Malaysia; wool from Australia and South America; and a wide range of machines, textiles, instruments and books from Britain, western Europe, and Japan.

United States trade has a worldwide distribution and impact: of its export total, about a third goes to Britain and western Europe; a fifth to Canada; another fifth to Japan, Southeast Asia, Australia, and New Zealand; a tenth to South America; and a tenth to Mexico, Central America, and the West Indies. Of almost equal importance has been the widespread influence of United States aid: while, initially, this helped United States trade by being tied to the use of American equipment, it is now much more free and enables countries to develop their agriculture or industry in the most satisfactory way.

The Mexican and the Central American segment. The Latin American portion of the continent includes some highly sophisticated regions, along with many as yet undeveloped areas. In Mexico's internal trade the capital region predominates, producing almost four-fifths of the nation's manufactures, which are then distributed through regional cities. Mexico City consumes much of the oil piped up from the coast, the metals of the Cordilleran mines, the cotton of the irrigated central and western basins, and hemp from Yucatán. Mexico's external trade, like that of Canada, is mainly in foods and raw materials. Agricultural exports include maize (corn), meat, hides, winter vegetables and fruits, and cotton and sisal-hemp and account for over two-fifths of total exports. Metals are extensively exported, mainly to the United States; with oil, they make up another fifth of the exports. Manufactured goods, including canned and prepared foods and textiles, are rapidly increasing, amounting to just over a fifth of the export total, and chemicals, wood products, and fish are also significant.

Imports consist predominantly of manufactured goods or else of parts or of materials needed for Mexican exports. Machinery, vehicles, and consumer goods are the chief items. The United States has the greatest share of Mexico's foreign trade, providing five-eighths of the imports and taking two-thirds of the exports. With the establishment of the Latin American Free Trade Association, however, an increasing amount of Mexico's trade has been oriented toward Latin America, which takes only about 8–10 percent of Mexican exports. Mexico is also trying to send more winter fruits and vegetables, textiles, and leather goods to Canada.

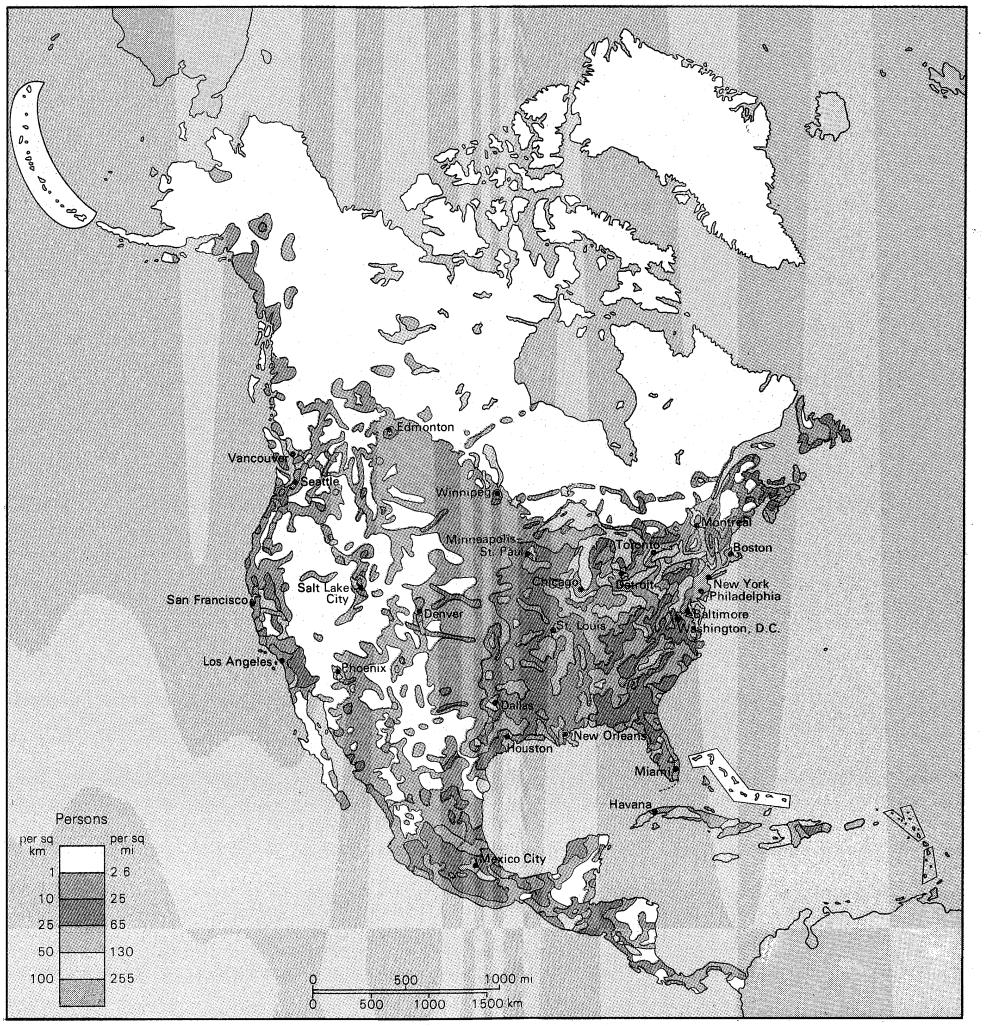
Mexico's trading partners

Central America has not as yet developed much trade. By far the greatest amount is in tropical fruits, fibres, and minerals (especially from the Caribbean), which are sent to the United States in exchange for American manufactured goods. Increasingly, virtually the whole of North America is being integrated in its economic development with the growth of the United States.

PATTERNS OF RESOURCE DEVELOPMENT

Favourable and unfavourable regions. The natural resources of North America and their developing use have led to the emergence of distinct national patterns, which are usually reflected in population distributions. As an area of late exploitation, North America does not have the high settlement densities of Europe and Asia; yet, with great natural wealth and an advanced technology, its population is growing rapidly and is concentrated in regions of comparative advantage, such as the coasts, the lowlands, and the humid and the temperate zones. The 2,500,000 square miles of shield and the 2,000,000 square miles of mountainous land remove nearly a half of the continent from continuous settlement. The frost-ridden areas of the Arctic coast plain, the Mackenzie, and the Hudson Bay lowlands restrict settlement still more. Similarly, the drought conditions in districts of intermontane America and Mexico detract from their use, while the disease-ridden, wet, tropical lowlands of Central America have remained comparatively empty. Further, the empty lands are getting emptier, as people who want to live well and in comfort crowd the already crowded regions. The more populated areas include the Atlantic coast plain from Nova Scotia to Florida; southern Canada in the humid, cool-temperate belt, where there are warm summers; the United States east of the Missouri, in the humid, warm-temperate zone, centred in the vast Mississippi–Ohio plain; the mild, moist Pacific coast from British Columbia south to California; and the temperate, yet warm, and well-watered Mexican Basin. Within these areas of advantage, growth is focussed still further on the great cities, especially those associated with the Hudson–Mohawk gap, the St. Lawrence–Great Lakes system, the Ohio Basin, the Middle Mississippi, Puget Sound and the Fraser Delta, and the California coast.

Geographic distributions. *Canadian share.* Canada's share of the total continental population is small, with



Population density of North America.

22,000,000 people out of a total of about 335,000,000. With an area of 3,560,000 square miles, its overall population density amounts to barely six persons per square mile. Most of Canada—the shield, the northern Appalachians and Cordilleras, and the tundra and boreal forest zones—thus is empty. Population is concentrated in the south, around the Bay of Fundy, the St. Lawrence–Lower Great Lakes, the southern prairies, and the Columbia and Fraser valleys. About 9 percent of the national population is engaged in primary production, nearly 26½ percent in manufacturing and construction, and 64½ percent in service occupations. Wealth is still based on iron and nickel and other metals, oil and natural gas, wheat, meat, fish, and forest products; most of its industries are developed from these resources. Mechanization and automation give a high per capita production, and Canada is only a little behind the United States in general affluence. It raises far more than it can consume and thus sells much abroad.

The makeup of Canada's population

A member of the British Commonwealth, its population is still 40 percent British by extraction and 60 percent English-speaking. It nevertheless grew out of the two founding nations of France and Britain and is 27 percent French by mother tongue. About 19 percent habitually speak French only. Some 30 percent of the people come from northern, southern, or eastern Europe; and most of these are English-speaking. About three-quarters of Canada's population is native-born, but immigration is still active, with rather more non-British (including Americans) than British settling in the country. Canada is thus a bilingual, multicultural nation, with strong attachments to Britain and Europe. It is economically dependent on the United States, which owns over 60 percent of the capital invested there and supplies 70 percent of Canada's imports. The country has a moderate population-growth rate of about 1.7 percent per annum. The most rapidly growing element is the native Indian population, now totalling about 240,000. Ontario and Quebec still lead in attracting population, but Alberta and British Columbia in the west are expanding rapidly. Over three-quarters of the Canadian people live in cities; indeed, about a quarter live in Montreal and Toronto, the leading industrial and business centres.

The role of the United States. The United States dominates the continent with approximately 210,000,000 people or 65 percent of the total continental population. Unlike the other countries, most of its area is habitable; although the higher parts of the Appalachians, the Cordilleran ridges, and the dry intermontane basins are empty. The average density of 59 persons per square mile is much higher than that of Canada, though still low by European or Asian standards. Its most populated parts are the mid-Atlantic states (340 per square mile), New England (167), and the Great Lakes states (148). In the early 1970s, Florida, Arizona, and California were the states that were drawing most of the increase of population. That influx was partly because of the many affluent people retiring in those areas and partly because of the emphasis on growth industries and services. America has gone further in the development of its service occupations than any other New World nation, with 63 percent of its population thus employed. Rather less than 32 percent are in manufacturing, and only 5 percent are in primary production jobs, like farming, fishing, and lumbering. America's consumption of food and goods is the highest per capita in the world.

Occupational structure of the United States

The United States, which long drew its settlers from Europe, by the 1970s had a population only 5 percent

foreign-born. Its population increase comes overwhelmingly from the native-born, among whom the most rapidly expanding groups are the Indians and the Negroes. Only about 30 percent of the black Americans remain in the rural South, most of them living in the Northern cities, predominantly in overcrowded ghetto areas. Washington and Newark are major cities with a black majority. Blacks also account for more than a third of the total population in Cleveland, Detroit, and Chicago and are strongly represented in New York, Pittsburgh, St. Louis, and Los Angeles. The inner-city concentration of blacks, one of the most noteworthy features of urban life in North America, is highlighted by a strong white population shift to the suburbs. About 65 percent of the current United States population growth is, in fact, in metropolitan suburbs, often exacerbating the problems of inner-city decline.

There are four main areas of metropolitan growth known popularly as Boswash (Boston–New York–Philadelphia–Washington) with 40,000,000 people, or 20 percent of the population; Chipitts (Chicago–Detroit–Cleveland–Pittsburgh) with 30,000,000, or 15 percent; Sansan (San Francisco–Los Angeles–San Diego) with 20,000,000 people, or 10 percent; and Jackdal (Jacksonville–Mobile–New Orleans–Houston–Dallas) with about 16,000,000 people, or 8 percent of the total. Fifty-five percent of the American people thus are concentrated within four huge urbanized belts, which comprise less than 10 percent of the country's area.

The role of Mexico, the Caribbean, and Central America. The combined areas of Mexico, the Caribbean, and Central America have a population of almost 100,000,000 (or 50,000,000, 33,000,000, and 17,000,000 people, respectively) and are rapidly growing. Compared with the United States growth rate of 1.2 percent per annum, Mexico is expanding at a rate of almost 3.5 percent per annum. Costa Rica, with a 3.4 percent increase per year, is expected to double its 1,800,000 1970 population within less than a generation. Crude birth rates of from 34 per 1,000 for Costa Rica, to 44 per 1,000 for Honduras compare with the 17 per 1,000 typical of both Canada and the United States. With death rates declining appreciably, a population explosion is under way. Yet, as elsewhere in the developing world, the economy can scarcely support the present population. In Mexico 41 percent of the population is engaged in primary production, and about 21 percent in industry; in Honduras the ratio at the 1961 census was even more unfavourable: 67 percent in primary occupations, and only 9 percent in industry. Great advances in manufacturing and services are needed. Massive United States capital investment and an aid program have been aimed at helping Central America and the Caribbean to help themselves; both factors have been major influences in the region's development, though the influence of their giant northern neighbour has not been without its local critics. Such progress tends to be concentrated in the capital cities. Mexico City has over 8,500,000 people in its metropolitan area and is the largest city between New York and Buenos Aires. Its influence is swiftly modernizing the state.

The population explosion and the economy

THE OUTLOOK

North America is a very dynamic social, demographic, and economic entity, with a geographical position that gave it a commanding role in the world once its huge reserves of metals and fuels and its rich resources of water, soil, and vegetation were developed by its peoples, both native and immigrant. Its peoples, who used their rapidly growing numbers and their varied traditions and skills to bring the continent to the forefront of history, present a formidable challenge to the peoples of the developing world as they, too, begin to seek an equitable share in global resources. The internal problems of the North American continent, some would claim, appear of greatly diminished significance in the face of this wider relationship with the evolving world community.

BIBLIOGRAPHY. General works on North America usually emphasize either a regional or a topical approach. O.P. STARKEY and J.L. ROBINSON, *The Anglo-American Realm* (1969), is an up-to-date regional treatment. G.H. DURY and R.S. MATHIESON, *The United States and Canada* (1970); and J. PATERSON, *North America*, 4th ed. (1970), combine the treatment of topics and regions. J.W. WATSON, *North America*, 2nd ed. (1967), is mainly topical, with a strong historical and human bent.

Geology and physical geography are covered by many texts, the more recent of which are: T.H. CLARK, *The Geological Evolution of North America* (1960), giving a broad picture of the growth of the continent; G.M. KAY (ed.), *North Atlantic: Geology and Continental Drift* (1969), stressing the role of the ocean basins in the grand relief of the Earth, with special reference to North America; C.B. HUNT, *Physiography of the United States* (1967); and H.E. WRIGHT and D.G. FREY, *The Quaternary of the United States* (1965), showing the influence of the last ice age on the landscape.

Biogeography is represented by a collection of works on climate, soils, vegetation, and wild life, since there is no overall text. W.E.D. HALLIDAY, "A Forest Classification for Canada," *Can. For. Serv. Bull. 89* (1950), gives the best account of northern forest types. R.J. PRESTON, *North American Trees* (1961), is a good description of the main trees and their habitats. The *United States Department of Agriculture Yearbooks* are invaluable sources, both from a scientific and economic point of view; see especially those on *Climate and Man* (1941); *Grass* (1945); *Trees* (1949); and *Soils* (1957). R.E. LEGGETT, *Soils in Canada* (1961), is the best account of northern soils.

The resource base is widely covered in governmental and other literature. The Canadian Geological Survey handbook, *Geology and Economic Minerals of Canada*, 4th ed. (1963, revised periodically), gives a good overall view of Canada's mineral wealth; the United States government offers excellent surveys of American resources in the following Information Circulars of the Bureau of Mines: *Supply and Demand for Energy in the United States, by States and by Regions*, pt. 1, *Coal*, pt. 2, *Utility Electricity*, pt. 3, *Dry Natural Gas*, and pt. 4, *Petroleum and Natural Gas Liquids* (1969); and in *The American Land* (1968), a publication of the Soil Conservation Service. Scholarly studies of the general resource situation are: G.W. WILSON, SCOTT GORDON, and STANISLAW JUDEK, *Canada: An Appraisal of Its Needs and Resources* (1965); matched by H. BROWN, *Resource Needs and Demands* (1970). Interesting forecasts of land use and development are given in H.G. BORLAND (ed.), *Our Natural World, the Land and the Wild Life of America As Seen and Described by Writers Since the Country's Discovery* (1969); and H.H. LANDSBERG, *Natural Resources of U.S. Growth: A Look Ahead to the Year 2000* (1964). Of special interest are A.N. LAYCOCK, "Water," in J. WARKENTIN (ed.), *Canada: A Geographical Interpretation* (1968); and A.WOHNAN, *Water Resources: A Report to the Committee on Natural Resources of the National Academy of Sciences-National Research Council* (1962). An important overview of the Latin American resource situation is given in J. GRUNWALD, *Natural Resources in Latin American Development* (1970).

Human resources include race, population, rural and urban development and changes in standards of living. D.J. BOGUE, *The Population of the United States* (1959), is still the standard American work; it should be supplemented by the government publications—*Population Challenge: What It Means to America* (1966), and *200 Million Americans* (1967). On early America, L.A. BRENNAN, *The American Dawn: A New Model of American Prehistory* (1970), gives a good review of theories of early men in North America. H.E. DRIVER, *Indians of North America*, 2nd ed. (1969), is a comprehensive survey of Indian cultures. Case studies of Indians and a summary of U.S. and Canadian relations with the Indians are provided by W.H. OSWALT in *This Land Was Theirs* (1966). Studies of the American Negro are legion, but few are geographical. The UNITED STATES DEPARTMENT OF COMMERCE, BUREAU OF THE CENSUS, *Changing Characteristics of the Negro Population* (1968), offers a useful geographical basis. The ethnic problem in the United States is handled well in M.A. JONES, *American Immigration* (1960). Changing rural trends are discussed in S.R. OGDEN, *America the Vanishing: Rural Life and the Price of Progress* (1969). Urban geography is well represented by J. GOTTMANN and R.A. HARPER, *Metropolis on the Move* (1967); and L.O. STONE, *Urban Development in Canada* (1967). More general statements of metropolitan trends are found in H. BLUMENFELD, *The Modern Metropolis* (1967); and R.A. MOHL and N. BETTEN, *Urban America in Historical Perspective* (1970).

Resource development has developed a vast field of literature. Useful from the geographic point of view are two general works: F.J. MARSCHNER, *Land Use and Its Patterns in the United States* (1959), the classic study of its kind; and

P.R. and A.H. EHRLICH, *Population, Resources, and Environment* (1970), a modern overall assessment. Several studies of transportation are valuable, including W.L. GARRISON, *Highway Development and Geographic Change* (1959); G.P. GLAZEBROOK, *A History of Transportation in Canada* (1964); K.J. KANSKY, *Structure of Transportation Networks* (1963); J.F. STOVER, *The Life and Decline of the American Railroad* (1970); and C.A. TAFF, *Commercial Motor Transportation* (1969). The growing concern with the environment is reflected in R.A. COOLEY and G. WANDESFORDE-SMITH, *Congress and the Environment* (1970); and in the UNITED STATES DEPARTMENT OF AGRICULTURE, *Outdoors USA* (1967); and *The Environmental Decade (Action Proposals for 1970's)* (1970). C.I. JACKSON, *The Spatial Dimensions of Environmental Management in Canada* (1970), is a model study. Mexico and Central America are well represented in two penetrating studies: F. JOHN MATHIS, *Economic Integration in Latin America* (1969); and C.T. NISBET, *Latin America: Problems in Economic Development* (1969).

Development patterns as they affect the geography of North America are set out in three key works: S.B. COHEN (ed.), *Geography and the American Environment* (1968); J. WARKENTIN (ed.), *Canada: A Geographical Interpretation* (1968); and P.E. JAMES, *Latin America,* 4th ed. (1969). A.N.J. DEN HOLLANDER and S. SKARD, *American Civilization* (1968), presents a masterly account of trends in American life. Recent trends in Central America and Mexico are described in L.B. FLETCHER, *Guatemala's Economic Development* (1970); D.E. RAMSETT, *Regional Industrial Development in Central America* (1969); and C.W. REYNOLDS, *The Mexican Economy: Twentieth Century Structure and Growth* (1970).

(J.W.W.)

North American Desert

A vast, irregular belt of inhospitable terrain stretches down the western side of the North American continent, covering 1,000,000 square miles (2,600,000 square kilometres) from southern Canada to northern Mexico and roughly corresponding to the sheltered and hence rain-starved intermontane region lying between the soaring barrier of the Rocky Mountains and the fertile coast ranges fringing the Pacific. The physical geography and human utilization of this huge area exhibit great internal variety, but its overall aridity, associated with an excess of evaporation over precipitation, great temperature extremes, frequent winds (sometimes producing dust storms and dust devils), localized storms, and a predominance of starkly eroded sunbeaten landscapes, which often have a harsh but breathtaking beauty, give it an unquestioned unity. Scientists, in naming this whole ecological complex, or biome, the North American Desert, are merely echoing the legend of a "Great American Desert" established as early as the 1820s by the vivid reportage of an expedition led by the pioneer explorer and engineer Stephen H. Long.

All forms of life, from lowly plants and insects to man himself, have had to struggle to survive in the region, and the North American Desert has thus had enormous importance in the development of the continent. Descendants of the earliest inhabitants, the desert-culture Indians, are still found in the area. Their ranks were swollen in the 19th century by tribes thrust westward in the great dispossession that followed the advance of Anglo-European settlement from the Eastern Seaboard into the continental interior, and the region is now the home—often in severely straitened circumstances—of the bulk of the United States Indian population. The legacy of much earlier population movements from farther south has lent a distinctly Spanish element to the area, while modern American settlement has added its own contribution in the form of sheep and cattle grazing, military installations, and small but often rapidly expanding oases of mining and manufacturing. All the peoples of the North American Desert, whatever their origin, have their lives molded by the basic and all-pervading lack of water.

The legacy of history

This article describes the North American Desert as a biome and also its overall human geography (for individual physical features see BASIN AND RANGE PROVINCE; COLORADO RIVER; DEATH VALLEY; GRAND CANYON; GREAT SALT LAKE; and RIO GRANDE; for a detailed treatment of human geography see the articles on the appropriate constituent states of the United States).

Exploration and scientific study. The Indians of the region accumulated a rich natural lore during the thousands of years of their adaptation to the desert environment, but it was left to Francisco Vázquez de Coronado and other 16th-century Spanish explorers to provide the first written descriptions of the region, particularly of the southwestern portion. The famous Lewis and Clark Expedition of 1804–06 described portions of the northern sector, and Stephen H. Long's pioneer work in the 1820s foreshadowed a host of reports often generated by the huge land grants made to railroads and land companies and written by 19th-century surveyor-engineers. In 1878 the geologist John Wesley Powell made a significant report on the arid west, accurately forecasting the detrimental consequences of imposing on arid regions ways of life more appropriate to humid lands. More diversified studies followed—the first arid-lands research laboratory was founded at Tucson, Arizona, in 1903—and contemporary studies include the important International Biological Program of ecological investigation.

The natural environment. *Regional divisions.* Differences in latitude, elevation, climate, topography, vegetation, soil, and human use allow the subdivision of the North American Desert into regions of cold midlatitude and hot midlatitude desert.

The basic divisions of the desert

The cold midlatitude desert regions cluster in the northern sectors and include the Columbia Plateau (extending across southeastern British Columbia and eastern Washington), the Great Sandy Desert (lying in the Harney Basin of eastern Oregon), nearly all of Nevada (including the volcanic Black Rock Desert), the Snake River Plains of southern Idaho, most of Utah (including the desert associated with the Great Salt Lake), the Painted Desert and the beginnings of the Grand Canyon in the Colorado Plateau west of the Rockies; straddling the Continental Divide, these regions extend from southern Montana into the Red Desert of Wyoming.

Around the southern end of the Sierra Nevada and the scorching landscapes of Death Valley (the continent's lowest point), the Great Basin Desert merges into the hotter midlatitude desert region characterized by the Mojave (Mohave) Desert, bounded on the east by the exotic Colorado River and on the southwest by the extremely hot, barren, and arid Colorado Desert and the sandy Yuma Desert. The last continues into Mexico and merges into the coastal Vizcaino Desert (Desierto de Vizcaíno) of Baja California. To the east, the undulating plains and foothills of the Sonoran Desert blend into the southern Arizona Upland (Saguaro) Desert, thrust further east into the huge Chihuahuan Desert, and stretch over the Mexican states of Chihuahua, Coahuila, and San Luis Potosí. The Sonoran Desert also extends northward into the arid portions of southeastern Arizona; it includes the dazzling gypsum dunes of White Sands National Monument, New Mexico, and crosses the Mexican Plateau and Trans-Pecos of Texas to the Rio Grande.

Landform characteristics. Most of the North American Desert—and all the salty-lake remnants—occupies areas covered by geologically recent (less than 2,500,000 years old) Quaternary deposits and mountains thrown up and folded by movements in Tertiary time (or about twice as old, with some plains and plateaus of Mesozoic sediments, up to 225,000,000 years old).

Because of long periods of erosion, the landforms produced from these rocks are characteristically sharp and angular (except in the very heavily eroded badlands regions) and contain some of North America's finest scenery. The action of wind, temperature changes, ephemeral streams, and floods have all been involved in this molding process. The individual deserts are characterized by plateaus, gorges, ravines, and alluvial fans washed out at the feet of mountains. Deserts of the bolson types contain playas (dried up lake remnants) and mud and salt lakes and flats. Deserts of the hammada type are characterized by extensive rocky surfaces with boulder or gravel coverings, sometimes blackened and wind scoured, with magnificent buttes, mesas, and other isolated mountain remnants rising high above the flat landscape. Stretches of shifting sands known as ergs—the extensive

Algodones Dunes of the Colorado–Yuma desert are a notable example—are found at lower elevations, with the shallow troughs of arroyos carrying intermittent streams from surrounding uplands to be lost in the sands.

Soils. The soils of the North American Desert have origins similar to those of more humid regions, but they are less enriched by organisms and less leached of constituents. Most belong to the "aridisols" dry-soil group, but local variations occur, reflecting differing salt and mineral composition and presence or absence of organic matter. With proper management, the more fertile ones can be quite productive. The rawest soils belong to the less developed group known as "entisols."

Plant and animal life. In the North American Desert, moisture—its quantity, quality, availability, and frequency—is the most critical factor for life. Local environmental factors are also significant in determining the nature of desert plant communities and their dependent animal life. Most desert plants are xerophytes (plants adapted to arid conditions) or phraetophytes (plants dependent on a permanent water supply) and survive only through their root systems and adaptive mechanisms for resisting drought. Sagebrush characterizes the Great Basin region, with Joshua trees, creosote bush, and burroweed typical of the Mojave. The Sonoran Desert has a thorn scrub of shrubs (mesquite, paloverde, ironwood, burrobush, smoke tree, and cat's claw) and a variety of moisture-preserving succulents. The Arizona Upland Desert is noted for the giant saguaro cactus, while the Chihuahuan Desert is characterized, notably in its eastern part, by a ground cover of open mesquite, a scattering of larger trees, and shrubby undergrowth, including the yucca, prickly pear, and other varieties of cactus. Plant life and associated algae, lichens, mosses, and insects become more complex as temperature and moisture conditions improve.

The North American Desert harbours an abundant variety of insects, including grasshoppers that occasionally reach destructive proportions. Lizards, snakes, and other reptiles, the most conspicuous animals, are dependent on plant fluids or devoured animals for moisture. Birds are largely independent of water sources (and are seen almost everywhere), as they live on insects and spiders as well as being preyers and scavengers. Rodents (mice, rats, squirrels, and rabbits) and bats are the most numerous mammals; essentially nocturnal, they remain underground during the heat of the day and, like the birds, obtain moisture from their food. Higher up the food chain are such carnivores as coyotes, bobcats, foxes, and skunks, and the largest desert mammal, found at higher elevations, is the bighorn sheep. Protective coloration, often remarkably complex, is an important feature of desert life.

The desert food chain

The human imprint. As ancient dwellings, rock paintings and carvings, and other archaeological remains testify, desert-culture Indians had developed a distinctive way of life within the approximate boundaries of the North American Desert thousands of years before the coming of the white man. Spanish explorers were the first to penetrate the southwestern area, and their legacy has molded much of the character of the region. It was only in the 19th century that a great wave of settlement, often attracted by the lure of mineral wealth, swept over the whole area on its way to the more fertile coastal regions, leaving a residue of settlement focussed on mineral wealth and irrigated regions and, more sparsely, in the vast areas given over to sheep and cattle grazing. Large areas of the contemporary landscape are occupied by Indian reservations, a legacy of the white man's continental expansion. Military installations, some associated with the testing of nuclear weapons, also take up huge areas. The various types of agriculture encompass dryland farming, sheep and cattle grazing, and more intensive developments on irrigated oases. Mineral exploitation has continued, often to the detriment of the natural environment, and manufacturing industry has become associated with growing urban settlement in the more favoured regions. The tourist trade has also grown immensely. In spite of the increasing development of dams, reservoirs, and canals, the lack of water is still a severe limitation to agricultural, urban, and industrial expansion, and the development of a low-cost water-supply system remains the key to an increased utilization of the entire North American Desert.

BIBLIOGRAPHY

Historical: F.V. COVILLE and D.T. MACDOUGAL, *Desert Botanical Laboratory of the Carnegie Institution* (1903), early plant, soil, and meteorological studies in southwestern U.S. desert areas; D.T. MACDOUGAL, *Botanical Features of North American Deserts* (1908), early desert investigations in the southwestern United States; and D.T. MACDOUGAL, *et al.*, *The Salton Sea: A Study of the Geography, the Geology, the Floristics, and the Ecology of a Desert Basin* (1914); J.W. POWELL, *Report on the Lands of the Arid Region of the United States, with a More Detailed Account of the Lands of Utah*, ed. by W. STEGNER (1962).

General: S. BROWN, "The Great American Desert," in *World of the Desert*, ch. 12 (1963); R. DUNBIER, *The Sonoran Desert: Its Geography, Economy, and People* (1968); E.C. JAEGER, *The North American Deserts* (1957); A. and M. SUTTON, *The Life of the Desert* (1966), an abundantly illustrated natural science volume on the North American desert.

Special studies: W.A. BURNS (ed.), *The Natural History of the Southwest* (1960), illustrated chapters on history, plants, reptiles, birds, and mammals; R.E. CAMERON, G.B. BLANK, and D.R. GENSEL, "Desert Soil Collection at the JPL Soil Science Laboratory," *Tech. Rep. Jet Propulsion Laboratory*, no. 32–977 (1966), an illustrated catalog of arid areas and soil profiles, primarily in the western United States—includes unique high altitude and volcanic areas; C. HODGE and P.C. DUISBERG (eds.), "Aridity and Man: The Challenge of the Arid Lands in the United States," *Publs. Am. Ass. Advmt. Sci.*, no. 74 (1963); INTERNATIONAL BIOLOGICAL PROGRAMME, *Analysis of Ecosystems: Desert Biome Proposal and Research Design* (1969), a summary of specific proposals by 300 scientists to undertake integrated ecosystem studies at selected sites in arid and semi-arid areas of the western and southwestern United States.

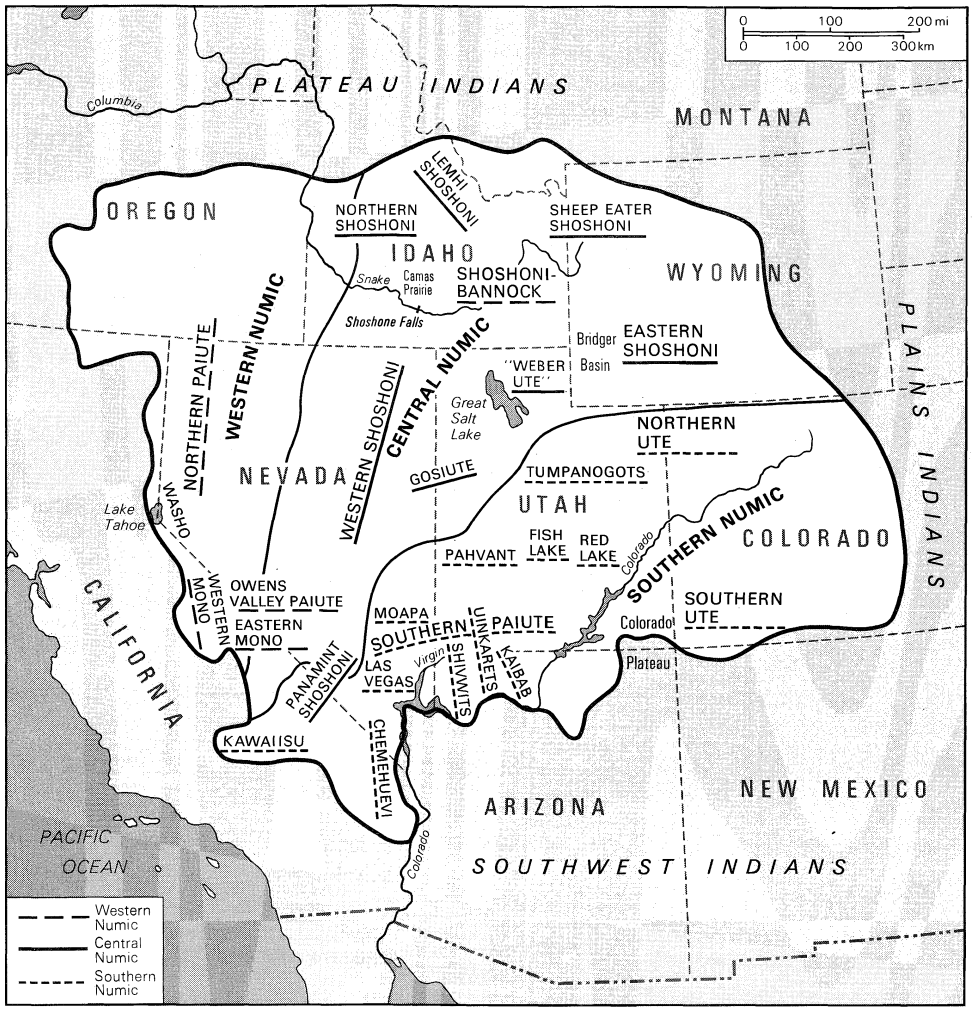
(R.E.C.)

North American Great Basin Indians

The Great Basin Indians in aboriginal times occupied a 398,000-square-mile (1,031,000-square-kilometre) area of interior western North America. The area includes the physiographic Great Basin—the interior mountain and basin region of present-day southeastern California, Nevada, southeastern Oregon, and western Utah—and the Snake River Plain of Idaho, the mountains to the northeast, the Bridger Basin of southwestern Wyoming, the Colorado Plateau area of Utah and Colorado, and the mountains of central and southern Colorado. The entire region is arid to semi-arid, annual average precipitation being four inches in the lowlands to 20–25 inches in the mountains. The precipitation falls primarily in the form of winter snow. Ecologically, the area is characterized by a vertical succession of life zones, each with a dominant xerophytic (desert-type) flora and related fauna.

Aboriginal population density was sparse, ranging from 0.8 to 11.7 persons per 100 square miles.

Peoples and languages. The languages spoken by the Indians are of two widely divergent language families. The Washo, whose territory centred on Lake Tahoe, speak a Hokan language related to languages spoken in parts of California, Arizona, and Baja California. The remainder of the Great Basin culture area was occupied by speakers of Numic languages. Numic, formerly called Plateau Shoshonean, is a division of the Uto-Aztecan language family, a group of related languages widely distributed in the western United States and Mexico. Linguists distinguish three Numic branches, Western, Central, and Southern, each branch having a pair of languages. Western Numic languages are Mono, spoken by the Eastern Mono and Owens Valley Paiute of California, and Northern Paiute, spoken by the several Northern Paiute groups of northeastern California, western Nevada (Paviotso), and southern Oregon and by the Bannock of southern Idaho. Central Numic languages are Panamint, spoken by the Koso, or Panamint Shoshoni, near Death Valley, California; and Shoshone, spoken by the Western Shoshoni of Nevada, the Gosiute of west-



Distribution of Numic languages and major groups of Great Basin area Indians.
From H. Driver *et al.*, *Indiana University Publications in Anthropology and Linguistics* (1953)

ern Utah, the now extinct "Weber Ute" of northern Utah, the Northern Shoshoni of Idaho, the Lemhi and Sheep Eater (Tukuarika) Shoshoni of the northeastern Idaho mountains, the Eastern (or Wind River) Shoshoni of western Wyoming, and the Comanche of the southern Plains. The Comanche separated from the Eastern Shoshoni in late prehistoric times, moved southward through the Rocky Mountains and became Plains Indians culturally. Southern Numic languages are Kawaiisu, spoken by the Kawaiisu band of southern California, and Ute, spoken by the several Southern Paiute bands, including the Chemehuevi of southeastern California and the Las Vegas, Moapa, Kaibab, Shivwits, and Uinkarets bands of southern Nevada, southern Utah, and northern Arizona. Ute is also spoken by the several Ute bands, the Fish Lake, Red Lake, Pahvant, and Tumpanogots of central Utah, the various Northern Ute bands of eastern Utah and the several Southern Ute bands of southern Colorado. The distinction between Southern Paiute and Ute is cultural rather than linguistic: Ute speakers who had horses in the early historic period are regarded as Ute; those without horses were Southern Paiute. The Numic peoples called themselves "Numa" or "Numu," meaning "people" or "human beings." The Washo called themselves "Washoe," meaning "Washo people" as distinguished from people of other tribes.

Linguistic and archaeological evidence indicates that the Washo had long been separated from other California Hokan-speaking groups, possibly for several millennia. Similar evidence indicates that the Numic-speaking peoples spread across the Great Basin from southeastern California sometime after the year 1000.

Social and cultural patterns. Great Basin Indians of the early historic period (1800–50) were divided into horse-using and non-horse-using groups. Horse-using groups generally occupied the northern and eastern sections of the Great Basin culture area. The Southern Ute and Eastern Shoshoni were among the first Indians north of the Spanish settlements of New Mexico to obtain horses, perhaps as early as 1680. There is some evidence that these bands acted as middlemen in the transmission of horses and horse culture from New Mexico to the northern Plains in the 1700s. As the Northern Shoshoni of Idaho obtained horses in the 18th century, they were joined by Northern Paiute speakers from eastern Oregon and northern Nevada to form the Shoshoni-Bannock bands of historic times. By 1800, the Southern and Northern Ute, the Ute of central Utah, the Eastern Shoshoni, the Lemhi Shoshoni, and the Shoshoni-Bannock were well equipped with horses, lived in skin tepees, and were oriented toward the Great Plains, the pursuit of bison, and warfare with other tribes. To the south and west in the Great Basin proper and on the western Colorado Plateau, the people did not take up the use of horses until 1850–60. The Washo did not use horses prior to white settlement, and rarely used them thereafter.

Basic social organization

The basic Great Basin social and cultural patterns were those of the nonhorse bands. The people were closely adapted to their arid environment. Small family bands moved through an annual cycle, exploiting available food resources in the various ecological zones of a particular valley and adjacent mountains. The exigencies of the food quest structured Great Basin society and culture. Food supplies were seldom adequate to permit groups of any size to remain together for more than a few days. Consequently, social organization was fluid and atomistic. For most of the year the people lived in small local groups, coming together into larger aggregates only for

certain brief periods—during rabbit drives or when fish were spawning, as the Washo did at Lake Tahoe in the spring, or during the pinyon nut season in the autumn. But despite periodic gatherings, there was no sustained sense of political cohesion or "tribalness," as that term is understood for other Amerindian groups.

The same fluidity of social organization was characteristic of the horse-using bands. Possession of horses permitted larger numbers of people to remain together for much of the year, but such aggregation did not lead to the development of formal tribal organizations. Among both horse- and non-horse-using groups, a particular leader was followed as long as he was successful in leading people to food, or in war. If he failed, people would leave to join other bands, or to form their own bands.

Kinship, marriage, and rites. The basic local social unit usually was one or more "kin cliques," consisting of a nuclear family (parents and their dependent children) or two brothers and their families, in addition to assorted other individuals related by blood or by marriage to someone in the core group.

Kin ties were reckoned bilaterally through both the mother's and the father's sides and were widely extended to distant relatives. Such extension permitted people to invoke kin ties and move from one group to another if circumstances warranted.

Marriage practices varied, with a tendency among some groups to marry true cross-cousins (mother's brother's or father's sister's child), or pseudo cross-cousins (mother's brother's or father's sister's stepchild). Both the sororate (compulsory marriage of a man to his dead wife's sister) and the levirate (compulsory marriage of a widow to her dead husband's brother), were practiced, as were their logical extensions, sororal polygyny and fraternal polyandry. Usually the latter was not formalized, consisting only of a man extending sex privileges with his wife to a brother for a time. Marriages were brittle and divorce frequent. Yet to survive, it was necessary to be married, as most men and women were throughout their adult lives. There was no set pattern of postmarital residence. A newly married couple might live with the bride's family for the first few years until children were born, but the availability of food supplies was the determining residence factor.

Children began to learn about and participate in the food quest as soon as they were old enough. There was little emphasis on puberty rites except among the Washo, who held a special dance and put a girl through various tests at the time of menarche.

Death rites were minimal. An individual was buried with his possessions or they were destroyed. The Washo abandoned or burned a dwelling in which a death occurred. Occasionally, old people who could not keep up with the group or who could no longer produce their share of the food supply were abandoned.

Technology and economy. The Numic people and the Washo built two types of shelters; semicircular brush windbreaks in the summers, and domed brush, bark-slab, grass, or reed-mat dwellings in the winter. The horse-using groups used Plains-style tepees but sometimes built grass or brush houses. Winter villages were sited along the edge of valley floors, near water, food caches, and firewood. Summer encampments were near food areas and were shifted as necessary. Horse-using groups camped along wooded stream bottoms near firewood and forage areas for their horses.

Tools and food-gathering

Tools were simple and portable: the bow and arrow, stone knife, rabbit stick, digging stick, several types of baskets and nets, and flat seed-grinding slab and handstone. Some Western Shoshoni and Southern Paiute groups made a coarse brown-ware pottery; some Northern Shoshoni made steatite jars and cups. In fishing areas, lines and hooks, harpoons, nets, and willow fish weirs were used. The Northern Paiute used duck decoys made of tule reeds covered with duck skins. Rodents were taken with snares and traps or pulled from burrows with long, hooked sticks. Rabbits were driven into nets and clubbed, or they were shot with bow and arrows. Antelope were driven into corrals and traps. Waterfowl were netted, trapped, or shot with bunt arrows (arrows with rounded heads, intended simply to stun). Deer, elk, and mountain sheep were taken by individual hunters with bow and arrows.

The people followed an annual round, exploiting plant and animal resources as they became available in the several ecological zones. Well over 70 percent of the food supply was vegetal. Over 200 species of plants were named and used, principally seed and root plants. Pinyon pine groves were found in upland areas of Nevada and central Utah, and large quantities of pinyon nuts were collected in the autumn and cached for winter use. Rabbit drives were also held in the autumn. The drives provided an occasion for larger numbers of people to come together for gambling, dancing, and courting. Winter was spent in small villages, living on cached foods and such game as might be taken. Early spring was a poor time; stored resources were often exhausted, and the people were forced to seek early greens and roots for food. Late spring and summer were devoted to collecting seeds, roots, insects, fishing where possible, and continued hunting.

Some Southern Paiute bands practiced limited horticulture along the Colorado and Virgin rivers. Some bands of Mono and Northern Paiute reportedly irrigated patches of wild seed plants to increase the yield.

The horse-using groups also followed an annual round but ranged over a much larger area. In some years, they ventured onto the Northern Plains for bison in the autumn, returning to the Bridger Basin, the Snake River area, or the Colorado Mountains for the winter. In the spring and summer, Shoshoni and Shoshoni-Bannock obtained roots from the Camas Prairie in Idaho and salmon from the Snake River, below Shoshone Falls. Deer, elk, and mountain sheep were taken when possible. Seed and root foods were collected as they became available.

Clothing consisted of sage bark aprons and breechclouts and rabbit-skin robes in the winter. The horse-using peoples wore Plains-style, tailored skin garments. Artwork was largely confined to basketry decoration. Among the horse-using bands, quill and beadwork decorated clothing and rawhide shields, and bags and containers were painted.

Trade was minimal among western Great Basin groups, although there is some evidence of the use of strings of shells as a medium of exchange in aboriginal times. Horse-using groups were more active, trading among themselves and with other tribes. The Eastern Shoshoni and some Ute bands participated in the fur trade between 1810 and 1840. Between about 1800 and 1850, mounted Ute and Navaho bands preyed on Southern Paiute, Western Shoshoni, and Gosiute bands for slaves, capturing and sometimes trading women and children to be sold in the Spanish settlements of New Mexico and southern California.

Religious concepts. Religious concepts derived from a mythical cosmogony, beliefs in "power" beings, and a belief in a dualistic soul. Mythology provided a cosmogony and cosmography of the world. Mythical animals, notably wolf, coyote, rabbit, bear, and mountain lion, were believed to be the progenitors of the modern animals. They lived prior to Indian life but were anthropomorphic, speaking and acting as people do in the present world. They created the world and were responsible for present-day topography, ecology, food resources, seasons of the year, and the distribution of Indian tribes. They set the nature of social relations—that is, defined how various classes of kinsmen should behave toward each other—and set the customs surrounding birth, marriage, puberty and death. Their actions in the mythic realm set moral and ethical precepts and determined the physical and behavioral characteristics of the modern animals. Most of the motifs and tale plots of Great Basin mythology are found widely throughout North America.

Power beings were animals, birds, or natural phenomena, each attributed with a specific natural power according to an observed characteristic. Some such beings were thought to be benevolent, or at least neutral, toward men. Others, such as Water Babies—small, long-haired creatures who lured men to their deaths in springs or

lakes and who ate children—were malevolent and feared. There were conceptions of various other vague beings, such as the Southern Paiute *unupits,* mischievous spirits who caused illness.

Shamanism

Shamans, or curers, were prominent in all Great Basin groups. Both men and women might become shamans. Shamans received their powers to cure disease, foretell the future, and, sometimes, to practice sorcery from a power being who came unsought to a prospective shaman. It was considered dangerous to resist being given a shaman's power, for those who did sometimes died. The being became a tutelary spirit, instructing an individual in curing and sources of power. Some shamans had several tutelary spirits, each providing instruction for specific types of treatment. Among Northern Paiute and Washo and probably elsewhere, a man who had received power apprenticed himself to an older, practicing shaman and from him learned rituals, cures, and feats of legerdemain associated with curing performances. Curing ceremonies were performed with family members and others present and might last several days. The widespread Amerindian practice of sucking an object said to cause the disease from the patient's body was often employed. Shamans who lost too many patients were sometimes killed.

In the western Basin, some men had powers to charm antelope and led communal antelope drives. Beliefs that some men were arrow-proof (and after the introduction of guns, bulletproof) are reported for the Northern Paiute and Gosiute but were probably general throughout the area.

Among the Eastern Shoshoni, young men sought power beings through a visionary experience. The active seeking of power beings through visions is a practice the Eastern Shoshoni probably learned from their Plains neighbours, although the characteristics of the beings sought were those common to Great Basin beliefs.

There was a concept of soul-dualism among most, if not all, Numic groups. One soul, or soul aspect, represented vitality or life; the other was the individual as he was in a dream or vision state. During dreams or visions, the latter soul left the body and moved in the spirit realm. At death, both souls left the body.

Impact of white settlement. Contact with white civilization drastically altered Great Basin societies and cultures. The Southern Ute were in sustained contact with the Spanish in New Mexico as early as the 1600s, but other Great Basin groups had no direct or continued contact with whites until after 1800. The fur trade, between 1810 and 1840, brought new tools and implements to the eastern bands. Settlement began in the 1840s, as did the surge of emigrants through the area on their way to Oregon and California. Mining, ranching, and farming activities destroyed or closed off traditional Indian food-gathering areas. Pinyon groves were cut for firewood, fence posts, and mining timbers. The Indians attempted to resist white encroachment. Mounted bands of Ute, Shoshoni, Shoshoni-Bannock, and Northern Paiute preyed on ranches and wagon trains and tried to drive the intruders away. The struggle culminated in several local "wars" and "massacres" in the 1850s and 1860s. After 1870, Indians were forced onto reservations or into small groups on the edges of white settlements, thus reducing their land base to a small fraction of its former size. This forced the abandonment of aboriginal subsistence patterns in favour of limited agriculture or stock raising, where possible, and wage work, especially as farm and ranch hands.

In 1870 and again in 1890, so-called ghost dances started among the Northern Paiute of western Nevada. The dances were millenarian in character. Prophets foretold that if the Indians danced and prayed, the whites would go away and the "old days" would be restored. The 1870 dance, led by a man named Wodziwob, centred in Nevada and California. The 1890 dance, led by Wovoka, or Jack Wilson, of Smith Valley, Nevada, spread to many Indian tribes in the western United States.

A Peyote Cult was introduced to the Ute and Eastern Shoshoni in the early 1900s by Oklahoma Indians. It later spread to other Great Basin peoples. Most peyote groups are now members of the Native American Church, a nationally recognized organization. Great Basin peyote rituals are a mixture of aboriginal and Christian elements. Ceremonies are led by "road chiefs"; that is, those who lead believers down the Peyote Road or Way. A ceremony, which lasts all night, includes singing, praying, and eating peyote buttons or drinking a concoction made therefrom, producing a mild hallucinogenic experience. The tenets of the Native American Church stress moral and ethical precepts and behaviour.

In postreservation times, the Eastern Shoshoni and Ute adopted the sun dance from the Plains Indians. The four-day dance is performed yearly to achieve health and valour for the participants, and partly as a tourist attraction.

Present-day reservations

Present-day reservations and the people living on them are: Wind River, Wyoming (Eastern and Sheep Eater Shoshoni); Fort Hall, Idaho (Shoshoni-Bannock and Lemhi Shoshoni); Uintah and Ouray, Utah (Northern Ute); Kaibab, Arizona (Southern Paiute); Skull Valley and Goshute, Utah (Gosiute); Moapa River, Nevada (Southern Paiute); Ruby Valley, Te-moak, Duckwater, and Yomba, Nevada (Western Shoshoni); Duck Valley, Nevada–Idaho (Western Shoshoni and Northern Paiute); Fort McDermitt, Nevada–Oregon (Western Shoshoni and Northern Paiute); Burns and Warm Springs, Oregon (Northern Paiute); Summit Lake, Pyramid Lake, Walker River and Stillwater, Nevada (Northern Paiute); and Fort Bidwell, California (Northern Paiute). There are also over 20 "colonies" of 25 to 200 people adjacent to cities and towns in Nevada, eastern California, and southern Utah. The Washo do not have a reservation as such but live in the Reno and Carson City, Nevada, colonies as well as at Dresslerville, Nevada, and Woodfords, California.

The Indian Reorganization Act of 1934 led to the establishment of local elected "tribal councils" for the various reservations and colonies. Councils have sought to develop various economic activities including ranching, light industry, and tourism.

Indian children were sent to federal day schools and boarding schools beginning in the 1880s. In the past few years, federal schools have been phased out and Indian children attend local schools and universities.

Great Basin Indian peoples retain some of their traditional culture in crafts, dances, and visiting patterns. Older people still speak the native languages, but many of the younger people fail to learn or use them. Many people still live on the reservations, but others have moved to towns and cities, where they are employed in many different capacities.

BIBLIOGRAPHY. There is no general monograph on all Great Basin Indians. CATHERINE S. FOWLER, *Great Basin Anthropology: A Bibliography* (1970), lists some 6,500 sources on the area. Summary articles on various aspects of Great Basin anthropology are contained in W.L. D'AZEVEDO *et al.* (eds.), *The Current Status of Anthropological Research in the Great Basin: 1964* (1966); and in E.H. SWANSON, JR. (ed.), *Languages and Cultures of Western North America* (1970). The earliest systematic study of Great Basin Indians was by JOHN WESLEY POWELL; see D.D. and C.S. FOWLER (eds.), *Anthropology of the Numa: John Wesley Powell's Manuscripts on the Numic Peoples of Western North America, 1868–1880* (1971). Modern ethnographic studies include J.H. STEWARD, *Basin-Plateau Aboriginal Sociopolitical Groups* (1938); R.H. LOWIE, *Notes on Shoshonean Ethnography* (1924); R.F. and Y. MURPHY, *Shoshone-Bannock Subsistence and Society* (1960); V.C. TRENHOLM and M. CARLEY, *The Shoshonis: Sentinals of the Rockies* (1964); JAMES F. DOWNS, *The Two Worlds of the Washo, an Indian Tribe of California and Nevada* (1966); and I.T. KELLY, *Southern Paiute Ethnography* (1964). Religious beliefs are treated by W.Z. PARK, *Shamanism in Western North America* (1938); and B.B. WHITING, *Paiute Sorcery* (1950). Important linguistic studies include E. SAPIR, *Southern Paiute Texts and Dictionary* (1930–31); W.H. JACOBSEN, JR., "Washo Linguistic Studies," and W.R. MILLER, "Anthropological Linguistics in the Great Basin," both in W.L. D'AZEVEDO (*op. cit.*). Great Basin prehistory is summarized in J.D. JENNINGS "The Desert West" in J.D. JENNINGS and E. NORBECK (eds.), *Prehistoric Man in the New World,* pp. 149–174 (1964).

(D.D.F./C.S.F.)

North American Indian Languages

The term North American Indian languages usually refers to those languages that are indigenous to the United States and Canada, and that are spoken north of the Mexican border. A number of language groups within this area, however, extend as far south as Central America. The present article will concentrate on the languages of Canada and the United States. (For further information on languages of Mexico and Central America, see MESO-AMERICAN INDIAN LANGUAGES).

The Indian languages of North America are both numerous and diverse. Their original number has been estimated at 300; these tongues were spoken by a native population of approximately 1,500,000. The number of languages still used was estimated at around 200 by the American linguist Wallace Chafe in 1962. Some of these had only one or two elderly speakers. The numbers continue to drop, but with some notable exceptions—*e.g.*, Navajo is steadily increasing in number of speakers. As a consequence of the growing trend toward extinction in the American Indian languages, the field of study is becoming more concerned with the past than the future. Even so, the rich diversity of these languages provides a valuable laboratory for linguistic theory; certainly the discipline of linguistics could not have developed as it has, especially in the United States, without the native American languages. In this article, the present tense will be used in referring to both extinct and surviving languages.

Within the diversity of the North American Indian languages, no general characterization is possible; various features of structure are common to them, but there is no feature or complex of features shared by all. At the same time, there is nothing primitive about these languages. They draw upon the same linguistic resources and display the same regularities and complexities as do the languages of Europe. If historical connections are sought among the Indian tongues, some languages clearly show numerous and systematic resemblances comparable to those between Spanish, French, and Italian. These similarities strongly suggest classification as a linguistic family. North American languages can then be grouped into some 57 families. On this level, too, the diversity of some areas is notable. Thirty-seven families lie west of the Rockies and 20 in California alone; California thus shows more linguistic variety than all of Europe. Some families seem to be related to each other in more remote historical groupings, often called phyla. Such classifications border on speculation, however, partly because data are lacking on many languages (because they are extinct or still unstudied), and partly because of the difficulty in distinguishing, at the deeper historical levels, between resemblances caused by common origin and those resulting from linguistic borrowing.

Language families and phyla

In any case, no theory of common origin for the North American languages has become established. Although most anthropologists believe that North America was populated mainly by people who migrated across the Bering Strait from Asia, attempts to relate native American languages to Asian languages have not gained general acceptance. (There is one possible exception—the relationship of Eskimo-Aleut to certain Siberian languages.) The linguistic diversity of North America suggests, indeed, that the area was populated as a result of several waves of migration by peoples of distinct linguistic stocks of Asia; these stocks may have no modern survivors.

Classification. The first comprehensive classification into families of the North American Indian languages was made in 1891 by the American John Wesley Powell, who based his study on impressionistic resemblances in vocabulary. A principle of nomenclature adopted by Powell has been widely used ever since: families are named by adding *-an* to the name of one prominent member; *e.g.*, Caddoan is the family including Caddo and other languages. For this most obvious level of relationship, the Powell classification remains essentially unchallenged. Various scholars, however, have attempted to group the families into larger units that reflect deeper levels of historical relationship. Of these efforts, one of the most ambitious and best known is that of Edward Sapir, which was first published in the *Encyclopaedia Britannica* in 1929. In Sapir's classification, all the languages are grouped into six phyla—Eskimo-Aleut, Algonkian-Wakashan, Na-Dené, Penutian, Hokan-Siouan, and Aztec-Tanoan—established on the basis of very general grammatical resemblances. In 1958, research of the American linguist Mary R. Haas revealed precise sound correspondences between the Algonkian languages and a "Gulf" group in the southeastern United States that Sapir had assigned to the Hokan-Siouan phylum. Since that time, various reconsiderations of Sapir's groupings have been proposed. A classificatory map published by Charles F. and Florence M. Voegelin in 1966 offers one such classification, and it is likely to serve as a standard reference point for some time. Although preserving Sapir's Eskimo-Aleut, Na-Dené, Penutian, and Aztec-Tanoan groups, it also proposes reconstituted Macro-Algonkian, Macro-Siouan, and Hokan phyla, and allows nine families to remain unclassified, pending further research.

Sapir's six phyla

The Table, based on the Voegelin map, gives approximate indications of the aboriginal home territories and of the number of speakers remaining in 1962 as estimated by Chafe (see also ESKIMO-ALEUT LANGUAGES; MESO-AMERICAN INDIAN LANGUAGES).

Language contact. The Indian languages of North America, like all languages in the world, have always existed in contact with other tongues. From this situation bilingualism, or multilingualism, has resulted; the extent is determined by sociological factors. The Indian languages show varying degrees of linguistic acculturation; *i.e.*, there may be borrowing between languages not only of vocabulary items, but also of phonological, grammatical, and semantic features. In aboriginal times, in areas where bilingualism was most important (*e.g.*, the Northwest), there tended to be well-defined linguistic areas in which languages of diverse genetic affiliations came to share numerous structural characteristics through the process of borrowing. As noted above, such phenomena create difficulties for attempts at genetic classifications. In a few cases, situations of language contact have given rise to a pidgin or compromise language that is composed of elements from various sources and is used as a second language, especially in trading. An example is the Chinook Jargon of the Northwest; this came to be used by many whites and absorbed many loanwords from French and English before its eventual obsolescence.

Pidgin languages

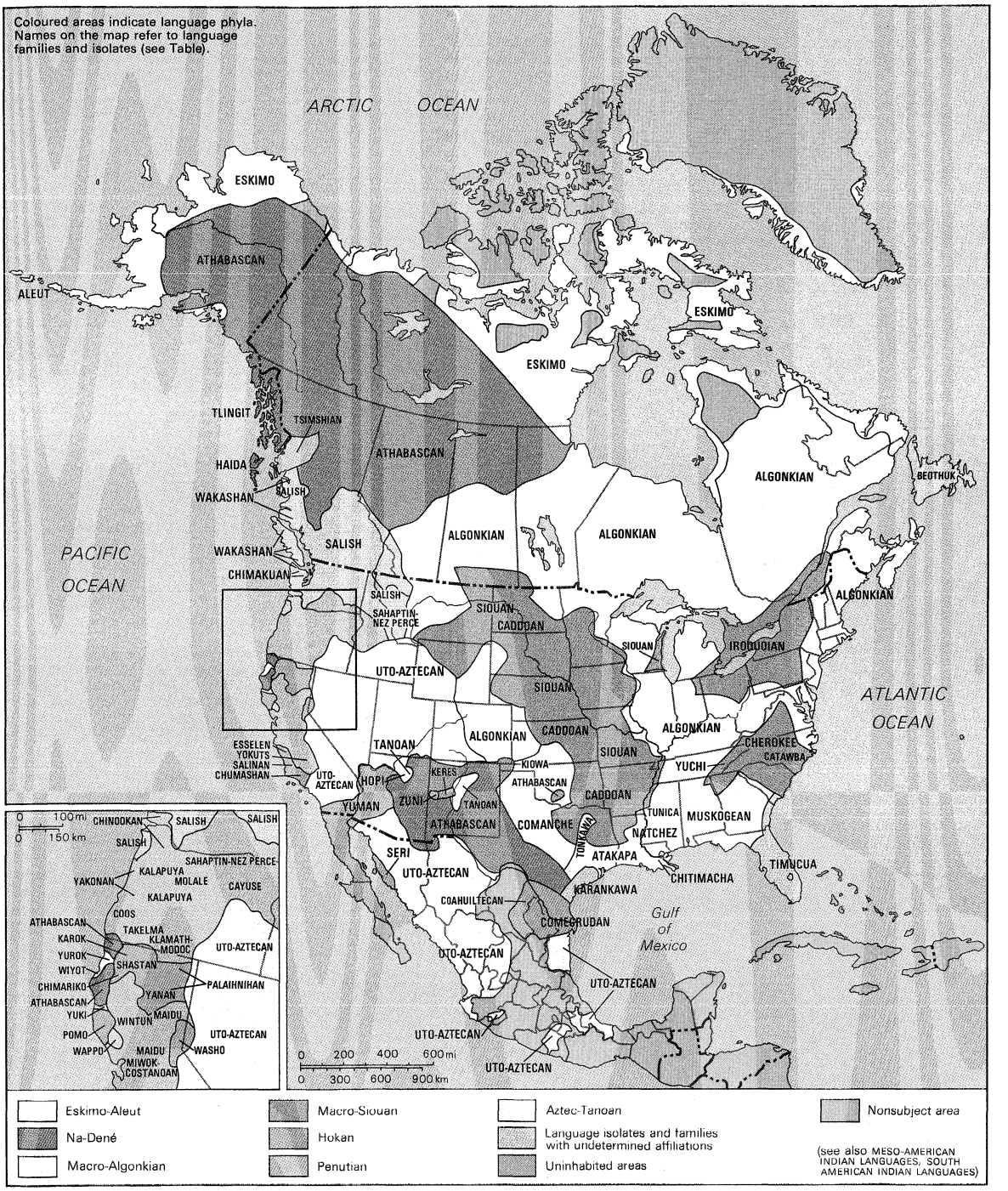
In more recent times, contact of Indian languages with European languages—French, English, Spanish, and Russian—has again resulted in bilingualism. With the Indian languages generally relegated to a socially subordinate position (and with many of them headed for extinction), borrowing, however, has involved the relatively superficial level of vocabulary more often than the deeper levels of language structure, such as the sound system or grammar. The effects on European languages are apparent mainly in place names like Massachusetts and Seattle and in names like squash and abalone for native American plants and animals. Among the Indians, the type and degree of linguistic adaptation to European culture has varied greatly, depending on sociocultural factors. For example, among the Karok of northwestern California, a tribe that suffered harsh treatment at the hands of whites, there are only a few loanwords from English (*e.g.*, *ápus* "apples"), a few calques or loan translations (the "pear" is called *vírusur* "bear," because English "pear" and "bear" are merged in Karok pronunciation), but a large number of new formations from native materials; *e.g.*, a hotel is called *am-naam* "eating place."

Grammar. The term grammatical structure as used here refers to both the traditional categories of morphology—how words are made up—and syntax—how words are combined into sentences. It should again be emphasized that in grammar, as well as in phonological or semantic structure, neither the American Indian languages nor any other languages in the world display anything that could be called primitive in the sense of undeveloped or rudimentary. Every language has a structure

North American Indian Languages*

phyla, families, languages	location	speakers remaining
American Arctic-Paleosiberian		
Eskimo-Aleut		
Chukchi-Kamchatkan (in Siberia)		
Na-Dené		
Athabascan		
Dogrib, Bear Lake, Hare	N.W.T.	1,400
Chipewyan, Slave, Yellowknife	N.W.T.	4,400+
Kutchin	Yukon, Alaska	1,200
Tanana, Koyukon, Han, Tutchone	Alaska	1,800+
Sekani, Beaver, Sarsi	Alberta	450+
Carrier, Chilcotin	B.C.	1,500+
Tahltan, Kaska	N.W.T.	300+
Tanaina, Ingalik, Nabesna, Ahtena	Alaska	1,400+
Eyak	SC Alaska	†
Chasta Costa, Galice, Tututni	SW Oregon	†
Hupa	NW Calif.	50
Kato, Wailaki	NC Calif.	†
Mattole	NC Calif.	‡
Tolowa	NW Calif.	†
Navajo	Ariz., N.M.	80,000+
Western Apache	W Arizona	8,000+
Chiricahua, Mescalero Apache	S New Mexico	1,100+
Jicarilla Apache	N New Mexico	1,000
Lipan Apache	Texas	†
Kiowa Apache	Oklahoma	†
Tlingit	SE Alaska	1,000+
Haida	B.C.	700
Macro-Algonkian		
Algonkian		
Cree, Naskapi, Montagnais	E Canada	35,000
Menominee	Great Lakes area	300+
Fox-Sauk-Kickapoo	south of Great Lakes	1,500
Shawnee	SC U.S.	300+
Potawatomi	Michigan	100+
Ojibwa, Ottawa, Algonkin, Salteaux	S Ontario	41,000+
Delaware	C Atlantic coast	10+
Penobscot, Abnaki	New England	50+
Malecite, Passamaquoddy	New England, Maritime Provinces	900+
Micmac	Maritime Provinces	3,000+
Blackfoot	Mont., Alberta	5,000+
Cheyenne	E Wyoming	3,000+
Arapano, Atsina, Nawathinehena	E Colorado	1,000+
Yurok	NW Calif.	10+
Wiyot	NW Calif.	1
Muskogean		
Choctaw, Chickasaw	N Mississippi	12,000+
Alabama, Koasati	Alabama	300+
Mikasuki, Hitchiti	NW Florida	700
Muskogee (Creek), Seminole	Georgia	10,300+
Natchez	N Louisiana	‡
Atakapa	SW Louisiana, SE Texas	‡
Chitimacha	S Louisiana	‡
Tunica	N Louisiana	‡
Tonkawa	E Texas	†
Macro-Siouan		
Siouan		
Crow	E Montana	3,000
Hidatsa	North Dakota	500+
Winnebago	Wisconsin	1,000+
Omaha, Osage, Ponca, Kansa, Quapaw	central plains	1,200+
Dakota (Sioux)	northern plains	15,000+
Tutelo, Ofo, Biloxi	Gulf Coast	‡
Catawba	Carolinas	‡
Iroquoian		
Seneca, Cayuga, Onandaga	New York	2,600+
Mohawk	New York	1,000+
Oneida	New York	1,000+
Wyandot (Huron)	SE Ontario	‡
Tuscarora	North Carolina	100+
Cherokee	southern Appalachians, Oklahoma	10,000
Caddoan		
Caddo	Arkansas	300+
Wichita	Texas, Okla.	100+
Pawnee	Kansas	400+
Arikara	Dakotas	200+
Yuchi	southern Appalachians	100+
Hokan		
Yuman		
Walapai, Havasupai, Yavapai	NW Arizona	900+
Mohave, Yuma	lower Colorado River	2,000
Delta Yuman (Cocopa)	delta of Colorado River	300+
Diegueño, Kiliwa	S Calif., Baja Calif.	10+
Seri	Sonora	200
Pomo	NC Calif.	
Northern Pomo		30
Northeast Pomo		1
Central Pomo		30
Southwest Pomo		50
Southeast Pomo		†
Southern Pomo		30
Palaihnihan	NE Calif.	
Achomawi		10+
Atsugewi		†
Shastan	NE Calif.	†
Yanan	NC Calif.	‡
Chimariko	NW Calif.	‡
Washo	EC Calif., Nevada	100
Salinan	WC Calif.	†
Karok	NW Calif.	100+
Chumashan	S Calif.	1
Comecrudan	S Texas, NW Mexico	‡
Coahuiltecan	S Texas, NW Mexico	‡
Esselen	WC Calif.	‡
Branches in Meso-America		
Penutian		
Yokutsan	SC Calif.	10+
Maiduan	NC Calif.	10+
Wintun	NC Calif.	
Patwin		10+
Wintu, Nomlaki		20+
Miwok-Costanoan		
Sierra Miwok	SC Calif.	50
Coast-Lake Miwok	C Calif.	†
Costanoan	WC Calif.	‡
Klamath-Modoc	SC Oregon	100+
Sahaptian		
Sahaptin (Klikitat, Umatilla, Walla Walla, Warm Springs, Yakima)	NC Oregon	1,400+
Nez Perce	WC Idaho	500+
Cayuse	NE Oregon	‡
Molale	NC Oregon	‡
Coos	SW Oregon	†
Yakonan	WC Oregon	
Alsea		‡
Siuslaw, Lower Umpqua		‡
Takelma	SW Oregon	‡
Kalapuya	WC Oregon	†
Chinookan	NW Oregon, SW Wash.	20
Tsimshian	WC B.C.	3,000
Zuni	WC New Mexico	3,000+
Latin American branches		
Aztec-Tanoan		
Kiowa-Tanoan		
Tiwa	NC New Mexico	2,200+
Tewa	NC New Mexico	2,000
Towa	NC New Mexico	1,200
Kiowa	Oklahoma	2,000
Uto-Aztecan		
Mono	EC Calif.	100+
Northern Paiute (Paviotso), Bannock, Snake	NE Calif., SE Ore., N Nev., S Idaho	2,000
Panamint, Gosiute, Shoshone	C Nev., N Utah, SW Wyo.	5,000+
Comanche	N Texas	1,500
Kawaiisu, Ute, Chemehuevi, Southern Paiute	SE Calif., S Nevada, S Utah, SW Colo.	3,000+
Hopi	N Arizona	3,000+
Tubatulabal	SC Calif.	†
Luiseño	S Calif.	100+
Cahuilla	S Calif.	10+
Cupeño	S Calif.	†
Serrano	S Calif.	†
Pima-Papago	S Arizona	13,000+
Latin American branches		
Unclassified		
Keresan	New Mexico	7,000
Yukian	NC Calif.	
Yuki		†
Wappo		†
Beothuk	Newfoundland	‡
Kutenai	Mont., Idaho, B.C.	300+
Karankawa	SE Texas	‡
Chimakuan	NW Washington	
Quileute		10+
Chemakum		‡
Salish		
Lilloet	C B.C.	1,000+
Shuswap	E B.C.	1,000+
Thompson	C B.C.	1,000+
Okanagon, Sanpoil, Lake, Colville	S B.C.	1,000+
Pend d'Oreille, Flathead, Spokan, Kalispel	N Idaho	600+
Coeur d'Alene	N Idaho	100
Middle Columbia, Wenatchee	E Washington	200
Tillamook	NW Oregon	1
Twana	NW Washington	†
Upper Chehalis, Cowlitz, Lower Chehalis, Quinault	W Washington	10+
Southern Puget Sound Salish	W Washington	50+
Straits Salish	SW B.C.	500
Halkomelem	SW B.C.	1,000+
Squamish	SW B.C.	100+
Comox, Sishistl	Vancouver Is.	†
Bella Coola	WC B.C.	200+
Wakashan		
Nootka	Vancouver Is.	1,000+
Nitinat	Vancouver Is.	10+
Makah	NW Washington	500
Kwakiutl	WC B.C.	1,000
Bella Bella, Heiltsuk	WC B.C.	100+
Kitamat, Haisla	WC B.C.	100+
Timucua	Florida	‡

*Phyla given in boldface type; families given in italics (including those consisting of single languages); single languages, or dialect groups so closely related that they can be treated as single languages, given in roman type. †Minimal number of speakers; *i.e.*, under 10. ‡Extinct.



Distribution of North American Indian languages.

From C.F. and F.M. Voegelin, *Map of North American Indian Languages;* copyright 1966 by University of Washington Press

as complex, as subtle, and as efficiently adaptable to cultural needs as that of Latin or English, for example.

The North American Indian languages display great diversity, so that it is not possible to characterize them as a group by the presence or absence of any particular grammatical peculiarities. At the same time, there are some characteristics that, though not unknown elsewhere in the world, are sufficiently widespread to be considered typical of the continent or of particular linguistic areas within North America. The phenomenon of polysynthesis, in which many sentence elements are expressed within the boundaries of a single word by compounding and affixation, is especially characteristic of Eskimo and Algonkian, but is also found elsewhere. An illustration from the Algonkian group is the Menominee form *nekees-pesteh-wenah-nɛɛwaaw* "but I did see him on the way." Incorporation, the compounding of a noun with a verb, is rarely used in English (*e.g.*, "to baby-sit") but is common in some Indian languages; *e.g.*, Mohawk *ke-wēna-weiẽhō* "I-language-understand." (The symbols used that are not found in the Latin alphabet have been adopted from phonetic alphabets.)

Polysynthesis and incorporation

Some especially common characteristics of North American languages are the following:

1. In verbs, the person and number of the subject are commonly marked by prefixes; *e.g.*, Karok has *ni-'áhoo* "I walk," *nu-'áhoo* "he walks." In some languages, the prefix simultaneously indicates the object as well as subject; *e.g.*, Karok *ni-mmah* "I see him," *ná-mmah* "he sees me."

2. Tense and aspect of verbs are usually marked by suffixes, as in many languages throughout the world. But in some areas—*e.g.*, among the Athabascan languages—prefixes are used. For example, Chipewyan *hɛ-tsaɣ* means

"he is crying," *ɣĩ-tsaɣ* is "he cried," and *ɣwa-tsaɣ* is "he will cry."

3. In noun forms, the concept of possession is widely expressed by prefixes indicating the person and number of the possessor. Thus Karok has *ávaha* "food," *nani-ávaha* "my food," *mu-ávaha* "his food," etc. When the possessor is a noun, as in "man's food," a construction like *ávansa mu-ávaha* "man his-food" is used. Many languages have inalienable nouns, which cannot occur except in such possessed forms. These generally designate such things as kinsmen or body parts; *e.g.*, Luiseño, a language in Southern California, has *no-yó'* "my mother," *o-yó'* "your mother," but no word for "mother" in isolation.

4. Nouns in many languages have forms with a meaning of location; *e.g.*, Karok *áas* "water," *áas-ak* "in the water." Such a construction is reminiscent of the case forms of Latin, and case systems do indeed occur in California and the southwest. For example, Luiseño has the nominative *kíiča* "house," accusative *kíiš*, dative *kíi-k* "to the house," ablative *kíi-ŋay* "from the house," locative *kíi-ŋa* "in the house," instrumental *kíi-tal* "by means of the house."

The following five grammatical features are less typically North American, but are nevertheless distinctive of many areas. First person pronouns in many languages show a distinction between a form inclusive of the addressee—"we" denoting "you and I"—and an exclusive form —"I and someone other than you." Some languages also have a distinction in number between singular, dual, and plural pronouns. Reduplication, the repetition of all or part of a stem, is widely used to indicate distributed or repeated action of verbs; *e.g.*, in Karok, *imyah* means "breathe," *imyáhyah* means "pant." In Uto-Aztecan languages, reduplication sometimes is associated with plural nouns, as in Pima *gogs* "dog," *go-gogs* "dogs." In many languages, verb stems are distinguished on the basis of the shape or other physical characteristics of the associated noun; thus in Navajo, in referring to motion, *'á*^{n} is used for round objects, *tá*^{n} for long objects, *tí*^{n} for living things, *lá* for ropelike objects, etc. Similar distinctions may refer to dual and plural number. Karok has *ikpuh* "one swims," *iθpuh* "two swim," *ihtak* "several swim."

Inclusive and exclusive pronouns

Verb forms also frequently specify the location or direction of an action by the use of prefixes or suffixes. In Karok, for example, from *paθ* "throw" is derived *páaθ-roov* "throw upriver," *páaθ-raa* "throw uphill," *paaθ-rípaa* "throw across-stream," and as many as 38 other similar forms. Some languages also specify the instrument of an action, generally by prefixation; *e.g.*, Pomo *phi-de-* "to move by batting with a stick," *phu-de-* "to move by blowing," *pha-de-* "to move by pushing with the end of a stick." Lastly, many languages have evidential forms of verbs that indicate the type of validity of the information reported; such distinctions may assume the importance played by tense and aspect in European languages. Thus Hopi distinguishes *wari* "he ran, runs, is running" as a reported event, from *warikŋwe* "he runs (*e.g.*, on the track team)," which is a statement of general truth, and from *warikni* "he will run," which is an anticipated event. In other languages verb forms consistently discriminate hearsay from eye-witness reports. Such a system might be very welcome in other societies; *e.g.*, especially as regards the reliability of news reports.

Phonology. The languages of North America are as diverse in their systems of pronunciation as they are in other ways. In terms of the number of contrasting sounds (phonemes), the Northwest Coast is characterized as a linguistic area by the unusual richness of its systems. A language like Tlingit has approximately 50 consonants and vowels (a comparable count for English would number 35). By contrast, Karok has only 23. The richest sound inventories seem to occur where bilingualism was commonest, and sounds were borrowed between languages.

Consonant features

The large number of consonants that is found in many Indian languages is based on the use of a number of phonetic contrasts that are relatively unfamiliar in European languages. In English, different consonants are produced by vibrating the vocal cords (which results in voiced sounds) or by not vibrating them (which gives unvoiced sounds); by shutting off the air momentarily, thus producing stops, or by letting the airstream pass through the mouth with friction (producing fricatives); and by placing the tongue in a variety of positions. The Indian languages also use these mechanisms, but sometimes others as well. The glottal stop, an interruption of breath produced by closing the vocal cords (as in the middle of English *oh-oh!*) is a common consonant. A related phenomenon, widespread in western North America, is the use of glottalized consonants, as when a *t* is produced with near simultaneous closure and reopening of the vocal cords. This is recorded with an apostrophe; it differentiates terms like Hupa (Athabascan) *teew* "underwater" from *t'eew* "raw."

The number of consonantal contrasts is also frequently expanded by distinguishing a larger number of tongue positions than do most European languages. Many languages distinguish two types of velar sounds (sounds made with the back of the tongue)—a *k* much like an English *k*, and a uvular *q*, produced further back in the mouth. Some languages even differentiate three such *k* sounds—front, middle, and back. Labiovelars, velar sounds that have simultaneous lip-rounding, are also common. Thus Tlingit has 21 phonemes made in the velar area alone: *g*, *k*, uvular *G*, *q*, glottalized *k'*, *q'*, labiovelar g^{w}, k^{w}, k'^{w}, G^{w}, q^{w}, q'^{w}, in addition to the corresponding fricatives ɣ and *x*, with uvular *X*, glottalized *x'*, *X'*, and labiovelar x^{w}, X^{w}, x'^{w}, X'^{w}. In comparison, English has only two sounds, *k* and *g*, made in the same area of the mouth.

Another class of sounds common in North America, especially in the West, is that of the laterals, which are produced by stopping the breath with the central part of the tongue but allowing it to escape at the sides. Alongside the common lateral *l*, such as exists in English, many Indian languages have a voiceless counterpart, similar to the Welsh *ll*; this sound is approximated by the *thl* in northwestern place names such as Cathlamet. To this some languages also add glottalized varieties, as well as a close-knit *tl* unit, which may in turn be aspirated or glottalized, so that there may result, as in Navajo, a total of five distinguishable lateral sounds.

In some Indian languages, as in English, stress is significant in distinguishing the meaning of words. In others, musical pitch plays a linguistic function, as it does in Chinese; *e.g.*, in Navajo, *bíní'* is "his nostril," *bìní'* is "his face," and *bìní'* is "his waist." (High and low pitches are indicated with the acute and grave accents, respectively.)

A peculiarity of some northwest coast languages is their use of complex consonant clusters, as in Bella Coola *tlk'^{w}ix^{w}* "don't swallow it." Some words even lack vowels entirely; *e.g.*, *nmnmk'* "animal."

Phonological change

Processes of phonological change, in which differences of sound are associated with grammatical distinctions (as with English *f* and *v* in "half," "halves," "to halve"), are also found in North American languages. In some languages, for example, consonantal change is related to diminutive meaning: thus Luiseño *r* changes to *d* in *ŋarúŋru-š* "pot," *ŋadúŋdu-mal* "pot-small." Vowel harmony, a process whereby vowels change to resemble adjacent ones, is further attested in North America. Yurok in northwestern California, for example, has an unusual *r* vowel, comparable to the sound in English "bird"; when this occurs in a suffix, stem vowels change to agree with it, thus *lo'óɣè* "black" + *-'r'y* (animate suffix) yields *lr'rɣr'r'y* "black animal."

Vocabulary. The word stock of American Indian languages, like those of other languages, is composed both of simple stems and of derived constructions; the derivational processes commonly include affixation (the use of prefixes, suffixes, etc.) in addition to compounding in some languages. A few languages use internal sound change, similar to the case of English "song" from "sing"; *e.g.*, Yurok *pontet* "ashes," *prncrc* "dust," *prncrh* "to be gray." New vocabulary items are also acquired by borrowing, as mentioned above.

It should be noted that, in languages generally, the

meaning of a vocabulary item cannot be adequately inferred from a knowledge of its historical origin or from knowing the meaning of its parts. For example, the name of an early 19th-century trapper, McKay, entered Karok as *mákkay*, but with the extended meaning of "white man." It was then compounded with a native noun *váas* "deerskin blanket" to give the neologism *makáy-vaas* "cloth"; this in turn was compounded with *yukúkku* "moccasin" to give *makayvas-yukúkku* "tennis shoes." At each stage of vocabulary formation, meaning is determined not simply by etymology but also by arbitrary extensions or limitations of semantic value.

Semantic structure of Indian languages

It is in the area of semantic structure that American Indian vocabulary is likely to present some surprises to the investigator. It is frequently observed that the immense diversity of the physical universe is reduced by every society to a manageable set of classifications embodied in its vocabulary. But there are few universals in such classification, and every language makes its unique semantic divisions. One language may make many specific discriminations in a particular area, while another is content with a few general terms; the difference is correlated with the importance of the semantic area for the particular society. Thus English is highly specific in classifying bovines (bull, cow, calf, heifer, steer, ox), even to the point of lacking a general cover term in the singular (what is the singular of cattle?), but for other species it has only cover terms like camel, llama. North American Indian vocabularies, as would be expected, embody semantic classifications that reflect native American environmental conditions and cultural traditions.

Interest in the semantic classifications of American Indian languages, especially in Hopi, has been particularly stimulated by the work of the American investigator, Benjamin Lee Whorf. When English discriminates "airplane," "aviator," and "flying insect," Hopi generalizes with a single term *masa'ytaka*, roughly "flier"; but when English uses a single general term, "water," Hopi differentiates *pāhe* "water in nature" from *kēyi* "water in a container."

The vocabularies of different languages may differ not only in the categorization of particular items but also in the general principles of semantic organization; such differences may be found even between neighbouring languages in a single culture area. English, for example, tends to exhaust the universe of flora and fauna with multilevelled hierarchical classifications such as "plant, bush, berry bush, gooseberry bush" or "animal, insect, louse, body louse," but the languages of northwestern California, by contrast, have relatively few generic terms and many vocabulary items that do not fall into any such hierarchy. The generic terms of Yurok refer, roughly, to "quadruped mammal," "fish," "snake," "bird," "tree," "bush," "grass," "flower," and "berry"; the organization in the neighbouring Tolowa language is simpler, lacking "quadruped mammal" and "fish." In such frameworks, a term like Yurok *wrrɣr* "body louse" cannot be subsumed in the larger classes of "louse" or "insect" because none exist. The placing of terms in semantic pigeonholes tends to be replaced, in these semantic systems, by identifying them in terms of similarity. A Yurok speaker, asked to identify a flowering bush for which he knows no name will describe it not as "a kind of bush," but as *sahsip seɣon* "similar to wild lilac." Such evidence suggests that the semantic structures of some American Indian vocabularies are based on classes defined less by their boundaries than by their centres.

Kinship terms

Another type of semantic structuring is illustrated by certain systems of kinship terms. In Fox, an Algonkian language, the term for maternal uncle also includes maternal grandmother's sister's son's son (a kind of second cousin). This can be accounted for by recognizing some very simple rules, rules that apply to the other terms of the kinship system as well: (1) siblings of the same sex, as linking relatives, are reckoned as equivalent; (2) a father's sister, as a linking relative, is equivalent to a sister, and conversely, a mother's brother's child is equivalent to a mother's brother. Then a mother's mother's sister's son's son, by rule 1, is equivalent to a mother's mother's son's son; but because one's mother's son is one's brother, this is the same as a mother's brother's son; and this in turn, by the converse of rule 2, is equivalent to a mother's brother. It is clear that the semantic systems of American Indian languages exhibit not only structures of hierarchy and similarity but also rules of semantic equivalence.

Language and culture. The exotic character of American Indian semantic structures, as manifested not only in their vocabularies but also in the relationships expressed by their morphological categories and syntactic patterns, has led a number of scholars to speculate on the relationships between language, culture, and habitual thought patterns or "world view." It was hypothesized that the unique organization of the universe that is embodied in each language might act as a determining factor in the individual's habits of perception and of thought, thus forming and maintaining particular tendencies in the associated nonlinguistic culture. As Edward Sapir put it,

> Human beings do not live in the objective world alone, . . . but are very much at the mercy of the particular language which has become the medium of expression for their society . . . The fact of the matter is that the "real world" is to a large extent unconsciously built up on the language habits of the group . . . We see and hear and otherwise experience very largely as we do because the language habits of our community predispose certain choices of interpretation.

The Whorfian hypothesis

This idea was further developed, largely on the basis of work with American Indian languages, by Sapir's student Benjamin Lee Whorf, and is now often known as the Whorfian hypothesis. Whorf's initial arguments focussed on the strikingly different organization of experience that can be found between English and Indian ways of saying "the same thing." From such linguistic differences, Whorf infers underlying differences in habits of thought. It then remains to show how these habits are manifested in nonlinguistic cultural behaviour. Thus, Whorf points out that, in Hopi, words referring to units of time (*e.g.*, "day") differ from other nouns in that they have no plural form; furthermore, they cannot be counted with the cardinal numerals ("one," "two," etc.) but only with the ordinals ("first," "second," etc.). From this he infers that when the English speaker speaks of "ten days," as if the days were an aggregate of separate units, the Hopi speaker, on the other hand, thinks in terms of the cyclic recurrence of a single phenomenon. Whorf attempts to support this idea by reference to Hopi ceremonial behaviour, which involves repeated preparation for future events. If, in the Hopi view, each day is really a recurrence, rather than something new, then it is reasonable to believe that the daily repetition of ceremonial acts will have a cumulative effect on the future. As Whorf says, the Hopi belief is diametrically opposed to the English proverb that "Tomorrow is another day."

More investigation is necessary to either prove or disprove the Whorfian hypothesis. In any case, the diversity of American Indian languages and cultures has continued to provide a rich laboratory for investigation. A particularly interesting problem is found in the area of northwestern California, where several small tribes have very similar cultures, but use languages of very diverse types. These are Karok, genetically classified as Hokan; Yurok and Wiyot, which are Algonkian; and Hupa and Tolowa, Athabascan languages. By the Whorfian hypothesis, one might expect that the difference in languages would have produced a greater diversity in the cultures; or failing that, one might expect the languages to have grown more similar to each other. In fact, both linguistic diversity and cultural uniformity seem to have made modest accommodations to each other. As an example of Whorfian linguistic determinism, the systems of biological taxonomy of Yurok and Tolowa, referred to in the previous section, may be noted. The Yurok have a larger number of generic classifications, which means they have more choice in nomenclature, because either a generic or a specific term can be used. This is consistent with the high degree of choice afforded in Yurok grammar, in which word order is nearly free and many morphological categories are optional. The sparser taxonomy of Tolowa offers less

choice, corresponding to a much more rigid grammatical structure.

Language, culture, and prehistory

A different kind of relationship between language and culture is of more interest to the student of North American prehistory, namely, the fact that language retains traces of historical changes in culture and so aids in reconstructing the remote past. Here again the pioneering work was done by Sapir, who pointed out, for instance, that the original home from which a group of related languages or dialects has dispersed is more likely to be found in the area of great linguistic diversity; *e.g.*, there are much greater differences in the English dialects of the British Isles than of the more recently settled areas such as North America or Australia. To take an American Indian example, the Athabascan languages are now found in the Southwest (Navajo, Apache), on the Pacific Coast (Tolowa, Hupa), and in the Western Subarctic. The greater diversity of the Subarctic languages leads to the hypothesis that the original centre of Athabascan migration was from that area. This northern origin of the Athabascans was further confirmed in a classic study by Sapir in which he reconstructed parts of prehistoric Athabascan vocabulary, showing, for example, how a word for "horn" had come to mean "spoon" as the ancestors of the Navajo migrated from the far north (where they made spoons of deerhorns) into the Southwest (where they made spoons out of gourds). The correlation of such linguistic findings with the data of archaeology holds great promise for the study of American Indian prehistory.

Writing and texts. Although a writing system was in use among the Mayas of Meso-America at the time of first European contact, none was known in North America. All writing systems that have been used for North American Indian languages have resulted from the stimulus of European writing, or have actually been invented and introduced by whites. Perhaps the most famous system is that invented by Sequoyah, a Cherokee, for his native language. It is not an alphabet but a syllabary, in which each symbol typically stands for a consonant–vowel sequence. The forms of characters were derived in part from the English writing system, but without regard to their English pronunciation. Well suited to the language, the syllabary fostered widespread literacy among the Cherokee until their society was disrupted by government action; its use, however, has never died out, and attempts are now being made to revive it.

Other writing systems, invented by missionaries, teachers, and linguists, have also included syllabaries; *e.g.*, for Cree, Winnebago, and some northern Athabascan languages. Elsewhere, alphabetic scripts have been used, adapted from the Roman alphabet by the use of additional letters and diacritics. White educational policy, however, has generally not encouraged literacy in Indian languages. A rich oral literature of American Indian myths, tales, and song texts has been in part published by linguists and anthropologists, and there is now increasing encouragement for the training of Indians to transcribe their own traditions—*e.g.*, among the Navajo. It is possible that there may yet be a flowering of American Indian literature, not only in spoken but also in written form.

BIBLIOGRAPHY. J.W. POWELL, "Indian Linguistic Families of America North of Mexico," *U.S. Bureau of American Ethnology, 7th Annual Report*, pp. 1–142 (1891), the first comprehensive classification; FRANZ BOAS, *Handbook of American Indian Languages*, 3 pt. (1911, 1922, and 1933–38), a classic introduction, with sketches of sample languages; HARRY HOIJER *et al.*, *Linguistic Structures of Native America* (1946), a summary of work on language classification and sketches of languages; C.F. and F.M. VOEGELIN, *Map of North American Indian Languages*, rev. ed. (1966), presents the classification used in this article; T.A. SEBEOK (ed.), *Current Trends in Linguistics*, vol. 10, *North America* (1971), surveys of different aspects of the field, with extensive bibliographies (see esp. the valuable article of JOEL SHERZER, "Areal Linguistics in North America"); M.R. HAAS, *The Prehistory of Languages* (1969), a discussion of the principles in the historical study of American Indian languages; WALLACE CHAFE, "Estimates Regarding the Present Speakers of North American Indian Languages," *International Journal of American Linguistics*, 28:162–171 (1962), and 31:345–346 (1965), gives data used in the present article; EDWARD SAPIR, *Selected Writings in Language, Culture, and Personality* (1949), articles on the relationship of language and culture in aboriginal North America; B.L. WHORF, *Language, Thought, and Reality: Selected Writings* (1956), classic articles on American Indian language and world view.

(W.O.B.)

North American Peoples and Cultures

The American Indians had their origins in Asia and are basically Mongoloid in physical type. The New World may be dismissed as the home of early hominid development because no fossil progenitors of modern man have been found and the evolution of the Primates clearly occurred in the Old World. The date of first arrival in North America as yet has not been accurately established, but it is assumed to have occurred sometime during the last glacial period, about 20,000 to 35,000 years ago (though some authorities have recently argued for an even earlier date, some 50,000 years ago or more). By the time that Europeans arrived in the 15th century, the descendants of these and later waves of migrants had spread over the Americas and developed cultures adjusted to various ecological conditions.

Comparison of New and Old World traits

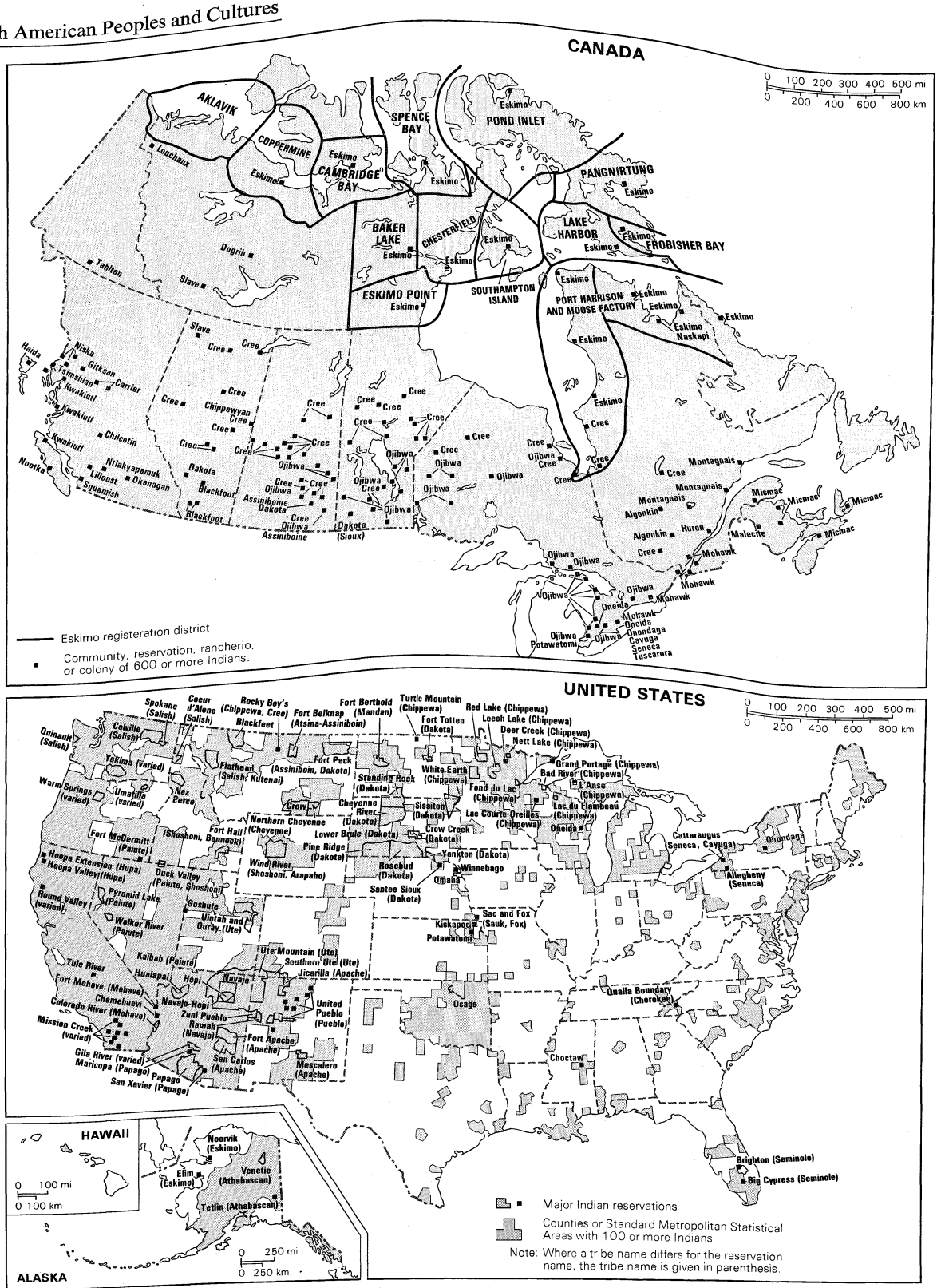
The waves of Asian incomers to the New World possessed a series of traits that were relatively ancient and were shared with most cultural groups in the Old World. These included the use of fire and the fire drill; the domesticated dog; stone implements of many kinds; the spear thrower, harpoon, and simple bow; cordage, netting, and basketry; various rites and healing beliefs and practices. Important traits lacking in the New World but known in the Old World included various significant domesticated animals, plants, and artifacts—including cattle, sheep, goats, pigs, horses, camels, and reindeer; wheat, barley, and rice; the wheel and the plow; iron; and stringed instruments. Most New World cultures depended on hunting and gathering, but the economic base of the American higher cultures came to be horticulture, with maize, beans, and squash as the staple crops. These crops were cultivated from the St. Lawrence River, in the north, to the Río de la Plata, in South America. The plants were tended by hand, using only a simple digging stick or a hoe.

The North American continent into which the Asian migrants descended is divided roughly into three major physiographic landforms: the high Cordilleras in the west, the relatively lower Appalachian Highlands and Piedmont in the east, and, between them, the Great Plains, which spearhead from the Arctic Ocean to the Gulf of Mexico in a great triangle. The Western Cordilleras are a series of parallel north–south ranges cutting from Alaska to Central America; in the area of Canada and the United States they enclose, from north to south, a large plateau of grasslands and forests, an even broader, arid plateau known as the Great Basin, and a desert plain and range area of Arizona and New Mexico (as well as northwestern Mexico). The Cordilleras along the Pacific Coast separate a plethora of basins and plateaus from the coastlands. The central Great Plains, in the extreme north, consists of Subarctic land that is swampy and coniferous and similar to the taiga and tundra of Siberia; to the south, the great Mississippi drainage system divides the relatively drier high plains to the west from the low, well-watered prairies and rolling hills to the east. The Appalachian Highlands and Piedmont contain, to the north, the great eastern woodlands and, to the south, a series of highlands and foothills descending to lowlands on the coast of the Atlantic and the Gulf of Mexico.

THE PEOPLE

Mongoloid characteristics of American Indians

Physical types. Although American Indians are fundamentally Mongoloid, considerable variation is found. The generally uniform physical features are these: the hair is usually straight, coarse, and uniformly black; the skin is reddish brown, the eyes dark, and the body hair scant; the cheekbones are prominent; and the facial size is generally large. Such features as cephalic index, nasal form, and stature, however, are extremely variable.



Distribution of North American peoples and cultures (top) in Canada and (bottom) in the United States.

The uniform features are definitely Mongoloid; the variable features are more difficult to assess. Some scholars believe that the early migrant populations were essentially Mongoloid and that the variations came about through adjustment to American environments. Other scholars argue that the New World was peopled by a variety of physical types, with later mixtures but some marginal survival in isolated regions.

Archaeologically, the earlier populations were generally long headed (dolichocephalic) and showed fewer Mongoloid characteristics. These early peoples were slight in build with well-developed brow ridges, and many represent either a proto-Mongoloid type or an unspecialized early Caucasoid form related to the Ainu of Japan.

The distribution of blood groups among the American Indians may eventually aid in solving the problem of their origins. Thus, blood type B is generally absent in the aboriginal population of the Americas (though its inci-

dence is high among Asian Mongoloids), and type A is found mainly in North American Indians.

Modern genetic theory would explain much of the variation found in terms of such factors as mutation, selection, admixture, and random genetic drift. In the small-scale groups involved in the early peopling of the New World, relatively rapid changes were possible and could account for all of the variation found.

Population and languages. Estimates of the aboriginal population are based on information supplied by explorers, traders, missionaries, and other early reporters and are only as good as the reporters' observations were trustworthy. A more serious impediment to an accurate count is that some tribes, by the time they were visited, had already been depopulated by European diseases and weapons. The estimates in the Table, it should be noted, are given an approximate dating.

The American anthropologist Alfred Louis Kroeber submitted a population total for the area north of Mexico almost identical to that given in the Table, since he used the same figures except for California. He subgrouped the material, however, to accord with subsistence areas rather than geographical boundaries. In addition, he ranked the areas according to population densities, expressed in numbers of persons per square kilometre: California area, 43.40; northwest Pacific Coast, 28.30; southwestern United States, 10.70; Columbia–Fraser river area, 7.15; eastern area, 6.95; Arctic coast, 4.02; Great Basin, 2.47; and northern area, 1.35. Although agricultural areas of the east and southwest contained the greater population (about 405,000 in all), Kroeber believed that the predominantly fishing economy of the Pacific Coast (Bering Strait to southern California) had greater relative density of population. His estimates were Pacific Coast, 25.2 persons per square kilometre; agricultural areas, 10.1; remaining area north of Mexico, 2.2.

The population of a little over 1,000,000 for North America north of Mexico contrasts with the estimated 5,000,000 for Mexico and Central America and with the estimated 25,000,000 for the Western Hemisphere as a whole. (These uncertifiable estimates must, however, be approached with caution.)

Aboriginal Indian Population North of Mexico

	date	estimated population
North Atlantic		
New England, New York, New Jersey, Pennsylvania	1600	55,600
South Atlantic		
Delaware, Maryland, Virginia, West Virginia, the Carolinas, except Cherokee country	1600	52,200
Gulf States		
Georgia, Florida, Alabama, Mississippi, Louisiana, Arkansas, Tennessee, Cherokee country	1650	114,400
Central States		
Ohio valley from Alleghenies to Mississippi, Chippewa in Canada	1650	75,300
The Plains (Canada to Gulf)		
Northern	1780	100,800
Southern	1690	41,000
Columbia River basin		
Washington, most of Oregon, northern half of Idaho	1780	89,300
California	1769	260,000
Central mountain		
Nevada, Utah, parts of surrounding states	1845	19,300
New Mexico and Arizona	1680	72,000
Subtotal United States (except Alaska)*		849,000
British America		
Eastern Canada, central Canada, British Columbia	1600–1780	190,950
Subtotal British America*		221,000
Alaska	1740	72,600
Greenland	1721	10,000
Total		1,153,450

*From analysis according to modern political divisions as of dates given.
Source: James Mooney, "The Aboriginal Population of America North of Mexico," Smithsonian Miscellaneous Collection, vol. 80, no. 7 (1928).

The outstanding characteristic of American Indian languages is their diversity. There were more than 60 language families in North America, comprising over 500 languages, but these have been reduced to a smaller number of superstocks by modern linguists. The American linguist and anthropologist Edward Sapir, for instance, proposed six linguistic groups for North America (including the Arctic): Eskimo-Aleut, Algonkian-Wakashan, Na-Dené, Penutian, Hokan-Siouan, and Aztec-Tanoan (see NORTH AMERICAN INDIAN LANGUAGES). No American language has any genetic relationship to any language group in the Old World that has yet been fully demonstrated. It may be concluded from this that the ancestors of the Indians left the Old World so long ago that any relationship was lost through linguistic change.

Diversity of languages

Culture areas. At the time of European contact there were perhaps as many as 240 different tribal entities in North America. Groups of these tribes, however, have been classified by anthropologists into a more convenient limited number of culture areas, determined very much by physiographic or environmental differences: the Subarctic, the Northwest Coast, California, the western Plateau, the western Great Basin, the Southwest, the Plains, the Eastern Woodlands, and the Southeast. This article, in dealing with the major culture areas of North America, excludes the Arctic Eskimo and Aleuts (see ARCTIC PEOPLES AND CULTURES) and the peoples of Mexico and other parts of Middle America (see MIDDLE AMERICAN PEOPLES AND CULTURES).

Subarctic. The population of the American Subarctic was always relatively small, given the vast area from Alaska to Labrador. The various tribes—Algonkian speakers in the east and Athabascan speakers in the west—were hunters, fishers, and gatherers of wild plant foods; and social organization tended to be simple and territorially very limited. The largest cohesive group tended to be a small kinship group or no more than a band or village of related families (see AMERICAN SUBARCTIC CULTURES).

Northwest Coast. The peoples of the Northwest Pacific Coast depended for their livelihood almost entirely on salmon, supplemented by other fish, sea mammals, shellfish, birds, and some food plants; they also took advantage of the woods of the forests for constructing homes and canoes and developed the use of bark fibres and mountain-goat wool to make clothing and blankets. So rich were their resources and so intensively were the resources exploited that the density of population, as calculated by Kroeber, was greater on the Northwest Coast than in most places in North America north of Mexico. Although social organization centred on the village rather than a larger tribe, it tended to be fairly sophisticated and highly stratified (see NORTHWEST COAST INDIANS).

Importance of fishing

California. From very early times until European contact, California was marked by a great complexity of tribal groups and languages. In some instances a group of not more than 500 individuals, speaking a distinct language, would live near another group of similar size that spoke a different language, and neither seemed to impinge on the other. The diversity is partly attributable to the complex of mountains and valleys and coastlands. In spite of the great complexity of linguistic and demographic factors, however, the California culture area presented no striking deviations. Agriculture was not practiced except along the Colorado River, since throughout the vast central portion and southward along the coast a method of leaching acorn pulp and converting it into flour supplied abundant and constant food. Elsewhere, fishing and hunting were equally productive (see CALIFORNIAN INDIANS).

Plateau. The northern plateau area, bounded by the Rocky Mountains and the Coast and Cascade ranges, was a land of vast rolling treeless areas, dense forests, and snow-covered mountains. Because the area is drained largely by two great river systems, the Fraser and the Columbia, most of the peoples of the area depended primarily on fishing for salmon, though other fish, game, and wild plants were also taken. Although the material accomplishments of the Plateau Indians were modest, their political achievements were sometimes impressive.

The primary political unit was the village, but a sense of larger tribal and cultural unities led sometimes to representative government, with village hierarchies, tribal chieftainships, and even confederations (see NORTH AMERICAN PLATEAU INDIANS).

Great Basin. The Great Basin, a large area centering on Nevada and Utah, though perhaps fertile during and just after the great Ice Age, became arid and impoverished in modern times. Before European contact, the Shoshonean-speaking inhabitants were divided into loosely affiliated family units that subsisted on wild seeds, small mammals, and insects. Each family was independently nomadic during most of the year and joined other families only briefly for certain hunts or dancing. Some groups acquired horses in the 18th and early 19th centuries, however, and formed bands of mounted hunters and warriors and adopted many cultural characteristics of the Plains Indians (see NORTH AMERICAN GREAT BASIN INDIANS).

Southwest. The Southwest, centering on Arizona and New Mexico, but including parts of Utah, Colorado, Texas, and Mexico, has been the site of a number of agricultural and hunting and gathering peoples. Best known are the Pueblo Indians, from the Zuni and Hopi on the west to the Rio Grande groups on the east, who built architecturally remarkable multiple apartment houses of adobe and stone masonry and developed to a high state agriculture, arts, and crafts. Their ancestors occupied the Southwest thousands of years ago, and their high cultural development began in the 1st millennium AD with the arrival of influences from Mexico. Two other major groups, the Athabascan-speaking Navajo and the Apache, undoubtedly came from the far Canadian north and probably did not initially reach the Southwest until AD 1000 or later. The Navajo borrowed extensively from the Pueblo Indians—agriculture, weaving, and arts—but the Apache remained basically hunters and gatherers; only a few groups engaged in supplementary cultivation of corn and other vegetables (see SOUTHWEST AMERICAN INDIANS).

Plains. Until the late 16th century the Great Plains were occupied only sparsely or intermittently. Toward the year 1600, however, Spanish horses were introduced and spread northward from the region of New Mexico, reaching almost the entire Plains area by 1750. Horses revolutionized the hunting of bison, making it much more profitable, and not only improved the conditions of resident peoples but also seem to have drawn peoples from surrounding areas into the plains to develop a new way of life. Thus, most tribes thought by Europeans to be typical nomadic horse Indians—such as the Cheyenne, Arapaho, and Dakota (Sioux)—were newcomers to the area and had been farmers and village dwellers not many generations before their first European contacts. In any event, the Plains Indians, with their bands organized into assemblages of tribes, their buffalo hunts and tepee villages, and their elaboration of warfare and raiding (much facilitated by the horse), became the tribes of North America often regarded now as "typical" American Indians (see NORTH AMERICAN PLAINS INDIANS).

Revolution of the horse

Eastern Woodlands. The peoples of the eastern woodlands, largely Iroquoian and Algonkian speakers, were semi-sedentary, living in villages and cultivating maize, beans, and squash. The forests provided much of their material cultures—their wigwams and longhouses made of coverings of bark sheets, their dugout and bark canoes, and their clothing made of wild animal skins—and added game and fish to their diet. A village band of a few hundred persons was the social and economic unit, though several bands might be loosely organized into tribes. The honour of warfare was engrained in most of the Eastern Woodlands Indians (see EASTERN WOODLANDS INDIANS).

Southeast. The southeastern Indians, most of them Muskogean speakers, were primarily agricultural, planting maize, pumpkins, beans, cane millet, tobacco, and other crops; they also gathered wild nuts and fruits and hunted deer or, in the west, bison. Their settlements were straggling; the "town" contained a square, on which were public and religious buildings; "villages" were often outlying. The towns were autonomous and essentially constituted tribes. They might unite into confederacies, such as those of the Creek and Choctaw, directed by councils; such confederacies might break up and recombine. In contrast to the fairly well organized political institutions, economic life was rather unsophisticated; there was little property, almost no stored wealth, and limited trade before the coming of Europeans (see SOUTHEAST AMERICAN INDIANS).

PREHISTORY

The earliest records of the peopling of North America are scanty, and it is difficult to characterize their culture beyond calling it a hunting and gathering economy. The first settlers seem to have crossed the Bering Straits region from Asia during the expansion of the glacial sheets of the Ice Age (or Pleistocene Epoch). As the great ice sheets developed and expanded, they not only covered major land areas in the Northern Hemisphere but also brought considerable areas of the continental shelves above sea level. In the Arctic this provided a tundra coastal plain across which man could move from Asia to North America. The amount of the Earth's moisture incorporated into the ice probably lowered the sea level hundreds of feet. Asia and America were thus not separated by the gradual rise of the sea until about 9,000 or 10,000 years ago; likely sites of the earliest migrants are now below sea level.

Early migrations from Asia

The Americas were the last major land mass, with the possible exception of Australia, to be occupied by prehistoric man, who first had to develop the cultural equipment to exist in the Arctic area. Once this adjustment was made, he was able to move by way of ice-free, open-land routes into the Mackenzie Basin and down into central North America. In addition to the Mackenzie route southward, at a later time the Yukon River valley also offered an ice-free route, and still later (8,000–10,000 years ago) the Liard and Peace river systems were available for intromontane travel. The Pacific Coast slope was probably available for travel at about the same time. Some migrations may also have occurred by way of the Aleutian Islands, but this would have taken place at a considerably later date.

Early Cultures. The earliest well-defined cultures in the New World have been placed by radiocarbon dating at about 10,000 to 8000 BC. At this period, two distinct traditions in North America are known: the Paleo-Indian big-game hunters of the West, the Great Plains, and eastern North America; and the Desert culture peoples of the western Basin–Range region.

Paleo-Indian hunting cultures. In spite of regional differences in detail, there was a remarkable similarity in the economic complex of the hunters. They lived in a variety of environments, from mountain passes and valleys in the west to the then better watered grasslands of the Plains and the varied forest and park-land environment of the eastern woodlands. The variety of their bone tools indicates that one of their major food supplies came from animals, the hides of which provided clothing. In the western Plains and the Southwest they hunted such extinct North American animals as the camel, ground sloth, tapir, mammoth, and horse.

In the Great Lakes area of the eastern woodlands they may have hunted mastodon, but other commoner animals, such as the elk and deer, presumably formed the bulk of their meat diet. Some of their bone and wooden tools were probably used for working and ornamentation. These early hunters had temporary shelters and moved about as small bands in search of game. Their physical type is not clearly known, but it was related to that of an eastern Asian Early Stone Age population and is less Mongoloid than many groups of American Indians of the historic period.

Archaeologically, the oldest remains of the Paleo-Indian tradition are found on kill sites, where large Pleistocene mammals were killed and butchered. The most distinctive artifact type of this horizon is the Clovis Fluted projectile point (named after the site of first discovery, near Clovis, New Mexico); this was a lance-shaped point

Oldest archaeological remains

of chipped stone that had had one or more longitudinal flakes struck from the base of each flat face. These points are accompanied by side scrapers and, in one instance, by long cylindrical shafts of ivory. They are most frequently associated with mammoth. A second Paleo-Indian horizon, which seems in part to be contemporary with the Clovis material and partially to postdate it, is the Folsom phase of the central High Plains (Folsom, New Mexico, being the site of initial discovery). It is characterized by lance-shaped points of more careful manufacture (including broader fluted surfaces) than Clovis, associated with the remains of extinct *Bison antiquus*. The Lindenmeier site, a Folsom campsite in northeastern Colorado, has yielded a wide variety of end and side scrapers, gravers, and miscellaneous bone artifacts. Clovis sites have been dated at about 9000 BC and Folsom sites at about 500 to 1,000 years later.

The Desert culture. In the western United States, from Oregon to northern Mexico and from the Pacific coast to the eastern foothills of the Rocky Mountains, there was a distinctive cultural adaptation to the dry, relatively impoverished upland environment. There, in the relative absence of large game resources, vegetation was exploited to a great extent, with the development of grinding tools and related equipment. The Cochise Desert culture (named from Cochise County in southern Arizona, where it was discovered) ran from about 8000 BC through several stages, down to the historic period in some areas.

Development of seed milling

The Desert culture people lived as small bands of wandering seasonal food gatherers, collectors, and hunters. They ate a wide variety of animal and plant foods and developed techniques for small-seed harvesting and processing; an essential feature of Desert assemblages was the milling stone, for use in grinding wild seeds. Their best known habitations were caves and rock shelters, and they had twined basketry, nets, mats, cordage, fur cloaks, sandals, wooden clubs, and digging sticks. They also had the spear thrower, with darts of pointed hardwood or with points of flint and later of obsidian. Their rough stone implements were shaped by percussion, and consequently many of their choppers and scrapers had an Earlier Stone Age appearance. Their projectile points, however, showed excellent craftsmanship and followed continent-wide styles. The domesticated dog, another migrant from Asia, was known by about 4000 BC in the Desert culture (though by this time the dog was also known elsewhere in North America).

The far west. On the far west coast in California, the marked variety of geographical situations developed a number of regional complexes dependent upon intensive exploitation of the local resources. None of these was agricultural. In the southern desert area the people subsisted upon plant seeds and small game and used crude flint tools, grinding stones, and (later) arrowheads. In the mountainous areas and in the better watered central areas, larger game animals such as the elk and deer, supplemented by acorns, fish, and birds, were the major food supply. By at least 2000 BC, in this central area, the utilization of the local resources plus cultural intrusions from the north resulted in full adaptation to the area. The coastal groups from north to south depended upon the sea for their food supply, some subsisting mainly on shellfish, some on sea mammals, others on fish, and still others on a mixture of all three.

In the north Pacific part of the United States and in western British Columbia, some of the early sites of the hunters have yielded fluted blades, crude choppers, and cutting tools. Between 9000 and 7000 BC there were varied economic activities but with an emphasis on hunting. By about 8000 BC there was a strong orientation toward salmon fishing, particularly during the salmon runs, and the peoples tended to emphasize the use of bone and antler tools. The burin, a chisel-like bone working tool, has been found in such sites, along with prepared cores and blades. During the postglacial warming period that culminated between 3000 and 2000 BC, the inhabitants of the drier areas without permanent streams took on more of the traits of the Desert culture to the south, while others turned toward riverine fishing and marsh resources or to food from the sea. In the 1st millennium BC, the so-called Marpole complex, a distinctive ground slate complex, was known in the Fraser River area, with basic resemblances to the northwest coast historic culture in maritime emphasis, woodworking, large houses, and substantial villages. The emphasis on ground slate and woodworking tools is like that in the Eastern Woodlands Archaic (see below) and recalls similar emphasis in certain northwestern Siberian cultures. In most of the areas of the Northwest Coast, clear indications of the beginnings of the historic cultures were not known until about AD 1300.

The Archaic cultures. *The Eastern Archaic.* With the retreat of the ice sheets in the north, beginning about 10,000 years ago, the cool, moist climate gradually became hot and dry in the Great Plains and Great Basin regions, with consequent extinction or migration of Pleistocene animal life. The High Plains were largely deserted by man for a considerable period. In the eastern woodlands area, partly as a result of the variety of forest environments, climatic differences, and physiographic features, there developed a series of regional readaptations to local food supplies. The change from the primarily hunting economy of the early American hunters was gradual and is clearly seen in slowly evolving projectile point and other implement changes. The pattern of life became one of mixed hunting and collecting, with some groups developing by 6000 BC a taste for riverine and coastal living with abundant fish and mollusk resources to supplement such vegetational products as acorns, seeds, berries, and tubers.

Change from hunting economy

During the long Eastern Archaic, from 8000 to 1500 BC, regional social and economic diversification was developed, and it was during the Archaic that significant early linguistic differentiation also probably occurred and during which varieties of physical types developed.

The typical Archaic house was a small circular structure with wooden posts for the wall and roof supports; the covering was probably bark. Cooking was done in the open by boiling in containers of wood, bark, or hides or by baking in pits or by roasting and grilling. Identification lists of mammal, fish, and bird bones from Archaic sites read like a listing of the early historic fauna. Various game-gathering devices, including nets, traps, and pitfalls, were used along with the spear and dart thrower. Fishhooks, gorges, and net sinkers were known, and in some areas fish weirs were built. River, lake, and ocean mollusks were consumed, and probably a great many native roots, berries, fruits, and tubers known in the early historic period were incorporated into the diet during the Archaic. The extensive lists of plant medicines recorded by the early colonists were probably a part of the primitive Archaic pharmacopoeia.

The large variety of chipped-flint projectiles, knives, scrapers, perforators, drills, and adzes reflect regional styles and changes during the long Archaic period. The late Archaic was distinguished by the gradual development of ground and polished, grooved stone axes, pestles, gouges, adzes, plummets, and forms attached to the spear thrower. This was a reflection of a growing versatility in the technology and economy. Trade and exchange are also known from the distribution of native copper implements from the Michigan–Wisconsin area to as far south as Louisiana and Florida and the finds of southeastern marine shells as far north as the upper Mississippi–Great Lakes area. An extensive system of trails and water routes was probably in existence during the Late Archaic.

Developing trade and technology

The great boreal forest zone of spruce, fir, and pine that now runs from New England and the maritime provinces of Canada westward to the Canadian plains and the Mackenzie Valley gradually acquired its present distribution following the retreat and melting of the Arctic ice cap. Its present distribution was reached by about 2500 BC. The forest cover and climate had a limiting effect on the cultural development and on the general pattern of hunting and fishing. These efforts were supplemented by some use of plant material.

In the upper Great Lakes area there was an Old Copper

culture, which has special interest because copper implements and weapons were made from the native copper of the Lake Superior basin. This culture appeared about 3000 BC and lasted about 2,000 years. It was a northern expression of the Late Archaic. Its tools and weapons, particularly in the adzes, gouges, and axes, clearly indicate an adaptation to the forest environment. In the area south of James Bay to the upper St. Lawrence about 2000 BC, there was a regional variant called the Laurentian Boreal Archaic and, in the extreme east, the Maritime Boreal Archaic. In this eastern area, slate was shaped into points and knives of forms similar to those of the copper implements to the west. Trade between the eastern and western areas has been recognized, and this evidence, along with general similarities of the culture, suggests that water transportation by canoe was known at this time.

Along the southern border of the central and eastern boreal forest zone between 1500 and 500 BC, there developed a distinctive burial complex, reflecting an increased attention to burial ceremonialism. These burials, many including cremations, were often accompanied by red ochre, caches of triangular blanks, fire-making kits of iron pyrites and flint strikers, copper needles and awls, and polished stone forms. The triangular points of this complex may have represented the introduction of the bow and arrow from the pre-Eskimo cultures east of Hudson Bay. The earliest Woodland pottery appeared in the Great Lakes area about 1000 BC. It is another of the culture traits derived from northeastern Asia and across northern Alaska to northwestern Canada. The route by which it reached the Great Lakes is not known.

The Plains Archaic. In the western Plains from about 8000 to 3000 BC the fluted blade points were no longer made, and many styles or types were produced that have been identified by such local names as Plainview, Angostura, Milnesand, Agate Basin, and Scottsbluff. These minor varieties of dart and spear point and their primarily hunting culture may be included in the term Plano. The Plano complex or culture type was a direct descendant from the fluted-blade early American hunters. Their primary game animal was the bison, for the larger animals of the preceding period had died out or were exterminated.

Plano culture

The stone complex associated with the Plano hunters was markedly similar from site to site over a considerable period of time during which the climate became increasingly warmer and until the major warm period was reached, about 3000 to 2000 BC. As the climate moderated, peoples of the Late Plano complex moved north into Saskatchewan and Alberta with the grazing game animals and, by 3000 BC, had reached the Arctic tundra zone in the Northwest Territories of Canada at Grant and Dismal lakes and Great Bear River. Important elements of this culture also moved east in the Mississippi valley and western Great Lakes area. Many of the sites of this culture type were kill sites with abundant bison bones that accounted for the number of implements and tools associated with hunting and leatherworking. In the tundra zone the major game animal was the caribou. Choppers, pounders, and milling stones have been found there.

Early agriculturalists. *Early southwestern planters.* Primitive agricultural practices began in Mexico by 6000 to 4000 BC and by approximately 2000 BC were known on the northern fringe of the Middle American culture area. Maize was not the only crop plant, for gourds, squash, peppers, cotton, and varieties of beans were also domesticated. Maize was grown in the southwestern United States by 2000 to 1000 BC, but most of the other domesticates did not arrive until just before and after AD 1. The early introduction of maize in the Southwest had no marked effect on cultural development, and the existence of pottery, storage pits, and domestic houses with semi-subterranean floors and lateral entryways were not known until about AD 1. These houses had wood uprights for walls, central roof supports, radiating beams, and wattle-and-daub plastered walls. The small settlements of the early Puebloan, or Basket Maker, people of the Four Corners area (namely northwestern New Mexico, southwestern Colorado, southeastern Utah, and northeastern Arizona) were among the first village agricultural societies in the Southwest.

Ohio Valley farmers. In eastern North America one of the earliest known phases in which corn cultivation appears to have had a role in subsistence is the Adena, which occupied the middle Ohio River Valley by about 800 BC (Adena takes its name from an estate near Chillicothe, Ohio, the site of a discovered burial mound). The stimulus of the Adena farmers was apparently instrumental in bringing about the spectacular Hopewell culture in the Illinois and Ohio valleys. (Hopewell is similarly named after a farmsite in Ohio). The success of the Hopewell peoples (200 BC to AD 200) seems to have been due largely to their combining elements of the preceding Archaic cultures with elements of the Adena culture and perhaps with some features of a local cultivating tradition. It is evident that the Hopewell culture included a well-organized village-based society in which surplus resources were used in the construction of elaborate earthworks and were concentrated as wealth by a restricted group of individuals. The most outstanding feature of Hopewell culture is a burial complex that called for the deposition of concentrations of wealth in tombs of one or several deceased individuals. The interment procedure was elaborate and involved the construction of a large log tomb, later burned and covered by an earth mound. Artifacts found within these burial mounds indicate that the Hopewell were able to obtain goods from widespread localities in North America. Obsidian and grizzly bear teeth were apparently derived from the Rocky Mountain region; copper from the northern Great Lakes; and conch shells and other exotic objects from the southeast and along the coast of the Gulf of Mexico. Ohio, particularly, served as a distributing centre for ceremonial goods and special products over a wide area in the eastern United States. The ceramics of the Hopewell appear to be based in two major traditions: one derived from northern Asia, which reached eastern North America by about 1000 BC, and the other from Middle America, where certain decorative techniques, characteristic of finger Hopewell pottery, existed several hundred years prior to the earliest appearance of the Hopewell culture. In less favourable areas of eastern North America, a "generalized Woodland" culture paralleled the Hopewell in time, probably based more on collecting than on cultivation for subsistence.

Adena and Hopewell cultures

There is a clear evidence of cultural regression between AD 200 and 700 in the north central United States following the Hopewell expansion and florescence. This is attributed to a minor cold phase that did not allow a continuation of agriculture in this area under the techniques then current. Although there was concurrent change in the south, this did not take the form of a lowering of the cultural level.

Mississippi Valley and peripheral woodlands. The last major cultural development in the eastern United States is called Mississippian because its primary centre was in the valleys of the Mississippi River and its major tributaries and in the southeast. This predominantly agricultural complex was a marked cultural advance over earlier stages in the east. Its initial growth and expansion was at approximately the same period (AD 700–1200) as that of the southwestern Puebloan complex. The initial growth was along the Mississippi between modern St. Louis and Vicksburg. It was stimulated by the introduction of concepts, religious practices, and improved agricultural procedures from northern Mexico, which resulted in a sedentary societal organization. By AD 1000, large villages were in existence with subsidiary villages and farming communities nearby. Regional specialized production in pottery, projectile points, house types, and other utilitarian products reflected the tribal groupings of the period. An outstanding feature of this culture type was the earthen temple mound, which served as a raised platform on which the major community buildings were placed. These council houses and temples served as the political and ceremonial centres. The platform mounds were placed on the sides of a central plaza that served as a ceremonial

The Mississippian culture complex

centre for the tribal community during important recurrent functions or during times of crisis. The more permanent buildings, both family and community, were of wattle-and-daub construction, usually rectangular in floor plan. In some areas large, circular charnel houses received the remains of the dead, but burial was normally made in large cemeteries or in the floors of dwellings. The size of the ceremonial tribal centres varied from 10 to 100 acres (four to 40 hectares). Important household industries involved the production of mats, baskets, clothing, and a variety of vessel forms for specialized uses. Food surplus was kept in ground storage pits and in storage cribs above the ground.

One of the more striking developments was the production of ceremonial costumes and ornaments, for use in the religious ceremonies that were conducted by an organized priesthood with a well-established ritual. The religious symbolism spread throughout the Mississippian complex, and a number of centres of production of specialized ceremonial items are known. Other innovations were walled fortifications with timber palisades and bastions surrounding the village, which reflected an increase in intergroup aggression and a tendency, continuing into the historic period, toward the development of confederacies. The intergroup conflicts apparently were primarily quests for prestige and revenge instead of a means of territorial expansion or economic control.

Spread of Woodland culture

Along the eastern and northern periphery, some tribes, while retaining the older Woodland complex, were somewhat influenced by the Mississippian culture. The extent of this influence seems to have depended on their nearness to the more advanced cultural complex and on their ability to maintain an agricultural economy along with hunting and gathering. There was a spread of Woodland culture from about 200 BC to AD 200 into the eastern part of the Plains from Oklahoma to North Dakota, with some sites, particularly in eastern Kansas, clearly forming a part of the Hopewellian complex. In the Plains there was evidence of corn and bean cultivation during this period, and later there was cultivation of gourds and squash, but between about AD 300–400 and 800 there was little occupation of the western part of the Plains by agricultural people because of the relative aridity.

After 800, however, Late Woodland populations had spread west to the eastern slopes of the Rockies and were in contact with eastward-moving Puebloan people. A favourable agricultural period was indicated by the marked increase in village size and in population density for the next 400 years, during which hospitable areas along major streams were occupied by various interrelated cultural groups collectively known as the Plains Mississippian cultures. Part of this complex was connected to the developing Mississippi complexes to the east by diffusion and, to some degree, by a migration of such groups as the Omaha and Ponca from the St. Louis area by about AD 1000.

Between AD 1500 and 1700 the High Plains from New Mexico to Wyoming and in eastern Oklahoma, Kansas, and Nebraska were pre-empted by horse-using, semi-agricultural peoples of the plains—the Apache and Comanche. Prehistoric village agriculturalists of a plains Mississippi tradition came into the historic period as the Pawnee, Arikara, Mandan, Hidatsa, Crow, and Wichita.

Southwestern village farmers. *Anasazi, Mogollon, and Hohokam cultures.* The southwestern village farmers were distributed from eastern Utah and southern Colorado through most of New Mexico and Arizona. The effective agricultural area varied with fluctuations in climate that profoundly affected the ability of the Indians

Pre-Pueblo cultures

to occupy marginal regions. Although corn and some other agricultural plants were introduced from Mexico between 2000 BC and AD 1, the first village complexes, with five to 15 pit or surface houses, ceremonial buildings, refuse pits, and pottery, did not appear until shortly before AD 1 in southern Arizona and New Mexico. Two of the major farming complexes began at this time: Mogollon was located in the mountainous belt of west central New Mexico and east central Arizona, while Hohokam was located in the desert area of the Gila basin of southern Arizona. The latter group depended upon irrigation for its crops, whereas Mogollon depended upon rainfall and stream diversion over floodplains. Mogollon became the pattern of agriculture that later was developed in the Anasazi or Puebloan culture, the third major farming complex of the Southwest.

The geographical expansion, population growth, and striking development of permanent villages with multiroom and multilevel buildings came during the period from AD 700 to 1200, which coincided with a minor climatic period of favourable distribution of rainfall for plant growth over the entire Southwest. For the same climatic reasons, there was an expansion of population and cultural movement from central and western Mexico into northwestern Mexico. Trade and cultural stimuli then moved from northwestern Mexico into the American Southwest at a time when the climate in both areas was most favourable for population and cultural growth. Indicating such cultural movement, cast copper bells, parrots, ball courts, shell trumpets, and pottery vessel shapes and designs have been found; they clearly reflect the transmission of religious beliefs and ceremonies. These southern influences were blended into local and regional complexes.

The Anasazi village agricultural complex had expanded by AD 900 to occupy northeastern Arizona, southwestern Colorado, and northwestern New Mexico. By AD 1100, expansion had taken place into the Virgin River valley of southeastern Nevada, north as far as the Great Salt Lake and northwestern Colorado, to the east into southeastern Colorado and to the Pecos and upper Canadian river valleys of New Mexico. During this period there was probably a development of priestly offices and of rituals and ceremonialism. The increasing population concentration in large pueblos was apparently organized into households according to lineage. Control of the agricultural activities was presumably in the hands of clan leaders, who were also the priests who officiated in the rain-producing ceremonies. During this period some of the larger village populations ranged from 300 to more than 1,000 people.

Primarily because of increasing aridity there was a marked retraction of Anasazi culture between 1100 and 1300. As a result, a concentration of the pueblos took place in northeastern Arizona, along the Rio Grande and its immediate tributaries, and in the present Zuni area of western New Mexico. The Anasazi groups maintained their societies by sand-dune farming with floodwater and some canal irrigation. The increased importance and elaboration of religious rain-producing ceremonies between 1300 and 1540 is deduced from paintings on walls and from symbolic pottery decoration.

Mogollan and Hohokam cultures

The Mogollon complex in its early phases, from 200 BC to AD 700, consisted of relatively small villages of pit houses grouped near a large ceremonial structure. No organization of the village structures into a pattern is apparent, however, and trash disposal was random. Although the initial impetus for sedentary village life appeared early in the Mogollon area, there was a period of apparent cultural quiescence about AD 400 to 600. With the growth and spread of the Anasazi complex in the period after 700, the main flow of culture was from that area, and Mogollon villages from AD 900 to 1100 were a blend of local development strongly influenced from Anasazi. During the climatic deterioration after AD 1200, much of the Mogollon territory in southwestern New Mexico was abandoned.

The Hohokam culture of southeastern Arizona was primarily limited to main river valleys. Agriculture was made possible by extensive irrigation canals that required cooperation between villages. The people lived in villages of scattered pit houses made of brush and mud that were dispersed along the streams and canals. Their main settlements and major culture growth took place also during the period AD 700–1200. Following this for 200 years, there was a blend with Anasazi and Mexican elements and a tendency toward the construction of more compact settlements surrounded by compound walls with a few massive multiroom and two-story buildings. There is relatively little evidence of trade and influences from north-

western Mexico. Such historic groups as the Pima and Papago are descended from the Hohokam people.

Pueblo culture. Best known of the prehistoric and historic southwestern peoples are the Pueblo Indians proper, whose ancestors built great cliff villages now seen in ruins and equally remarkable multiple apartment houses of adobe and stone masonry. Some of the latter are still occupied, and the Pueblo Indian inhabitants speak languages and observe ceremonies that are at least pre-Spanish in origin.

Pueblo culture periods

The beginnings of Pueblo culture, in the 1st millennium AD, are obscure. The traditional type of aboveground, straight-line, or crescent-shaped multiple house continued to be built, two rooms wide; stone masonry, however, began to replace the earlier pole-and-mud and adobe construction. Agriculture, including several varieties of corn, may have been augmented at that time by the cultivation of a native long-staple cotton. Pottery was not much changed, but it included a greater variety of shapes and decoration. Basketry was much less common. These early phases of Pueblo culture are termed Developmental Pueblo.

The great Classic Pueblo period followed in about AD 1050–1300, a period most popularly associated with the term Pueblo. It was the time of the great cliff houses, such as Mesa Verde, and the large apartment-like structures in Chaco Canyon (Pueblo Bonito) and elsewhere. An actual shrinking in area took place as inhabitants of the outer fringes moved in to build the large dwelling units. Also, because a number of outstanding structures were built in quite inaccessible canyons and mesa walls, there is the possibility that hostile strangers had reached the outlying districts. The most notable advance over previous periods was in architecture and pottery. Masonry walls were greatly thickened, dressed stones being used in many localities to bear the greater weight of massive structures. These community structures had from 20 to as many as 1,000 rooms and from one to four stories. Each of the larger houses was in effect a single village. Windows and doors were quite small, and usually no openings were made in the lowest rooms, which were entered by ladder through the roof. Floors were terraced or set back, and the terraces were much used as outdoor living space. Roofs were constructed to carry great weights by laying heavy beams covered with a mat of smaller poles and brush, then laying on a coat of adobe six to eight inches thick. Some semi-subterranean ceremonial chambers, known as kivas, were enlarged to as much as 80 feet (25 metres) in diameter. Craftsmanship in pottery reached a high level, and specialization became so pronounced in the different centres, as in Chaco canyon, Mesa Verde, Kayenta, and a number of others, that the style of each can be recognized easily. To the earlier black-on-white and red-on-white designs were added polychromes of three or more colours applied more lavishly. Cotton cloth, blankets, and bags were woven, and yucca fibre also entered into various articles of clothing and such utility objects as mats. Feather-cloth robes were worn in cold weather.

Abandonment of the cliff houses and large community buildings marked the close of the great Pueblo period. In part this may have resulted from incursion into the northern part of the territory by nomadic Athabascans (Navajo and Apache) and a prolonged drought that occurred in the late 13th century. It is also possible that lack of central leadership led to internal dissensions.

The next period (AD 1300–1700), called Regressive Pueblo, was characterized by a general movement southward and eastward, and new villages, some larger than those of Classic Pueblo, were built on the Little Colorado, Puerco, Verde, San Francisco, Rio Grande, Pecos, upper Gila, and Salt rivers. Pottery showed new developments; geometric patterns were largely replaced by naturalistic representations of birds, animals, insects, and the human figure; glazing was frequently used. The modern Pueblo period is usually dated from the beginning of permanent Spanish settlement at the close of the 17th century. From 1540 on, when the Spaniards first entered the Pueblo country, the number of Pueblo settlements declined considerably, though much of the culture and many of the skills in agriculture and crafts continued down to present times.

EVOLUTION OF CONTEMPORARY NORTH AMERICAN INDIANS

The great diversity that marked Indian cultures also persisted into modern times. After the arrival of the Europeans, some tribes disappeared by amalgamation with others or by wars and epidemics, and some languages perished. The process of adaptation to white encroachments and of acculturation to European ways worked at varying rates within the various Indian societies, transforming and devitalizing segments of custom and practice. Nevertheless, many Indian populations north of Mexico continued to manifest an unexpected viability. Although the total Indian population of the area in the 1970s was only about one-half or two-thirds of what it was at the time of the first European contacts, it was increasing, at least in the United States, faster than the general population—this, despite rates of poverty, death, and disease far greater than those for the general population. As for Indian attitudes, customs, or values, it is obviously not possible to state in quantitative terms how much they continue to function; nevertheless, their persistence does seem undeniable. It is clear that American Indians are not a vanishing race, as was once thought.

Colonial policies. The formulation of public policy toward the Indians was of concern to the major European colonizing powers. The Spanish tried assiduously to Christianize the natives and to remake their living patterns. Orders were issued to congregate scattered Indian villages in orderly, well-placed centres, assuring the Indians at the same time that by moving to such centres they would not lose their outlying lands. This was the first attempt to create Indian reservations. The promise failed to protect Indian land, according to the Franciscan monk and historian of Mexico, Juan Torquemada, who reported about 1599 that there was hardly "a palm of land" that the Spaniards had not taken. Many Indians who did not join the congregations for fear of losing what they owned fled to mountain places and lost their lands anyway.

Early Spanish policies

The Russians never seriously undertook colonization in the New World. When Peter I the Great sent Vitus Jonassen Bering into the northern sea that bears his name, interest was in scientific discovery, not overseas territory. Later, when the problem of protecting and perhaps expanding Russian occupation was placed before Catherine II the Great, she declared (1769):

> It is for traders to traffic where they please. I will furnish neither men, nor ships, nor money, and I renounce forever all lands and possessions in the East Indies and in America.

The Swedish and Dutch attempts at colonization were so brief that neither left a strong imprint on New World practices. The Dutch government, however, was probably the first (1645) of the European powers to enter into a formal treaty with an Indian tribe, the Mohawk. Thus began a relationship, inherited by the British, that contributed to the ascendancy of the English over the French in North America.

France handicapped its colonial venture by transporting to the New World a modified feudal system of land tenure that discouraged permanent settlement. Throughout the period of French occupation, emphasis was on trade rather than on land acquisition and development, and thus French administrators, in dealing with the various tribes, tried primarily only to establish trade relations with them. The French instituted the custom of inviting the headmen of all tribes with which they carried on trade to come once a year to Montreal, where the governor of Canada gave out presents and talked of friendship. The governor of Louisiana met southern Indians at Mobile. The English, reluctantly, found themselves competing on the same basis with annual gifts. Still later, United States peace commissioners were to offer permanent annuities in exchange for tribal concessions of land or other interests.

Early English policies

In contrast to the French, the English were primarily interested in land and permanent settlements; beginning quite early in their occupation, they felt an obligation to bargain with the Indians and to conclude formal agree-

ments with compensation to presumed Indian landowners. The Plymouth settlers, coming without royal sanction, thought it incumbent upon them to make terms with the Massachuset Indians. Cecilius Calvert (the 2nd Baron Baltimore) and William Penn, while possessing royal grants in Maryland and Pennsylvania respectively, nevertheless took pains to purchase occupancy rights from the Indians. It became the practice of most of the colonies to prohibit indiscriminate and unauthorized appropriation of Indian land. The usual requirement was that purchases could be consummated only by agreement with the tribal headman, followed by approval of the governor or other official of the colony. At an early date also, specific areas were set aside for exclusive Indian use. Virginia in 1656 and commissioners for the United Colonies of New England in 1658 agreed to the creation of such reserved areas. Plymouth Colony in 1685 designated for individual Indians separate tracts that could not be alienated without their consent.

In spite of these official efforts to protect Indian lands, unauthorized entry and use caused constant friction through the colonial period. Rivalry with the French, who lost no opportunity to point out to the Indians how their lands were being encroached upon by the English; the activity of land speculators, who succeeded in obtaining large grants beyond the settled frontiers; and, finally, the startling success of the Ottawa chief Pontiac in capturing English strongholds in the old Northwest (the Great Lakes region) as a protest against this westward movement, together prompted the king's ministers to issue a proclamation (1763) that formalized the concept of Indian land titles for the first time in the history of European colonization in the New World. The document prohibited issuance of patents to any lands claimed by a tribe unless the Indian title had first been extinguished by purchase or treaty. The proclamation reserved for the use of the tribes "all the Lands and Territories lying to the Westward of the sources of the Rivers which fall into the Sea from the West and Northwest." Land west of the Appalachians might not be purchased or entered upon by private persons, but purchases might be made in the name of the king or one of the colonies at a council meeting of the Indians.

This policy continued up to the termination of British rule and was adopted by the United States. The Appalachian barrier was soon passed—thousands of settlers crossed the mountains during the American Revolution—but both the Articles of Confederation and the federal Constitution reserved either to the president or to Congress sole authority in Indian affairs, including authority to extinguish Indian title by treaty. When French dominion in Canada capitulated in 1760, the English announced that "the Savages or Indian Allies of his most Christian Majesty, shall be maintained in the lands they inhabit, if they choose to remain there." Thereafter, the proclamation of 1763 applied in Canada and was embodied in the practices of the dominion government. (The British North America Act of 1867, which created modern Canada, provided that the parliament of Canada should have exclusive legislative authority with respect to "Indians, and lands reserved for the Indians." Thus, both North American countries made control over Indian matters a national concern.)

United States policy. The first full declaration of U.S. policy was embodied in the Northwest Ordinance (1787), which stated:

> The utmost good faith shall always be observed toward the Indians, their lands and property shall never be taken from them without their consent; and in their property, rights, and liberty, they shall never be invaded or disturbed, unless in just and lawful wars authorized by congress; but laws founded in justice and humanity shall from time to time be made, for preventing wrongs being done to them, and for preserving peace and friendship with them.

This doctrine was embodied in the act of August 7, 1789, as one of the first declarations of the U.S. Congress under the Constitution.

The final shaping of the legal and political rights of the Indian tribes is found in the opinions of Chief Justice John Marshall, notably in decision in the case of *Worcester* v. *Georgia*:

> The Indian nations had always been considered as distinct, independent, political communities, retaining their original natural rights, as the undisputed possessors of the land, from time immemorial. . . . The settled doctrine of the law of nations is, that a weaker power does not surrender its independence—its right to self-government—by associating with a stronger, and taking its protection. A weak state, in order to provide for its safety, may place itself under the protection of one more powerful, without stripping itself of the right of government, and ceasing to be a state.

The first major departure from the policy of respecting Indian rights came with the Indian Removal Act of 1830. For the first time the United States resorted to coercion, particularly in the cases of the Cherokee and Seminole tribes, as a means of securing compliance. The Removal Act was not in itself coercive, since it authorized the President only to negotiate with tribes east of the Mississippi on a basis of payment for their lands; it called for improvements in the east and a grant of land west of the river, to which perpetual title would be attached. In carrying out the law, however, resistance was met with military force. In the decade following, almost the entire population of perhaps 100,000 Indians was moved westward. The episode moved Alexis de Tocqueville to remark in 1831:

Indian Removal Act of 1830

> The Europeans continued to surround [the Indians] on every side, and to confine them within narrower limits . . . and the Indians have been ruined by a competition which they had not the means of sustaining. They were isolated in their own country, and their race only constituted a little colony of troublesome strangers in the midst of a numerous and dominant people.

The territory west of the Mississippi, it turned out, was not so remote as had been supposed. The discovery of gold in California (1848) started a new sequence of treaties, designed to extinguish Indian title to lands lying in the path of the overland routes to the Pacific. The sudden surge of thousands of wagon trains through the last of the Indian country and the consequent slaughtering of prairie and mountain game that provided subsistence for the Indians brought on the most serious Indian wars the country had experienced. For three decades, beginning in the 1850s, raids and sporadic pitched fighting took place up and down the western Plains, highlighted by such incidents as the Custer massacre by Sioux and Cheyenne Indians (1876), the Nez Percé chief Joseph's running battle in 1877 against superior U.S. army forces, and the Chiricahua Geronimo's long duel with authorities in the Southwest, resulting in his capture and imprisonment in 1886.

Indian wars

Toward the close of that period, the Ghost Dance religion, arising out of the dream revelations of a young Paiute Indian, Wovoka, promised the Indians a return to the old life and reunion with their departed kinsmen. The songs and ceremonies born of this revelation swept across the northern Plains. The movement came to an abrupt end December 29, 1890, at Wounded Knee Creek, South Dakota. Believing that the Ghost Dance was disturbing an uneasy peace, government agents moved to arrest ringleaders. Sitting Bull was killed (December 15) while being taken into custody, and two weeks later units of the U.S. 7th Cavalry at Wounded Knee shot down more than 200 men, women, and children who had already agreed to return to their homes.

A further major shift of policy had occurred in 1871 after congressional discussions lasting several years. The President, with the advice and consent of the Senate, had continued to make treaties with the Indian tribes and to commit the United States to the payment of sums of money. The House of Representatives protested, since a number of congressmen had come to the view that treaties with Indian tribes were an absurdity (a view earlier held by Andrew Jackson). The Senate yielded, and the act of March 3, 1871, declared that "hereafter no Indian nation or tribe" would be recognized "as an independent power with whom the United States may contract by treaty." Indian affairs were brought under the legislative control of the Congress to an extent that had

not been attempted previously. Tribal authority with respect to criminal offenses committed by members within the tribe was reduced to the extent that murder and other major crimes were placed under the jurisdiction of the federal courts.

Radical land allotment legislation

The most radical undertaking of the new legislative policy was the Dawes General Allotment Act of 1887. By that time the Indian tribes had been moved out of the mainstreams of traffic and were settled on lands that they had chosen out of the larger areas that they had formerly occupied. Their choice in most cases had been confirmed by treaty, agreement, act of Congress, or executive order of the president. The tribes that lived by hunting over wide areas found reservation confinement a threat to their existence. Generally, they had insisted on annuity payments or rations, or both, and the U.S. peace commissioners had been willing to offer such a price in return for important land cessions. In time the view came to be held that reservation life fostered indolence and perpetuated customs and attitudes that held Indians back from assimilation. The strategy offered by proponents of this theory was the Allotment Act authorizing the president to divide the reservations into individual parcels and to give every Indian, whether he wanted it or not, a particular piece of the tribally owned land. In order not to make the transition too abrupt, the land would be held in trust for a period of 25 years, after which ownership would devolve upon the individual. With it would go all the rights and duties of citizenship. Reservation land remaining after all living members of the tribes had been provided with allotments was declared surplus, and the president was authorized to open it for entry by non-Indian homesteaders, the Indians being paid the homestead price.

A total of 118 reservations was allotted in this manner, but the result was not what had been anticipated. Through the alienation of surplus lands (making no allowance for children yet unborn) and through patenting of individual holdings, the Indians lost 86,000,000 acres (35,000,000 hectares), or 62 percent, of a total of 138,000,000 acres in Indian ownership prior to 1887. A generation of landless Indians resulted, with no vocational training to relieve them of dependence upon land. The strategy also failed in that ownership of land did not effect an automatic acculturation in those Indians who received individual parcels. Through scattering of individuals and families, moreover, social cohesiveness tended to break down. The result was a weakening of native institutions and cultural practices with nothing offered in substitution. What was intended as transition proved to be a blind alley.

Indian population had been dwindling through the decades after mid-19th century. The California Indians alone, it was estimated, dropped from 100,000 in 1853 to not more than 30,000 in 1864 and 19,000 in 1906. Cholera in the central Plains in 1849 struck the Pawnee. As late as 1870–71 an epidemic of smallpox brought disaster to the Blackfeet, Assiniboin, and Cree. These events gave currency to the concept of the Indian as "the vanishing American." The decision of 1871 to discontinue treaty making and the passage of the Allotment Act of 1887 were both founded in the belief that the Indians would not survive, and hence it did not much matter whether their views were sought in advance of legislation or whether lands were provided for coming generations. When it became obvious after about 1920 that the Indians, whose numbers had remained static for several years, were surely increasing, the United States was without a policy for advancing the interests of a living people.

Indian Reorganization Act of 1934

A survey in 1926 brought into clear focus the failings of the previous 40 years. The investigators found most Indians "extremely poor," in bad health, without education, and lacking adjustment to the dominant culture around them. Under the impetus of these findings and other pressures for reform, Congress adopted the Indian Reorganization Act of 1934, which contemplated an orderly decrease of federal control and a concomitant increase of Indian self-government and responsibility. The essentials of the new law were as follows: (1) allotment of tribal lands was prohibited in the future, but tribes might assign use rights to individuals; (2) so-called surplus lands not pre-empted by homesteaders might be returned to the tribes; (3) tribes might adopt written constitutions and charters of incorporation embodying their continuing inherent powers to manage internal affairs; and (4) funds were authorized for the establishment of a revolving credit program, for land purchases, for educational assistance, and for aiding the tribes in forming organizations. Moreover, the act could be rejected on any reservation by referendum.

The response of the Indian population to the 1934 act was indicative of their ability to rise above adversity. About 160 tribes, bands, and Alaska villages adopted written constitutions, some of which combined traditional practices with modern parliamentary methods. The revolving credit fund helped Indians build up their herds and improve their economic position in many other ways. Borrowers from the fund were tribal corporations, credit associations, and cooperative groups that loaned to individual Indians and to group enterprises on a multimillion-dollar scale. Educational and health services were also improved through federal aid.

Originally, the United States exercised no guardianship over the person of the Indian; after 1871, when internal tribal matters became the subject of national legislation, the number and variety of regulatory measures multiplied rapidly. In the same year that the Indian Reorganization Act was passed, Congress significantly repealed 12 statutes that had made it possible to hold Indians virtual prisoners on their reservations. Indians were then able to come and go as freely as all other persons. The Snyder Act of 1924, extending citizenship to all Indians born in the United States, opened the door to full participation. Few Indians took advantage of the law, and because of their lack of interest a number of states excluded Indians from the franchise. Organization of tribal governments following the Reorganization Act, however, seemed to awaken an interest in civic affairs beyond tribal boundaries, and when Indians asked for the franchise, they were generally able to secure it eventually, though not until 1948 in Arizona and New Mexico, after a lengthy court action.

The federal courts consistently upheld the treaties made with Indian tribes and also held that property may not be taken from Indians, whether or not a treaty exists, "except in fair trade." The latter contention was offered by the Hualapai Indians against the Sante Fe Railway. The company was required by the courts in 1944 to relinquish about 500,000 acres it thought had been granted to it by the U.S. The lands had been occupied since prehistory by the Indians, without benefit of treaty recognition, and the Supreme Court held that, if the occupancy could be proved, as it subsequently was, the Indians were entitled to have their lands restored. In 1950 the Ute Indians were awarded a judgment against the United States of $31,750,000 for lands taken without adequate compensation. A special Indian Claims Commission, created by act of Congress on August 13, 1946, received many petitions for land claims against the United States and awarded, for example, about $14,789,000 to the Cherokee nation, $10,242,000 to the Crow tribe, $3,650,000 to the Snake-Paiute of Oregon, $3,000,000 to the Nez Percé, and $12,300,000 to the Seminole.

Search for new policies

The period from the early 1950s to the 1970s was one of increasing federal attempts to seek new policies regarding the Indians, and it was also a period in which Indians themselves became increasingly vocal in their quest for a better measure of human rights and the correction of past wrongs. The first major shift in policy came in 1954, when the Department of the Interior began terminating federal control over those Indians and reservations deemed able to look after their own affairs. From 1954 to 1960, support to 61 tribes and other Indian groups was ended by the withdrawal of federal services or trust supervision. The results, however, were unhappy. Some Indian groups, in extreme poverty, lost more acreage in deals for the private exploitation of Indian land and water resources. Indians in certain states became subject exclusively to state laws that were less liberal or sympa-

thetic than federal laws. Finally, the protests of Indians, anthropologists, and other interested groups became so insistent that the program was decelerated in 1960. In 1961 a trained anthropologist was sworn in as commissioner of Indian Affairs, the first anthropologist ever to hold that position. Federal aid expanded greatly, and in the ensuing decade Indians were specifically brought into various federal programs for equal economic opportunity. Indian unemployment remained severe, however—in some areas, as much as 10 times the rate for the general population.

During these decades, new Indian organizations, such as the National Congress of American Indians, arose with authority to speak for the various tribal groups, and for the first time in American history Indians were taking a truly active role in deciding their own futures. American Indians came more and more into public attention in the 1960s and 1970s as they sought (along with other racial minorities) to achieve a better life. One of the most dramatic demonstrations was their temporary seizure of the abandoned island of Alcatraz. Although the response of the federal government has been uncertain and there is still a groping for better policies, there does seem to be better recognition of the plight of the Indians and of the need for more favourable policies and programs.

BIBLIOGRAPHY. Some of the more notable general surveys of the American Indians are CLARK WISSLER, *Indians of the United States*, rev. ed. (1966), and *The American Indian: An Introduction to the Anthropology of the New World*, 3rd ed. (1938); H.E. DRIVER, *Indians of North America*, 2nd ed. rev. (1969); A.L. KROEBER, *Cultural and Natural Areas of Native North America* (1939); RUTH M. UNDERHILL, *Red Man's America* (1953); PAUL S. MARTIN, GEORGE I. QUIMBY, and DONALD COLLIER, *Indians Before Columbus* (1947); FRED EGGAN (ed.), *Social Anthropology of North American Tribes*, 2nd ed. (1955); and E.B. LEACOCK and N.O. LURIE (eds.), *North American Indians in Historical Perspective* (1971). An extensive listing of books and articles on particular Indian groups is given in GEORGE P. MURDOCK, *Ethnographic Bibliography of North America*, 3rd ed. (1960). Names, locations, and historical events are summarized in FREDERICK WEBB HODGE (ed.), *Handbook of American Indians, North of Mexico*, 2 vol. (1912, reprinted 1968); and JOHN R. SWANTON, *The Indian Tribes of North America* (1952).

For archaeology and prehistory some useful introductory works are J.D. JENNINGS and EDWARD NORBECK (eds.), *Prehistoric Man in the New World* (1964); DIAMOND JENNESS (ed.), *The American Aborigines* (1933); KENNETH MACGOWAN and J.A. HESTER, JR., *Early Man in the New World* (1962); and H.M. WORMINGTON, *Ancient Man in North America*, 4th ed. (1957). For physical anthropology some important surveys are W.S. LAUGHLIN (ed.), *Papers on the Physical Anthropology of the American Indian* (1951); T.D. STEWART, "A Physical Anthropologist's View of the Peopling of the New World," *SWest. J. Anthrop.* 16:259–273 (1960); and T.D. STEWART and M.T. NEWMAN, "An Historical Résumé of the Concept of Differences in Indian Types," *Am. Anthrop.*, 53:19–26 (1951).

There are no general surveys of all the Subarctic Indians, but FREDERICK JOHNSON (ed.), *Man in Northeastern North America* (1946); and DIAMOND JENNESS, *The Indians of Canada*, 7th ed. (1967), provide a useful introduction to most of them. Only two general surveys of the Northwest Coast Indians are available, both by the same author, PHILLIP DRUCKER: *Indians of the Northwest Coast* (1963) and *Cultures of the North Pacific Coast* (1965). For Californian Indians, useful surveys are contained in A.L. KROEBER, *The Handbook of the Indians of California* (1925); JACK FORBES, *Native Americans of California and Nevada* (1968); and ROBERT HEIZER and ALAN ALMQUIST, *The Other Californians* (1971). The Plateau Indians are treated in E.H. SWANSON, *The Emergence of Plateau Culture* (1962); and in two works by V.F. RAY: *Cultural Relations in the Plateau of Northwestern America* (1939) and *Culture Element Distributions: XXII, Plateau* (1942). There are no general surveys of all the Great Basin Indians, but some information can be found in E.H. SWANSON (ed.), *Languages and Cultures of Western North America* (1970). Important accounts of the southwestern Indians are contained in EDWARD H. SPICER, *Cycles of Conquest* (1962); FRED EGGAN, *Social Organization of the Western Pueblos* (1950); EDWARD P. DOZIER, *The Pueblo Indians of North America* (1970); and CLYDE KLUCKHOHN and DOROTHEA LEIGHTON, *The Navaho* (1946). The Plains Indian tribes, individually and generally, have been the subject of many volumes by many hands; among the useful introductory books are CLARK WISSLER, *North American Indians of the Plains*, 3rd ed. (1927); and R.H. LOWIE, *Indians of the Plains* (1963). For the Indians of the northeast or the eastern woodlands there are a number of works, including FREDERICK JOHNSON (ed.), *Man in Northeastern North America* (1946); ROBERT E. and PAT RITZENTHALER, *The Woodland Indians of the Western Great Lakes* (1970); and GEORGE T. HUNT, *The Wars of the Iroquois* (1940). For the southeast, JOHN R. SWANTON, *The Indians of the Southeastern United States* (1946, reprinted 1969), is the standard introduction.

North American Plains Indians

The Indians of the North American Plains are popularly regarded as the typical American Indians. They were essentially big-game hunters, the buffalo being a primary source of food and equally important as a source of materials for clothing, shelter, and tools. Until supplanted by the white man from the 16th century onward, they occupied the area between the Mississippi River and the Rocky Mountains, which includes portions of both the United States and Canada. It is a vast grassland stretching from northern Alberta and Saskatchewan in Canada to the Rio Grande border of Texas.

The climate is in general a continental one, with a wide seasonal range. Temperatures in winter may go below 0° F (−18° C) and in summer as high as 110° F (43° C). The plant cover varies with the amount of moisture, the tall grass of the prairies in the east giving way at about the 100th meridian to the shorter grass of the High Plains in the west. The area is drained principally by the Missouri–Mississippi river system.

Peoples and languages

The peoples of the Plains are designated by the languages they speak. It is permissible to call them "tribes" or "nations," bearing in mind, however, that in some cases, for example the Dakota (popularly known as Sioux), the designation covers several completely autonomous political divisions. The northern and southern divisions of the Cheyenne retained their unity as a tribe, while the Pawnee on the other hand comprised at least four independent groups. Many of the tribes of the Plains, such as the Cheyenne, migrated into them from the prairies and woodlands of the east. In addition, some of the tribes to the west of the area—the Ute and Jicarilla Apache, for instance—were influenced to a degree by the Plains culture and can be regarded marginal to the area.

Six distinct language families or stocks were represented in the Plains area, although none of them was confined to it. The speakers of the several languages within a stock might or might not be geographically contiguous. Some of the languages, moreover, were more closely related to each other than to others within the same stock. Thus languages belonging to the Algonkian stock included the Blackfoot (Piegan–Blood–Northern Blackfoot), Arapaho–Atsina (Gros Ventre), and Plains Cree, Plains Ojibwa, all in the Northern Plains, while Cheyenne, also an Algonkian language, was in the central part of the area. The Siouan language stock embraced Mandan, Hidatsa, Crow, Dakota–Assiniboin, Omaha–Ponca–Osage–Kansa, and Iowa–Oto–Missouri. The Pawnee–Arikara and Wichita were Caddoan languages, whereas Wind River Shoshoni and Comanche were of the Uto-Aztecan stock. The Athabascan (Na-Dené) stock was represented by the Sarcee (Sarsi) in the northern part of the area and by the Kiowa–Apache in the southern. Finally the Kiowa–Tanoan stock was represented in the area by one language, Kiowa.

Sign language provided a common, if limited, means of communication among tribes speaking different languages. This was a system of fixed hand and finger positions symbolizing ideas, the meanings of which were known to the majority of the tribes of the area.

General cultural characteristics

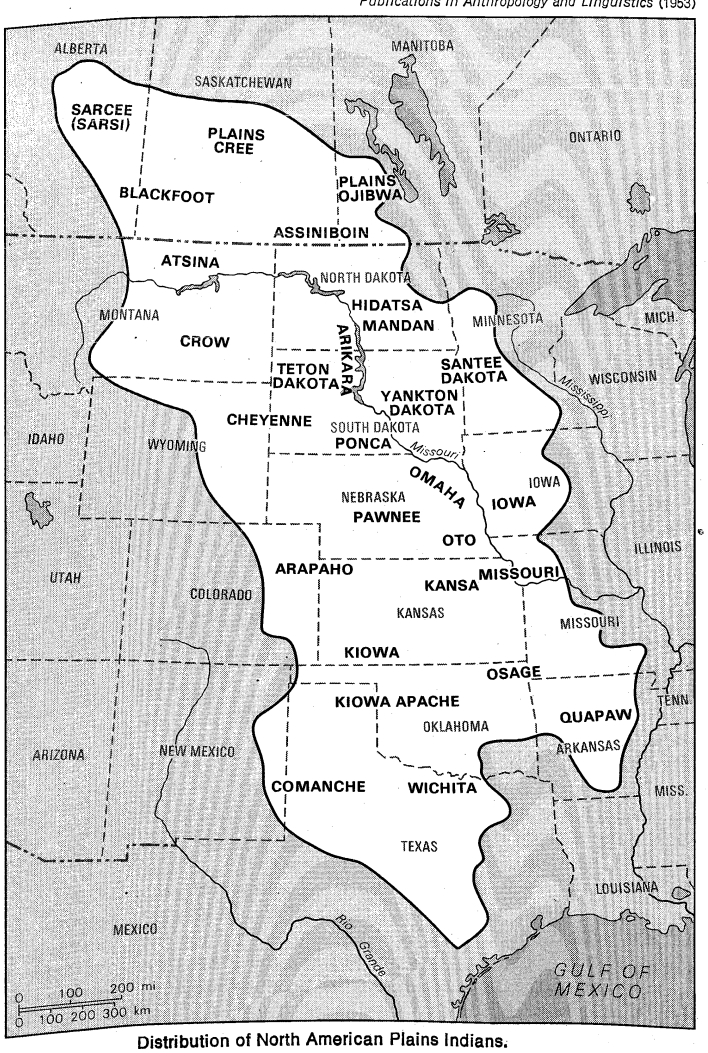
In aboriginal times the only domestic animal was the dog, which served some tribes for food but was in general use as a pack animal. Dogs were also made to draw the travois, a vehicle consisting of two poles in the shape of a V, the point of the V dragging on the ground; the ends were attached to the animal, and a platform was put across to carry the burden. The Spaniards introduced the

horse in the 16th century, in the southwest, and its use gradually spread northward. By the middle of the 18th century or earlier, all of the Plains Indians were equestrian. There was also a flowering of what one authority has termed luxury developments—"showy clothing, embroidered footgear, medicine-bundle purchases, elaborate rituals [culminating in the sun dance], [and especially] gratuitous and time-consuming warfare." The fighting seldom involved major tribal forces; it was carried out mainly by raiding parties of a few warriors, to avenge a death, to steal horses, and especially to gain glory. Touching an enemy's body was considered among many groups of greater moment than killing or scalping him.

In many of its aspects, Plains culture emphasized intense individualism and aggression, placing a high valuation on violent experience. Rivalry and hostility between men were part of the pattern, so much so that military organizations had to be established to maintain order at tribal assemblies. Nevertheless, a man who became eminent in war could enhance his status by showing generosity to the poor, sharing his goods with relatives, engaging in lavish hospitality, and behaving cooperatively with others.

The tribes who were marginal to the Plains generally borrowed only the external traits of Plains culture and not the religious, ceremonial, or social customs. On the other hand, the southern tribes of the Siouan language family, along with the Pawnee, Mandan, Hidatsa, and Arikara, while possessing most of the typical Plains Indian cultural traits, had a number of others, such as horticulture, pottery, and residence in fixed villages for part of the year; they may be said to constitute a subculture within the Plains Indians.

From H. Driver *et al.*, *Indiana University Publications in Anthropology and Linguistics* (1953)



Distribution of North American Plains Indians.

SOCIAL ORGANIZATION

Local and territorial units. Among the nomadic tribes the local units were bands—*i.e.*, groups of people wandering together in search of food. The size and composition of the bands varied. Comanche bands were loosely organized, each centring its activities in a vaguely defined area within the tribal territory. Only the larger bands received permanent recognition and a name. The bands did not fight one another, but neither did they act in concert as a tribe. The Teton Dakota tribe comprised seven independent bands, the largest being the Oglala. The Cheyenne, however, were more highly organized than the Teton; their ten bands sent representatives to a council of 44 peace chiefs, whose decrees were binding on all the bands. Only in late spring and summer, as a rule, did the nomadic band or tribe come together as a unit on the open Plains, when they engaged in the communal hunt and held major ceremonies. During the remainder of the year, the members dispersed in small groups to more sheltered areas.

Villagers

The semi-sedentary prairie dwellers lived in villages. The three Hidatsa villages were each independent of the others, whereas the Skidi Pawnee villages were united. The seasonal round of the village peoples may be illustrated by the Arikara, who spent the period between planting their crops and harvesting them on the summer hunt living as nomads. After the harvest small groups went to the Plains for another hunt, returning not to their permanent villages but to protected areas in the bottomlands, going back to the permanent villages to plant their crops in the spring.

Family and kinship units. Each local band or village was composed of families and wider kinship units. Every group had regulations governing the mating of kin. Some, such as the Atsina and Blackfoot, did not tolerate marriage between consanguineal relatives, no matter how distant the tie, and others proscribed marriage within varying degrees of relationship. On the other hand, marriage between those who were already relatives by marriage was often prescribed or preferred; the custom of a man marrying the widow of his deceased brother was widespread, as was also the replacement of a deceased wife by her sister. Most marriages were monogamous, but a man might have several wives, sisters being preferred.

Ideally, marriages were arranged between the families of the bride and groom, the latter usually paying a bride-price; sometimes, as among the Mandan, each side provided exactly equivalent gifts. Virginity was highly prized among most of the tribes, particularly the Cheyenne. Among the Blackfoot, women known to be chaste were selected for roles in important ceremonies. The double standard prevailed, however, and men in all of the tribes were expected to make sexual conquests. A husband who had evidence of his wife's infidelity might disfigure her by cutting off her nose or demand compensation from her lover. Romantic love and elopement were not unknown, but attitudes varied; the Teton tolerated the couple on their return, while the Cheyenne considered the girl disgraced forever.

Most Plains tribes had definite rules governing conduct between in-laws, such as the widespread "mother-in-law taboo" in which a man and his wife's mother showed their mutual respect by not speaking to, or in some cases not even looking at, each other. In some tribes, as among the Arapaho, the taboo was extended to include a woman and her father-in-law. The Atsina and a few other tribes required brothers-in-law to be very circumspect in their speech, avoiding any reference to sex no matter how indirect. Yet many of the tribes adhering strictly to this respect taboo permitted the greatest freedom between a man and his sister-in-law. Among the Crow they were expected to romp with each other and to talk to each other in vile language; the Atsina encouraged mutual practical joking and teasing; the Blackfoot allowed the same freedom as between man and wife. It is notable that, according to marriage rules on the Plains, the parties to this joking relationship were potential mates.

Family lines

While some of the Plains societies reckoned descent bilaterally—that is, equally in both the male and female

lines—others reckoned descent exclusively in either the male or female line. This did not mean that there was no recognition of the other parent and his or her relatives. The Hidatsa had a matrilineal clan system (*i.e.*, they traced their descent through females back to a common ancestress), yet there were important relationships to the father and his clansmen; they were always treated with respect and often presented with gifts; before battle a man would ask his clan father to paint his face; the clan father would give personal names to a clansman's son; and ceremonially the father's clan folk played an important part in performances such as the sun dance.

The Mandan and Crow also had matrilineal clan systems. The Oto and Missouri are reported to have had them as well, but there is little information on these groups. The patrilineal clan system was characteristic of the Iowa, Kansa, Omaha, Osage, and Ponca. There is some question as to whether or not the Blackfoot and Atsina had clans, since there was a tendency for the child to be a member of his father's subdivision, and marriage within this subdivision was frowned upon. The organization of the eastern Dakota and the Assiniboin is likewise in doubt.

Among some groups certain clans regarded themselves as more closely related to each other than to other clans. Among the Kansa the 16 clans were grouped into seven larger units. Often the larger units, or phratries, had no important function, although presumably in some cases they regulated marriage. Occasionally there was a further, higher grouping of phratries into two complementary units, or "moieties," as anthropologists call them. The Ponca moieties were composed of two phratries, each consisting of two clans. A similar dual division also existed among societies lacking clans and phratries. The Pawnee, for instance, were divided into the Summer People and the Winter People.

Social rank and warfare. There were no hereditary social classes, but there was ranking of individuals. A poor man, with the help of supernatural power, could win wealth and standing, mainly through prowess at war; but the son of a wealthy family would have an advantage over an orphan.

Most tribes ranked war exploits, but they did not all evaluate particular deeds alike. The taking of scalps was common, but many tribes considered scalping of lesser merit than counting coups—touching one's enemy. Stealing a valuable horse that had been picketed at its owner's lodge was considered a feat.

Most tribes had a number of clubs or associations, religious and secular. Among the latter were the military societies that often functioned as police for the tribal hunt. Some organizations were rivals. Among the Crow, for example, there were two outstanding societies, the Lumpwoods and the Foxes, that were of equal rank but competed violently in feats of war. The Arapaho, Atsina, Blackfoot, Mandan, and Hidatsa differed from the other Plains tribes in that their military societies were ranked in an ordered series. Distinctive regalia and membership privileges in each society were purchased collectively by one group of roughly the same age from an older group. Then the sellers as a group bought from the next older group, the exchange continuing until the oldest group sold out completely and retired from the system. The number of societies varied. The Hidatsa at one time had as many as ten military societies.

Women participated in many of the associations, often playing important roles. Among the Mandan and Hidatsa, women's societies existed similar to the men's graded societies.

Child rearing

Socialization and education. Training began early for Plains children, as part of their play. A very small boy would be given a bow and blunt arrows; as he grew stronger, he would receive larger, heavier bows and be shown how to stalk small game and to hit moving targets. Groups of boys engaged in shooting matches and sham battles, the winners receiving acclaim from their elders; the losers were praised if they had fought bravely. Competition marked almost all of the boys' games.

Girls were taught domestic skills. A father might make toy scraping tools for a girl, which her mother would teach her to use. She would learn to sew by making clothes for her doll and be given a toy tepee to put up while her mother was erecting the big one. In general the line between work and play was not sharply drawn.

The young were encouraged to behave in desired ways by praise and reward, special attention being given in many of the tribes to the first success. Thus an Oto father publicly gave away property to honour his son when the boy first walked, when he brought in his first small game, when he killed his first deer, and when he returned from his first war party. When a Crow boy killed his first big game animal, he was given public recognition; a song celebrating the achievement was sung at a ceremony similar to that which would mark his return from a first war party. Progress toward maturity was generally rewarded by removing restrictions and granting special privileges. Blackfoot boys who won shooting matches were allowed to wear feathers in their hair. As soon as he went on his first war party, the Cheyenne boy was relieved from the duty of herding horses and also from the necessity of listening to long lectures on proper behaviour.

Relatives helped to train children. Grandparents were often consultants and advisers. In a number of tribes, the mother's brother and the father's sister played important roles as mentors and disciplinarians. Among the matrilineal Hidatsa, the maternal uncle was responsible for the direction and supervision of his nephews; he guided them and punished them, but also praised them. Arapaho parents relied on the father's sister to instruct a girl in proper behaviour and to reprimand her when necessary. Physical punishment was seldom employed. Praise and reward for achievement seem to have been generally emphasized more than ridicule and admonishment for failure.

Though the quest for supernatural power through a vision or dream was important among all of the Plains tribes, the experience was not sought primarily at puberty, as was the case in other areas. Again, while fear of menstrual blood was universal, only a few tribes, including the Cree, marked the occurrence of the girl's first menses with an adolescent rite.

ECONOMIC SYSTEMS

Housing

The nomadic tribes lived throughout the year in portable dwellings, while the semi-sedentary peoples used them only seasonally. The tent, or tepee, was conical in shape, the foundation being either three or four poles with other poles placed around them to form a circular base. The cover was made from dressed buffalo skins carefully fitted and sewn together. Since the fireplace was in the centre, a smoke hole was left at the top that could be closed in bad weather.

When the whole tribe assembled, a camp circle was usually formed, leaving the space in the centre for ceremonial structures. Among some peoples, such as the Cheyenne and Atsina, each subgroup had a defined place in the circle. Among many tribes, too, the orientation of the lodges and the opening of the circle was toward the rising sun.

The earth lodge of the semi-sedentary peoples was a permanent structure much larger than a tepee; it was dome shaped, roofed with earth, and entered by a covered passage. The Pawnee, the Mandan, and some other tribes excavated the floor so that their dwellings were partly subterranean. The Osage and Wichita houses differed from those of the other horticultural tribes; the dwellings of the Osage were oval in ground plan, composed of upright poles arched over on top, interlaced with horizontal withes, and covered with mats or skins, while the Wichita houses were conical in shape and thatched with grass.

Technology

The nomads depended for subsistence primarily on big game: buffalo, antelope, deer, and elk. These were also important in the diet of the semi-sedentary tribes. While the animals could be hunted by individuals, the usual methods involved the whole tribe in driving the game down a cliff or into a corral or encircling it by fire. The introduction of the horse increased hunting efficiency, allowing larger numbers of game animals to be killed

more quickly. The mounted hunter continued to prefer the bow and arrow over guns.

The semi-sedentary tribes practicing horticulture raised principally maize (corn), but also beans, squashes, and sunflowers. The plots were cultivated by the women, using only a rake, a digging stick, and a hoe made from the shoulder blade of an elk or a buffalo.

Animal skins were used for clothing. On the Northern Plains, men wore a shirt, leggings reaching to the hips, moccasins, and a buffalo robe—the robe being painted to depict the war deeds of the owner. Women's clothing consisted of a long dress, leggings to the knee, and moccasins. Among the villagers and some southern nomads, men traditionally left the upper part of the body bare. Clothes were decorated with porcupine-quill embroidery, fringe, and, in later times, beadwork. Ordinarily the head was not covered, the feathered warbonnet and other elaborate headgear being reserved for ceremonial occasions.

Receptacles of various kinds were made from rawhide and leather. Traditional tools were of bone, horn, antler, and stone—before the introduction of metal by Europeans. Pipes were usually of stone. Basketry and pottery were characteristic products of the villagers; some of the nomads, including the Cheyenne, Comanche, and Arapaho, were said to have made flat coiled gambling trays, while the Atsina, Blackfoot, and Cree, among others, had traditions of having once made earthenware.

Property and exchange

Some anthropologists have argued that Indians could not have lived on the Plains before the introduction of the horse. Others, pointing to the fact that Francisco Vásquez de Coronado's expedition in 1541 encountered fully nomadic buffalo-hunting tribes on the Southern Plains who lacked horses and depended on dogs for transport, claim that the acquisition of the horse produced only minor changes. One consequence of the horse was the creation of great differences in wealth. Horse stealing became a major motive for warfare. The man who had horses to give away or to offer as bride-price was at a distinct advantage in social prestige.

There was very little intratribal trade in material goods, although there was much exchange of intangibles. The transfer of war medicine and of curing rites brought high prices in horses and other goods among practically all of the tribes. For the spiritual benefit believed to accrue from viewing the contents of a sacred pipe bundle (*i.e.*, a pack containing various sacred objects) of the Mandan, the individual had to pay, in the 1830s, the equivalent of what was then $100. Among the Hidatsa a person wishing to learn to chip flint and make arrows had to buy the rights and receive the instructions from those with ceremonial rights who possessed bundles carrying arrow-making songs.

Intertribal trade was fairly common, one form being that between nomads and villagers; *e.g.*, the exchange of skin robes for grain. The Cheyenne were middlemen in the trade of horses between the Indians of the Southern Plains and those of the north central Plains. Guns and other materials such as blankets, beads, cloth, and kettles, introduced into the northeast by the British and French, were highly valued by the Comanche, Kiowa, and other groups, who were willing to give horses in exchange for them.

BELIEF SYSTEMS

Communication with spirits

The Plains Indians did not distinguish sharply between the sacred and the secular nor between religion and magic. They attached much importance to visions. Success in life was believed to depend in large measure on the intervention of friendly spirits. The usual procedure for obtaining spirit help was to go to some lonely spot to fast, mortify the flesh, and beg for aid. If the suppliant was successful, the spirit would appear to him or be audible to him and would give him detailed instructions to follow to win immunity in battle, ability to cure illness, or various other kinds of power. The spirit might assume the form of an animal or a bird. Not everyone was successful in this quest, and, among the Crow and some other tribes, those with power were permitted to transfer it to others less fortunate. Among the Atsina and Teton Dakota, women might be vouchsafed visions, although they did not usually seek the experience.

All of the tribes had medicine men, or shamans, who had received supernatural powers. Arapaho, Atsina, and Cheyenne medicine men would walk on fire as a proof of their powers. More important was the ability of the shaman to cure illness. While, in most of the groups, ordinary illnesses such as dysentery or headaches would be treated with herbal remedies, a shaman was called in to diagnose and treat more serious illnesses. It was widely believed that illness was caused by intrusion of a foreign substance in the body and that the medicine man could cure the patient by extracting the object. If the medicine man failed, the reason was that there had been some unwitting infraction of the rules as laid down by his supernatural sponsor. He was not required to take every case; among the Teton Dakota he could refuse after examining a patient. Other services a medicine man might render included locating enemies and game animals and even finding lost objects.

In some tribes it is difficult to distinguish the role of medicine man, who had direct contact with the supernatural, from that of the priest who obtained his knowledge from other practitioners. The Cheyenne medicine man is thought to have been more of a priest than a shaman, since his main road to supernatural power was through acquisition of ritual knowledge from one who was already a priest, although some did seek power through visions. The same individual may have acted in some situations as a shaman and in others as a priest.

Among the tribes having a clear belief in a spirit superior to all other spirits were the Cheyenne, the Atsina, and the Pawnee. The Cheyenne, for instance, held that "the Wise One above" knew better than all other creatures; long ago he had left the Earth and retired to the sky. In smoking ceremonies the first offering of the pipe was always made to him. Some of the other tribes, such as the Crow, are not known to have believed in such a supreme deity.

Ceremonial and ritual were well developed on the Plains. They ranged from very simple rites to complicated proceedings requiring weeks of preparation, the final performance lasting for days. A number of common ritual elements were used alone or combined in various ways by the several tribes. Medicine bundles figured prominently in rituals throughout the area. In some cases the bundle was a personal one, the contents of which had been suggested to the individual by a supernatural sponsor, while in others it was a tribal property originating in the mythological past. It was handled reverently and opened according to definite rules. The opening of the Cheyenne sacred arrow bundle, for instance, was the focus of an elaborate tribal rite extending over four days.

The sacred number for most tribes was four, entering into the rituals in many ways. A less common number was seven. Many rituals centred on a kind of altar, a specially prepared space in a ceremonial structure for arranging sacred objects or smoking them with incense. The dimensions of the altar and the symbols that were used varied with the ceremony. Ritual purification in a sweat lodge was a widespread practice, required in connection with many ceremonies.

The sun dance

One important ceremony found among about 20 tribes is known inaccurately in English as the sun dance. The native terms varied: the Cheyenne phrase may be translated as "New Life Lodge"; the Atsina term means "Sacrifice Lodge." While the central features were the same among all the tribes, there were many differences in detail. The sun dance was always held in summer, when the whole tribe gathered, and was usually performed in fulfillment of a vow by someone who had promised it if he were relieved of some grave difficulty. The ceremony was an annual event among the Teton Dakota but occurred at quite irregular intervals among the Crow. The pledger was instructed by a priest or medicine man, and some weeks were needed for gathering food and other preparations. A ceremonial structure was built in the centre of the camp circle, and, before it was erected, offerings were placed in the fork of the central log. Within the structure was an altar upon which buffalo skulls

were laid. The pledger and other participants fasted and danced for several days, praying for power. A widespread, though not universal, feature of the ceremony was self-torture by some of the participants. The skin of the breast or back was pierced, and a wooden skewer inserted. One end of a rope was tied to the skewer, the other end being attached to the centre pole. The dancer leaned back till the line was taut and he strained until he tore himself loose. The Teton elaborated on this torture by dragging around buffalo skulls attached to skewers on the dancers' legs.

THE IMPACT OF THE WHITE MAN

Cultural changes

With the coming of the white man, the Plains Indians began to acquire manufactured articles such as guns, metal utensils, axes, knives, blankets, and cloth. This led to a decline of the native arts and crafts. Paradoxically, however, some aspects of social life were intensified as a result of the fur trade. Since women dressed the hides, the successful hunters secured more and more wives to do the dressing for them, and therefore polygyny increased on the Northern Plains. Religion was affected insofar as wealth brought by the fur trade encouraged more frequent transfer of medicine bundles, at higher prices.

With the coming of immigrant wagons, the building of railroads across the Plains, and the encroachment of white settlements, warfare became a unifying force. During the latter half of the 19th century, tribes that had formerly been hostile to one another often united against the intruders. Not infrequently the Indians were successful, although in the end they were overwhelmed. Eventually the buffalo disappeared, the system of status and rank collapsed, and the tribes were placed on reservations. The culture of the Indians was radically changed.

The United States government hoped to make the Indians into literate farmers, but the agents sent to teach them encountered many obstacles. The nomadic groups were loath to settle down, looking upon cattle as a poor substitute for buffalo. The reservation land was often unsuitable for agriculture. The semi-settled village peoples, among whom digging-stick cultivation was traditionally considered women's work, resisted the change in division of labour brought by the plow. Much confusion resulted when officials insisted on listing families by surnames, which Indians did not possess. Many misunderstandings arose among the matrilineal tribes when inheritance rules were changed so that land passed from father to son.

Schools were established on and off the reservation. Boarding schools had the advantage of facilitating the learning of the white man's language and customs. While some individual Indians adapted to the new conditions and were able to make their way among whites, those who returned from school to the reservations often found themselves in a difficult and marginal position.

Many Indians were Christianized. A new religious movement known as peyotism, combining pagan and Christian elements, spread among the Plains Indians in the latter part of the 19th century. It centred on a species of cactus that when eaten or imbibed caused hallucinations. Since it was considered dangerous by the government as well as by the missionaries, efforts were made to suppress it. But groups practicing the peyote religion were incorporated in 1918 as the Native American Church, and, by the 1960s, the church claimed 200,000 members.

BIBLIOGRAPHY. ROBERT H. LOWIE, *Indians of the Plains* (1963), is a short but authoritative general work. Accounts of particular tribes include the same author's *The Crow Indians* (1935); ALFRED W. BOWERS, *Mandan Social and Ceremonial Organization* (1950) and *Hidatsa Social and Ceremonial Organization* (1965); ALICE C. FLETCHER and FRANCIS LAFLESCHE, *The Omaha Tribe* (1911); JOHN C. EWERS, *The Blackfeet* (1958); GEORGE B. GRINNELL, *The Cheyenne Indians*, 2 vol. (1923); and E. WALLACE and E.A. HOEBEL, *The Comanches* (1952). GEORGE P. MURDOCK, *Ethnographic Bibliography of North America* (1960), covers the Plains in general and each tribe specifically. Records by early observers include JEAN LOUIS BERLANDIER, *The Indians of Texas in 1830*, ed. by JOHN C. EWERS (1969); and GEORGE CATLIN, *North American Indians*, 2 vol. (1926). GORDON MACGREGOR, *Warriors Without Weapons* (1946), is a study of the society and personality of the Teton Dakota under reservation conditions.

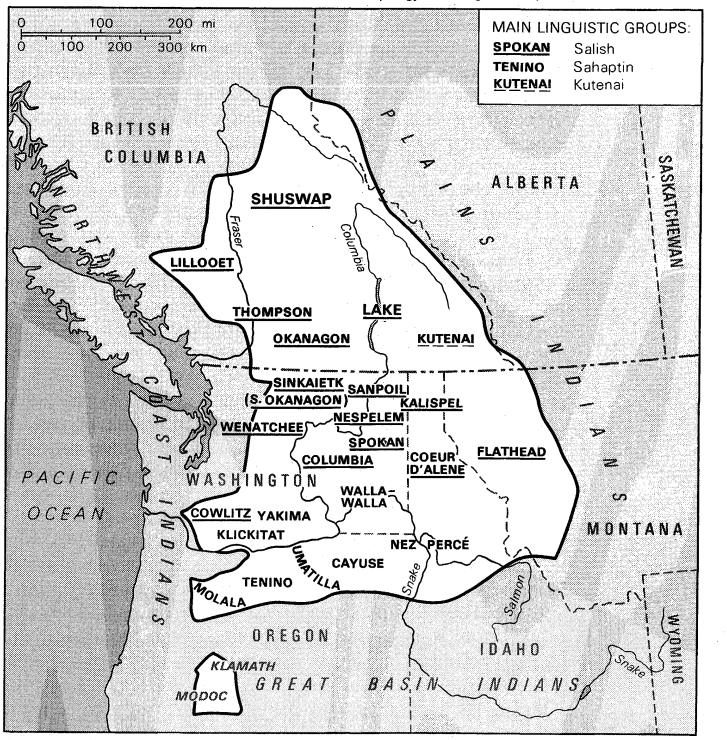
(R.F.-He.)

North American Plateau Indians

The North American Plateau is both a complex physiographic unit and a native cultural area. It is bounded on the west by the Canadian Coast Mountains and the Cascade Range, on the south by the Blue Mountains and the Salmon River (excepting a narrow corridor to California), on the east by the Rocky Mountains and the Lewis Range, and on the north by low extensions of the Rocky Mountains, such as the Cariboo Mountains. It may be defined as the drainage territory of the Columbia and Fraser rivers and as the high plateau between the main range of the Rocky Mountains and the coastal mountain system. In the south the natural area of the Plateau gradually merges with the Great Basin natural area: the boundaries between the corresponding culture areas are indeed also very imprecise. Previously, anthropologists included both culture areas as one, the Plateau.

The climate is a harsh, continental type. Temperatures range from −30° F (−34° C) in winter to 100° F (38° C) in summer. Precipitation is low, except in the mountainous areas, and forms a snow cover during the winter. There are three different provinces of vegetation, which correspond to three subcultures: the Middle Columbia area, a steppe of sagebrush and bunchgrass fringed by yellow pine on higher levels, is the territory of the Sahaptin groups and some Salish; the Upper Columbia area, a mainly wooded area with grassland in river valleys, is the home country of such Salish groups as the Okanagon and Flathead, and of the Kutenai; and the Fraser area, with a semi-open coniferous forest interspersed with dry grassland and a partly maritime flora, is the tribal ground of the northern Salish groups. The fauna is not rich, but there are deer and elk in the mountains and salmon and trout in the rivers.

From H. Driver *et al.*, *Indiana Universit Publications in Anthropology and Linguistics* (1953



Distribution of North American Plateau Indians.

Peoples and languages

The Indians of the Plateau belong mainly to four linguistic families: Salish, Kutenai, Sahaptin, and Klamath-Modoc (Lutuami). The majority of Plateau groups speak Salishan and Sahaptian. The Salish may be conveniently divided into Northern Plateau and Interior Salish (there are also Coast Salish on the Northwest Coast). To the

Northern Plateau group belong the Shuswap, Lillooet, and Thompson (Ntlakyapamuk) Indians; to the Interior group belong (mostly in the Upper Columbia area) the Okanagon (with Sinkaietk), the Lake (Senijextee), the Wenatchee, Sanpoil and Nespelim, the Spokan, Kalispel (with Pend d'Oreille), Coeur d'Alene, and Flathead. Some early works term all Salish "Flathead." The Sahaptin may be subdivided into three main groups: the Nez Percé, the Cayuse-Molala, and the Central Sahaptin (Umatilla, Yakima, Wallawalla, Tenino, and others).

TRADITIONAL CULTURE PATTERNS

The main characteristics of the Plateau cultures are best discernible against a historical background, for the Plateau cultural pattern was not stable. Opinion is divided as to whether its origins lay with the "desert culture" of arid Western North America, a primitive, seed-gathering culture, or with the "old Cordilleran culture" of the Plateau and North Pacific Coast, a culture with hunting, fishing, and gathering activities. It is certain, however, that the latter subsistence pattern predominated after 1500–1000 BC. By AD 1200–1300 the "classic" Plateau culture, characterized by permanent winter villages with semi-subterranean earth lodges along the main rivers and by summer camps with mat-covered conical lodges on the meadows, had emerged. There is evidence that the Plateau culture expanded as far south as the Snake River, including for some time Shoshone groups in Idaho. During the centuries that followed, the Plateau area was influenced by cultural elements from the highly specialized Northwest Coast culture. Thus, mat and plank houses, carving in wood and bone with animal motifs, and cremation and scaffold burial appeared. Part of this diffusion was possibly brought about by a Chinook group, the Wishram, who migrated from the coast into the Cascade Mountains.

"Classic" Plateau culture

During the 18th century there were influences from the east. The Shoshone had become mounted by this time and furnished their closest neighbours on the Plains and the Plateau with horses. White traders, from the beginning of the 19th century, testified that tribes such as the Nez Percé, Cayuse, Wallawalla, and Flathead had more horses than the tribes of the northern Plains.

Other elements of Plains culture came with the horse, particularly in the Middle Columbia area. Sahaptin Indians, for example, soon appeared in Plains beaded dresses and warbonnets and started to use tepees. Similar innovations occurred on the eastern periphery, in particular among the Flathead and the Kutenai. The northwestern Salishan groups, however, retained their original Plateau culture. Due to pressure from the Blackfoot, the Flathead and Kutenai had to withdraw from their home quarters on the plains of western Montana about 1800. They resettled in the intermontane valleys of the Rockies and from there made occasional buffalo hunts on the Plains in the company of other Plateau tribes, such as the Coeur d'Alene and Nez Percé.

The kind of military ethos found among the Plains Indians was not found uniformly among the Plateau Indians. The Thompson and Shuswap groups, and also the Sahaptin and Klamath further south, did make occasional war raids, dressed in elk hide or wooden slat armour and armed with bows and clubs. Other groups remained peaceful; the Flatheads, in particular, were well regarded by white settlers for courtesy, hospitality, honesty, and courage.

Social structures. Before the introduction of Plains culture, the village always formed the sociopolitical unit. The Thompson Indians, for example, had informal village meetings for decision making, and in matters of general interest the consent of all the villagers had to be obtained. The Sanpoil, on the other hand, had a more formal political structure: the village had a chief, a subchief, and a general assembly in which every adult had a vote (except for young men who were not married). The Nez Percé had a similar organization until the buffalo hunts on the Plains started. Each village had a chief whose office was hereditary, except in the case of poorly qualified sons. Sometimes groups from several villages came together at certain fishing sites or camas (edible lily) meadows, and on these occasions the leading men of the villages constituted an informal council. Early in the 19th century this organization was overruled when families from different villages joined to form bands for the autumn hunts on the Plains. The authority of the village chiefs lapsed as good hunters and fighters became band chiefs. As a result of pressure from missionary and governmental agencies, a tribal head chief was appointed in the 1840s, but he was unable to win any influence over the people. A truly tribal political organization existed among the Flathead, who had a head chief of great power and band chiefs under him. The head chief decided on matters of peace and war and was not bound by the recommendations of his council.

In many Plateau societies chiefs played a more prominent role than in Plains Indian culture, and they were also to a greater extent hereditary. But, although Sahaptin chiefs could exert their authority through whipping (perhaps a Spanish trait), social control was as a rule achieved through social pressure and public opinion. Nobody was coerced into following the advice of a chief or the decisions of a council meeting, and those who did not want to conform could move to another village or another band.

Kinship patterns

The simple, bilateral-descent system prevailed in typical Plateau groups. The average Plateau kin group consisted of the nuclear family and the closest relatives on the father's as well as the mother's side. This is the case among, for instance, the Tenino. Their kinship terminology reveals the close connection between family relatives of the same generation on both paternal and maternal sides, so that all female cousins are called by the same terms as those used for sisters. Marriages do not occur among first cousins (in distinction to the custom in clan-organized societies), and newly wedded couples may put up their residence with the father's or the mother's group. The Tenino also show a patterned kinship behaviour that has possibly existed in other Plateau groups, such as a "joking relationship" between a father's sister's husband and his wife's brother's child, and permitted sexual license between a man and his sister-in-law. All over the Plateau, marriages between one man and several wives (polygyny) were practiced, although they were not common.

It has been observed that kin term distinctions and ranked status distinctions tend to counteract each other: the Coast Salish have a ranked society with reduction of kinship terminology, whereas the majority of Plateau Salish have prevalence of descriptive kin terms and few status distinctions. Among the Sinkaietk or Southern Okanagon, chiefs were hereditary and the most important persons in the tribe in regard to moral influence, for the chief and his family were supposed to exemplify the virtues of the group. He was, on the other hand, not necessarily the wealthiest man in his group, although he was economically supported by his people. The chief had a female relative among his advisers. Such highly respected women also existed in other groups, such as the Coeur d'Alene, and bear witness of the independence of women in Plateau society (excepting the Plains-influenced Kutenai and Flathead). Although marked off as hereditary in his office, the Plateau chief did not separate himself from his group. The general spirit was one of equality and personal autonomy, particularly among the Interior Salish. The Northern Plateau Salish, however, and several other groups kept slaves, as did the Indians on the Northwest Coast, and traded them between each other. The tribes on the eastern fringe who shared the Plains values had a rank of honoured warriors and war chiefs.

The life cycle of the individual was marked by fixed ritual acts that opened the gateway to the different social roles he had to enact. One could say that it started before birth. Among the Sinkaietk, for example, a pregnant woman was not supposed to give birth to her child in her regular home but in a menstrual lodge or another separate lodge. The newborn baby spent its day strapped in a cradle of the flat board type. At the age of one the child was ceremonially conferred a name from the wealth of names in the family. The training of the child was left to the mother and grandmother, but even as a small boy a

Sinkaietk could accompany his father on fishing and small-game hunting trips, while the little girls helped their mothers about the house and gathered roots in the fields. Grandparents saw to it that the child was hardened by such practices as bathing in cold streams. Disobedience was rare but could sometimes result in the child being whipped.

Puberty rites

At puberty the boy was sent out to spend some days fasting on a mountain top and probably to receive a blessing vision from some spirit. Upon returning to the community, he took his place among the adult men. The girl who had her first menstruation was secluded in a menstrual lodge some distance from the village. Her hair was bound up in rolls, and she was only allowed to touch it with a small comb. Her face was painted red or yellow, and she wore undecorated clothing. She was not allowed to drink directly from a well but had to use a drinking tube, and she cleansed herself after the flow in a sweathouse. After a long time—one or several months—she finished her seclusion with prayers in the evening on a hill. Then she returned to the village, a full-grown woman.

Marriage was an entirely informal affair, as was divorce; a woman who was tired of her husband or had been expelled by him returned to her parents if they were alive. She could then remarry if she wished.

Two forms of burial predominated in the Plateau area, pit burials and rockslide burials. The pit burials took place in sand or gravel near the river banks and were often marked with piles of boulders. The rockslide burials were also located close to the river flats, with a cedar stake as a marker. Some cremation burials occurred in the Yakima Valley and at The Dalles and also in the Lillooet-Thompson area. The bereaved had to observe certain taboos, and a widow was supposed to dress poorly and wail at the grave, sometimes for as long a period as a year. There are reports that the house where the death occurred was torn down so that the dead person would not reappear there.

Economic life. The Plateau villages were generally located on waterways and particularly at rapids and other places where fish were abundant during the winter season. Each village had an upland for hunting; in contradistinction to the fishing localities, these uplands were mostly open for people from other villages as well. There were also permanent or semipermanent summer camps for hunting and root gathering in mountain valleys. River villages were permanent winter quarters and could at least temporarily lodge several hundred people. A Kalispel village, for example, numbered 300–400 and a Yakima village as many as 2,000.

Winter dwellings were of two main types, the semi-subterranean earth lodge and the mat-covered surface house. The latter was apparently more recent and existed only in the southern Plateau, where it had replaced an older earth lodge. It was replaced in its turn by the Plains Indian tepee. The average earth lodge was circular, with a pit 4–6 feet (1–2 metres) deep and a diameter of 10–40 feet (3–12 metres). The roof was conical or flat and was supported by leaning poles fastened to some central posts. The smoke hole in the top was also the entrance, the floor being reached by an inside ladder or notched log. The other type of dwelling was formed of two walls of varying length leaning together and covered with tule mats. It was a "longhouse" with a series of hearths in the middle, each one of them shared by two families, one on each side. During the summer people housed in conical mat lodges of small size or in simple windbreaks.

Food-gathering techniques

Fishing was the most important source of food. The Plateau Indians used one- or three-pronged fish spears, traps, and nets when taking their staple fish—eels, suckers, trout, and especially salmon. Large quantities of fish were dried on elevated wooden racks or kept in storage pits and preserved for winter consumption. Roots were dug with digging sticks provided with cross handles of antler or wood. The main root was the camas bulb (*Camassia esculenta*), but bitterroot, onions, wild carrots, and parsnips were also gathered. They were then cooked in earth ovens heated by hot stones. Berries—serviceberries, huckleberries, blueberries—were harvested as well. Hunting occasionally played an important role, even in the winter. Equipped with bow and arrows and perhaps a short spear, the Indian hunted deer first of all but also bear and caribou. In the winter he tracked the animals on long and narrow snowshoes; in the summer he could use a canoe—a dugout in the southern Plateau, a dugout or a bark canoe in the northern Plateau.

In historical times all Plateau peoples used tailored skin garments of the type well-known from the Plains. In prehistoric days both sexes wore a bark breechclout or apron and a twined bark poncho falling a little below the waist. During the cold season men wrapped their legs with fur, and women had leggings of hemp. Rabbit-fur robes or other skin robes were worn in winter. Sahaptin women had twined basket hats, whereas men everywhere had headbands; caps of fur and feathered headdresses appeared with the Plains influences. Both sexes braided their hair. The Chinook practiced flattening of the infant's head as sign of free birth. Curiously, the Flathead never shared this custom.

The village community owned the land, in particular the fishing sites. Household tools, weapons, traps and snares, and similar items were the property of individuals, except for larger weirs that were communal property. Food resources were in most places distributed according to needs. A more restricted system prevailed on the northern Plateau, where gift-giving ceremonies occurred, reminiscent of the potlatches of the Northwest Coast Indians: after some days of games and contests, gifts were distributed to the guests, who in their turn reciprocally handed over presents to their hosts. Although possessions were valued in many parts of the Plateau, the Klamath paid greater attention to them than any other group and held wealthy persons in great esteem. This value orientation, most probably derived from the Northwest Coast, contrasted with the general Plateau pattern of equality and sharing of necessities.

Belief and aesthetic systems. Religion was, like the rest of the culture, closely intertwined with Plateau ecology. In many ways religious beliefs echoed North American religions in general: there was a "great spirit," among the Okanagon conceived of as a bearded white man, and there were spirits of the atmosphere (winds, thunder, etc.) and a host of zoomorphic lesser spirits serving as personal guardian spirits.

Religious rituals

The main rituals were the guardian-spirit quest, the firstling rites, and the winter dance. The guardian-spirit quest was compulsory for boys and recommended for girls and was usually performed in connection with the puberty ceremony. The spirits who granted their blessings in lonely places were very specialized. Some made their clients into hunters, others into warriors or medicine men. Both boys and girls, but preferably the former, could become medicine men. Medicine men were much feared and sometimes very wealthy. They cured diseases by extracting the bad spirit that had entered the patient's body, and on the Northern Plateau they brought back souls that had been stolen by the dead, describing their feats in a dramatic pantomime.

The firstling rites concentrated on the first salmon or berries (roots, fruits) that had been caught or gathered during the summer season. The first salmon ceremony celebrated the arrival of the salmon run with the ritual cutting and eating of the first fish and the ritual throwing of the bones back into the water, in this way ensuring a good return next year. Some Salish had a "salmon chief" who surveyed the rituals. The Okanagon, Thompson, and Lillooet had similar rites for the first berries, which were eaten ceremonially, whereas they lacked the salmon ritual.

The winter or spirit dance, finally, was a ceremonial meeting at which participants personified their respective guardian spirits. The dramatic performances and the songs were, among the Nez Percé, thought to bring warm weather, plentiful game, and successful hunts.

Plateau mythology and folklore revolved around the culture hero and transformer, mostly the Coyote but in some places the Bluejay or another mythical personage. He is a beloved character in the stories, creator and trick-

ster at the same time. The Coyote cycle is well-known from adjacent areas as well.

There is nothing distinctive about Plateau art. On the contrary, most art historians divide it into a western branch, peripheral to the Northwest Coast, and an eastern branch, peripheral to the Plains. Plastic art is on the whole very rare, except in the vicinity of the Northwest Coast. Decorative art consisted of pictographic designs with a symbolic content, referring to supernatural beings and cosmic things. The same division between east and west characterizes musical styles.

DEVELOPMENTS SINCE THE COMING OF EUROPEANS

The preceding description of traditional Plateau culture demonstrates that the culture was neither static nor unitary but changed with time and place. The most dynamic development was introduced when the first impulses from white civilization penetrated the area: the coming of the horse in the beginning of the 18th century, the appearance of epidemics from 1780 onward, and the arrival of eschatological ideas, adapted in the Prophet Dance from perhaps the same time. The latter, which was the origin of the famous Ghost Dance, was a mixture of aboriginal and Christian elements: by dancing like the dead in the other world the Indians thought they could hasten the renewal of the world and the return of the dead. The Prophet Dance seems to have been a reaction against the increasing disruption of traditional culture through the new influences.

Early in the 19th century the fur trade brought Indian and white trappers from the east into the country, particularly to the northern Plateau. Catholic Iroquois trappers propagated Christian ideas among the Flathead, who thereafter visited St. Louis to call on missionaries. The great invasion of white settlers and gold seekers in the 1850s and 1860s and the ensuing Indian wars (of which the Nez Percé War of 1877 is the most famous) resulted in the reduction of Indian territories, the creation of a series of small reservations, and the isolation and deprivation of the Indians in "white" surroundings. Only during recent decades have Indians tried to take part in modern development outside their reservations.

The blending of aboriginal and white cultures has accelerated with every year and produced a variety of mixed cultures on the reservations, some more conservative, others more Europeanized. The Kutenai, for instance, have turned into ranchers or ranch hands during the warm season but use their fishing traps during the winter, a seasonal pattern that in a way conforms with the old culture. The Nez Percé, on the other hand, have at least partly become farmers. Their cultural assimilation has been furthered by political and religious factionalism. Summarily, it may be said that the Plateau Indians have retained their group feelings, part of their old economics, and in places much of their religion, whereas technology and material culture have long been characteristic of white poverty levels.

BIBLIOGRAPHY. The archaeological background has been described by E.H. SWANSON, *The Emergence of Plateau Culture* (1962); and B.R. BUTLER, *The Old Cordilleran Culture in the Pacific Northwest* (1961). A noteworthy paper is R.D. DAUGHERTY, *Archaeology of the Lind Coulee Site, Washington* (1956). General ethnological perspectives have been presented in two papers by V.F. RAY, *Cultural Relations in the Plateau of Northwestern America* (1939) and *Culture Element Distributions*, pt. 22, *Plateau* (1942). There are several good monographs on single tribes: J.A. TEIT, *The Thompson Indians of British Columbia* (1900); H.J. SPINDEN, *The Nez Percé Indians* (1908); LESLIE SPIER, *Klamath Ethnography* (1930); V.F. RAY, *The Sanpoil and Nespelem: Salishan Peoples of Northeastern Washington* (1932); and the work edited by LESLIE SPIER, *The Sinkaietk or Southern Okanagon of Washington* (1938). Demographic, linguistic, and sociopolitical analyses may be found in V.F. RAY, "Native Villages and Groupings of the Columbia Basin," *Pacif. NW. Q.*, 27:99–152 (1936); T.R. GARTH, "Early Nineteenth Century Tribal Relations in the Columbia Plateau," *SWest. J. Anthrop.*, 20:43–57 (1964); and W.W. ELMENDORF, "Linguistic and Geographic Relations in the Northern Plateau Area," *ibid.*, 21:63–78 (1965). Kinship systems are discussed in M. JACOBS, "Northern Sahaptin Kinship Terms," *Am. Anthrop.*, 34:688–693 (1932); and W.W. ELMENDORF, "System Change in Salish Kinship Terminologies," *SWest. J. Anthrop.*, 17:365–382 (1961). LESLIE SPIER, *The Prophet Dance of the Northwest and Its Derivatives: The Source of the Ghost Dance* (1935), has become a classic, but must now be checked against D.E. WALKER, JR., "New Light on the Prophet Dance Controversy," *Ethnohistory*, 16:245–256 (1969). Modern culture contact problems are dealt with in a skilful way in D.E. WALKER, *Conflict and Schism in Nez Percé Acculturation* (1968).

(A.G.B.H.)

North Carolina

One of the 13 original states of the United States, North Carolina lies on the Atlantic coast midway between New York and Florida. In recent decades it has become the leading industrial state of the South, and it is also, after Florida, the South's most populous state. More than one-half of the nearly 5,100,000 inhabitants recorded in 1970, however, lived outside urbanized communities, giving it one of the largest rural populations in the nation.

General character of the state

North Carolina's beginnings are tied closely to the earliest attempts at English colonization of the New World. Roanoke Island in the northeast, a part of the heavily indented and island-fringed coast, was the site of the famous Lost Colony that vanished sometime after the original landing in 1587. This eastern region today retains much of the flavour of colonial life, while the higher Piedmont region centred around the capital, Raleigh, has become the industrial and population hub, and the mountains of the west remain the focus of a lively folk culture and the home of the largest group of American Indians east of the Mississippi River.

Bounded on the north by Virginia, on the east by the Atlantic Ocean, on the south by South Carolina and Georgia, and on the west by Tennessee, North Carolina has an area of 52,586 square miles (136,198 square kilometres). Its 3,788 square miles of inland water, the fifth largest such area of any state, are concentrated both in the extensive marshlands of the coastal tidewater and in the lakes of the Piedmont and Appalachian regions. These three physical regions are related to major diversities in life styles among the people of the state, almost creating three North Carolinas within the common boundaries. (For information on related topics, see the articles UNITED STATES; UNITED STATES, HISTORY OF THE; and NORTH AMERICA.)

THE HISTORY OF NORTH CAROLINA

The proprietary and royal colony. Following the abortive attempts of Sir Walter Raleigh and others to colonize the coastal regions in the 1580s under patents from Queen Elizabeth I, the region remained Indian territory for many decades. A grant by King Charles I in 1629 for the lands south of Virginia brought the term Carolana into being, but no permanent settlement was made until farmers from Virginia moved into the Albemarle Sound area of the northeast in the 1650s. This resulted in a grant from Charles II in 1663 that created Carolina, but for years the settlers resisted the ineffective government imposed by the proprietors in England. Between 1712 and 1729, the separate province of "North Carolina" was ruled by a deputy dispatched from Charleston, which had become the centre of proprietary government. The boundaries between North and South Carolina were agreed upon in 1735 but not surveyed until 1815.

Difficulties of settlement and government

North Carolina's growth was hampered by restrictions on shipping imposed by Virginia on its already significant tobacco crop; by economic and religious quarrels with absentee proprietors that led to rebellions in 1677 and 1708; by war with the Tuscarora Indians (1711–13); and by coastal piracy involving Edward Teach (Blackbeard) and others. Unlike other colonies, which had grown up around the coastal towns that represented the first settlements, North Carolina had no town until Bath was settled by French Huguenots from Virginia after 1700. By 1729, when the colony came under royal rule, several other communities had been founded.

The decades of royal rule saw a turnabout in the colony's fortunes. The population rose rapidly, settlement

spread across the Piedmont, and the wealth and quality of life expanded toward that of the other colonies. A large slave population maintained an agricultural economy based on tobacco, rice, and naval stores from the extensive pine forests. Prior to the American Revolution, the seeds of an intense east-west hostility had flowered into several insurrections, but joint antipathy to British rule united North Carolinians and forced the flight of the royal governor in 1775.

Statehood. The Revolution in North Carolina comprised not only a miniature civil war involving the many Tories in the new state but also suppression of Cherokee uprisings in the west. Much of the state's energy was directed to resolving the conflicting interests of the eastern counties and those of the west until constitutional reforms in 1835 broke the dominance of the east. A period of great economic and social progress first under the Whigs and continued after 1850 by the Democrats was slowed by the increasing furor over slavery and ended by the Civil War.

The Civil War and its aftermath

Unlike South Carolina, whose strident proslavery voices led the South into secession, North Carolina left the Union reluctantly, seeking compromise to the last moment. Once committed, however, the state fought with the Confederacy and with it experienced the ignominy of defeat and the years of political corruption and social instability that characterized the postwar Reconstruction throughout the South. The "Bourbon Democrats" who controlled the state after readmission to the Union were oriented to the growing railroad and industrial interests, and they largely ignored the small farmer. Constitutional amendments in 1900 virtually disfranchised the state's Negro population.

North Carolina in the 20th century has been a part of the national experience of changing economic cycles and, especially since the 1950s, of racial tensions. The state has benefitted, however, from its rise to a ranking among the top one-third of industrial states, a factor that in terms of statewide wealth has tended to balance its huge rural population and agricultural activity; and from the concern with higher education as an integral part of the life of its citizens that has characterized it since its early years as a state.

THE NATURAL AND HUMAN LANDSCAPE

The physical and biological—as well as the human—characteristics of North Carolina's Coastal Plain (or Tidewater), Piedmont, and Appalachian Mountain regions are quite different. As the land reaches westward from sea level, it rises gradually to the Fall Line, a zone some 30 miles in width separating Coastal Plain from Piedmont. In the latter, the topography becomes irregular and rises about five feet a mile to the base of the Appalachians, a distance of about 140 miles. The mountains, many reaching well over 6,000 feet, have a worn, rounded appearance, reflecting their older geological origin than the rugged peaks of the American West. Mt. Mitchell, rising to 6,684 feet (2,037 metres), is the highest peak east of the Mississippi.

The state's regions. *The Coastal Plain.* Comprising some 45 percent of the state, the Coastal Plain divides into a gently rolling, well-drained interior and a swampy tidewater portion close to the coastline. The latter region was the first to be explored and settled. A long chain of islands, the Outer Banks, extends from Virginia to South Carolina, generally covered with sand dunes from a few feet to over 100 feet in height. Three capes—Cape Hatteras, Cape Lookout, and Cape Fear—jut out into the ocean in an area known as the "Graveyard of the Atlantic," a reference to the many ships that have gone down in the dangerous offshore waters. The entire area averages less than 20 feet above sea level, but in spite of the watery environment only small craft navigation is possible because of silting and shallow sounds and estuaries. The Intercoastal Waterway threads its way between the Outer Banks and the mainland on its way from New Jersey to the Gulf of Mexico. The inner Coastal Plain extends from 120 to 140 miles westward to the Piedmont.

Eastern North Carolina has been the citadel of the state's history since Raleigh's grand dream of colonization came to so mysterious an end. Close to Roanoke Island, are the sand dunes of Kitty Hawk, over which in 1903 Wilbur and Orville Wright ushered in the age of powered flight. Legends tell of fabulous pirate treasure buried beneath the dunes of the Outer Banks, off whose shores rusting smokestacks, masts, and boilers protrude from the water, testimony to the more than 2,000 ships that have gone down. Nearby Nags Head got its name, according to tradition, because unscrupulous settlers tied lanterns to horse's necks and drove them along the coast to lure unsuspecting sea captains to the reefs. On Ocracoke Island, so named by Blackbeard, visitors are astonished at the Elizabethan English of the residents, for whom "high tide" is "hoigh toide."

Diverse ways of life from coast to mountains

Further south in New Bern, the state's second-oldest town and one named by its Swiss settlers, is Tryon Palace, a restored palace and garden that has been called the most beautiful building in the colonial Americas. Along the southern coast, fishermen set out to battle the large deepwater fish of the close-by Gulf Stream, and in Edenton memories survive of the colonial ladies who held one of the first tea parties to protest the duties imposed by the British. Morehead City and Wilmington are the state's two deepwater ports, both significant in world trade, while in the area several major military installations add to the state's economic life.

The Piedmont Plateau. Containing about 38 percent of the state's area, the North Carolina Piedmont is a region of rolling, forested hills. The prominent ridges and hills of the eastern Piedmont may be the remains of an ancient mountain chain that paralleled the Appalachians, from which numerous spurs extend into the western Piedmont. The area is well drained by rivers flowing into the Coastal Plain or South Carolina. Dams on the Catawba and Yadkin rivers are important sources of hydroelectric power.

This region is a prime symbol of the "New South," in which modern industry has largely replaced the traditional agriculture. The pulsating main street of industry runs in a sweeping crescent westward and southward from Raleigh to below Charlotte, the state's largest city. Such cities as Durham, Greensboro, and Winston-Salem have made North Carolina the capital of the nation's tobacco industry. The colleges and universities that have been so influential in the state's history are centred in this region.

In spite of industry, the many antebellum homes in these cities maintain an aura of serenity, and cotton and tobacco fields are still found close to the city limits. The many lakes and the upper reaches of the rivers provide quiet havens for fishing and camping, and in many small towns the amply stocked general stores still serve the rural populations. Under the city streets of Charlotte—described by Lord Cornwallis, the English general of Revolutionary fame, as "a piddlin little place"—are the traces of early gold-mine shafts, which once produced many tons of the precious metal.

The mountains. The mountain region comprises a highly desiccated intermontane plateau bounded by two ranges of the southern Appalachians. On the east is the Blue Ridge, which rises steeply from the Piedmont to peaks of 3,000 to 4,000 feet and a few of 6,000 or more. In the far west, the Unaka Mountains contain the Great Smoky Mountains that roll westward into Tennessee. This region is divided into several cross ridges and a number of smaller plateaus and basins. One of the chief ridges is made up of the Black Mountain group. In all, 43 peaks rise over 6,000 feet and 80 over 5,000 in western North Carolina.

In North Carolina's mountains, ways of life change slowly. Many communities, relatively isolated since the early history of the state, have become self-sufficient. Wood carving, basketry, needlework, rug and bedspread making, and ceramics are among the many cottage industries whose crafts have been passed down through the generations. Today the serenity is beginning to be broken, however, both in the mountains and in the resort centre of Asheville. Winter and summer sports are becoming popular on the slopes, and the Pisgah National Forest is

Traditional cultures of the isolated communities

among the areas that attract a growing number of tourists and campers. One of the world's largest satellite tracking stations is located at Rosman.

Climate. North Carolina's climate ranges from that of the mountain region, which experiences a medium continental type, although summers are cooler and rainfall heavier, to the subtropical conditions of the state's southeastern corner. The growing season ranges from 275 days along the coast to 175 days in the mountains. Mean annual temperature ranges from 66° F (19° C) in the eastern region, to 60° F (16° C) in the central, and 55° F (13° C) in the mountains. July and August are the wettest months, and October and November are the driest. Annual rainfall varies from 46 to 54 inches (117 to 137 centimetres) on the coast, 44 to 50 inches (112 to 127 centimetres) in the Piedmont, and 40 to 80 inches (100 to 200 centimetres) in the mountains. Severe storms are rare and heavy snow infrequent. Occasional hurricanes occur along the coast, and there have been tornadoes inland.

Plant and animal life. Soil and vegetation are quite different over the state because of the geographical and climatic differences of the three main regions. Trees that once covered the landscape as dense forests have been cut and burned and now cover only 56 percent of the state. Erosion and leaching of the soils necessitates large amounts of lime to neutralize acidity and fertilizers to replace the leached nutrients.

A greater variety of plant life is to be found in North Carolina than in any other state in eastern North America. There are many species of hardwood trees. Sub-Arctic spruce and balsam fir are found in the mountains, and the subtropical palmetto and the carnivorous Venus's-flytrap in the south coastal area.

The common fauna of North America, including rabbits, squirrels, raccoon, oppossum, deer, and also bear and wildcats, are found within the state. The commonest birds are the cardinal, wren, mockingbird, chickadee, and many varieties of woodpecker and warbler.

Inland-water fish such as bluegills, crappies, bass, and sunfish are common. Brook and rainbow trout are found in the mountains.

THE PEOPLE OF NORTH CAROLINA

Archaeologists have found traces of human habitation in the state that date back some 16,000 years. It is estimated that when the first white explorers arrived there were between 35,000 and 50,000 Indians in the region. The Indians were finally crushed in the late 1830s with the forced removal of the Cherokee to lands west of the Mississippi, recorded in history as the "Trail of Tears" (1838–39). In the early 1970s only approximately 30,000 live in the state, the largest group east of the Mississippi River and the fifth largest group in the United States.

Early settlement. The coast was first explored by the French, Spanish, and English. The first English colony failed, and the fate of its members is still a mystery. Settlers came into North Carolina in the 1650s from the English colony at Jamestown, Virginia. Others came from Philadelphia and down through Virginia on the great wagon road from Pennsylvania through the Great Appalachian Valley into the Piedmont. Many came by sea lanes from Europe, all yearning for a plot of land and freedom from rigid class and religious restrictions.

Disease, poor roads, the closing of Currituck Inlet by silt, abandonment of worn-out fields, and limited trade with England slowed early settlement. The population in 1694 was 3,000, living mostly in the Albemarle Sound area. By 1729 the population was perhaps 30,000. The first United States census of 1790 showed the state had grown to over 390,000 persons.

Ethnic and religious backgrounds of early settlers

The early North Carolinians were a heterogeneous group, representing a variety of religious faiths, economic and social classes, and nationalities. The Anglican Church was established by law in the early 18th century, but there were also Presbyterians, Quakers, Moravians, Lutherans, Reformed, Baptists, Methodists, and a small number of Jews. Nationalities represented were English, Scottish, Irish, Welsh, French, German, and others.

The Negro component of the early North Carolina population, brought in from Virginia and South Carolina, was most important, as it furnished the labour for clearing and working the land and producing the naval stores of tar, pitch, rosin, and turpentine. The total slave population in 1790 was more than 100,000, one-third of the total population. Approximately 5,000 Negroes were free men. The labour-demanding crops of rice, indigo, tobacco, and cotton accounted for the spread of slavery in the state, especially after the perfection of the cotton gin.

Demographic trends. The 1970 census showed North Carolina to be the 12th most populous state in the nation. The population density was 104.1 persons per square mile, compared to 114.9 for the South Atlantic region, and 57.5 for the United States as a whole. Racially, the population is 76.8 percent white and 23.2 percent nonwhite. The black population declined from 27.5 percent in 1940 to 22.2 percent in 1970, mainly as a result of blacks seeking employment in the northern and western sections of the United States and because of the continued mechanization of agriculture. The migration, in and out, of whites has tended to equalize.

The state's white population has a considerably lower death rate than does the nonwhite, reflecting the wide variations between races in socio-economic status. Degenerative rather than infectious diseases have become the more important killers. Whites enjoy far greater protection from preventable diseases than do the nonwhites.

North Carolina's population is more rural than urban, despite the large industrial employment. The reason for this is that many industrial plants are located in small towns, and workers tend to commute long distances and live in rural areas. Rapid urbanization and the persistence of extremely rural areas accentuate the demographic contrasts in the state. In the eastern region, rurality is typified by tenancy and sharecropping, while in the mountain region the traditional folk rurality is continued on farms operated on a subsistence level.

THE STATE'S ECONOMY

North Carolina's economy depends on manufacturing and agriculture, but tourism is gaining in importance.

Components. North Carolina is endowed with numerous resources that are of great value to manufacturers. The state has one of the nation's largest known phosphate reserves; other important minerals include kaolin, mica, feldspar, granite, copper, limestone, marble, marl, olivine, talc, sand, gravel, and shale. Forestry and fishing are other important sources of income.

In annual industrial output, the state ranks 12th in the nation. Some 20 percent of the labour force is employed in the metalworking, electronic, chemical, paper and paper products, plastics, and food-processing industries.

The state ranks first nationally in farm population and second in the number of farms. The principal crops are tobacco, corn, grain, soybeans, peanuts, and hay. Home consumption of crops and livestock is higher than in any other state. Forest products are used for furniture and as a source of pulp for paper. An active reforestation program has resulted in a growth of forest reserves.

Tourism brought to North Carolina an estimated income of over $800,000,000 annually in the early 1970s. The industry has a diversified base, including the attractions of both ocean and mountains as well as the memorials to the state's past.

Transportation. Geographically, the state is one day's trucking time to New York or to the rapidly expanding Florida market. More long-distance interstate motor carriers are domiciled in the state than in any other. Eight commercial airlines, operating out of 14 airports, serve the state.

Marine traffic

North Carolina has two Atlantic gateways to world markets. Modern ports are found at Wilmington and Morehead City, both of which are equipped to handle any type of cargo. More than 50 steamship lines make calls, and both ports are served by rail and motor transport. In addition, 20 feeder ports are equipped to handle barges and small ships.

There are some 4,300 miles of rail line in the state, serving 92 of the 100 counties. Rail rates are consider-

ably lower than in the rest of the nation. The principal motor passenger carriers transport more than 3,000,000 passengers each year.

ADMINISTRATION AND SOCIAL CONDITIONS

Government. The structure of the government of North Carolina is based on constitutions of 1776 and 1868. There have since been numerous amendments.

State level. Administration of the state is supervised by elected executives, including the governor, lieutenant governor, the heads of numerous state agencies, all of whom serve four-year terms. The governor has great appointive powers but no veto over legislation—the only governor in the nation who lacks this power. All state fiscal and management agencies are consolidated in one department, the Department of Administration. The court system has, as its base, 30 district courts that deal with less serious civil and criminal cases. The superior courts, of which there is one for each judicial district, take the more serious criminal and civil cases. Superior court judges are elected in general elections for a term of eight years. There are eight special superior court judges appointed by the governor for four-year terms.

Above the superior courts are the Court of Appeals and the Supreme Court. The latter is the highest state court; it has seven justices elected for eight-year terms. The Court of Appeals was set up in 1967 to help relieve the North Carolina Supreme Court by hearing the less important cases. It has nine judges, all of whom are elected for eight-year terms.

Local government. North Carolina is divided into 100 counties. The county governments act for the state in providing education, health, and welfare services. Locally elected officials include county commissioners, the sheriff, the register of deeds, the clerk of the superior court, and the school board. There may be other elected officials in counties having larger populations. Town and city governments also provide local services.

The social milieu. In spite of its progressive and expanding economy, North Carolina remains below national averages in such areas as the per capita income of its population and its expenditures for social services. In addition, a wide economic gulf separates the black and white communities of the state.

Education. The public school system, supported by the state since 1933, has been improving steadily for several decades, though it is still below national levels. Other problems include a relatively low salary scale for teachers and a racial integration that is far from complete.

Institutions of higher education

In higher education, however, North Carolina has a number of institutions of national standing. The public university system is headed by the University of North Carolina whose main campus is in Chapel Hill; opened in 1795, it is one of the oldest state universities in the nation. Other campuses are located in Greensboro, Asheville, Charlotte, and Wilmington. The state also has 9 state universities, 13 community colleges, and more than 35 technical institutes in its system. These combined facilities make it the fifth largest university system in the nation. Among the dozens of private institutions around the state, most of them supported by various Protestant denominations, Duke University (established 1924; formerly Trinity College) in Durham is noted for its undergraduate and postgraduate programs.

Health and welfare services. State-funded hospitals cover a number of specialized areas such as crippled children, alcoholism, retardation and mental illness, cerebral palsy, and tuberculosis. An effective public health program has been in operation since 1877, and each county has a local health department. State aid is provided also to the aged or disabled, to families with dependent children, and for various counselling and other social service programs.

Race relations. In spite of continuing disparities between white and black living conditions, the black has had an impressive rise during the 20th century in the arts, sports, business, education, and politics. In the early 1970s, Chapel Hill had a black mayor, Charlotte a black mayor *pro tem*. The major areas of polarization remain in educational and religious institutions.

Services. Per capita direct general expenditure by the state and local governments in such areas as education, hospitals, sanitation, police, and welfare is $390, compared to an average in the United States of $578.

CULTURAL LIFE

The fine arts. An arts council was established in 1964 to assist in bringing the highest obtainable quality in the arts to the greatest number of people in the state and to expand the role of the arts. The council sponsors numerous projects, dance workshops and tours dominating its activity.

The North Carolina Museum of Art was the first in the country to be established by a state and to be supported mainly by state funds. The museum sends out travelling exhibits to schools, libraries, civic clubs, and other museums throughout the state. The North Carolina State Art Society donates funds for the purchase of additional works of art.

The North Carolina Symphony has the distinction of being the first state symphony in the country. The orchestra, which has 65 members, tours the state from October through April. More than half of the performances are free matinees for children. The North Carolina Symphony Society tries, through membership drives, to raise enough money to sponsor local adult and children's concerts and to stimulate interest in music.

The folk arts and pageantry. It is in the field of the folk arts and of historical pageantry that North Carolina excels. The many cottage industries of the western mountains combine with those of the coastal water-oriented communities to offer some of the richest folk culture in the United States. Summerlong outdoor pageants are held in Manteo on Roanoke Island, where the drama *The Lost Colony* revives the colonizing escapades of Sir Walter Raleigh in the court of Elizabeth and on the soil of Roanoke itself; in Boone, where *The Horn of the West* recreates such characters as Daniel Boone; and in Cherokee, where *Unto These Hills* is played by the descendants of the Cherokee Indians upon whose history the saga is based. A major force in the cultural life is the Carolina Playmakers. The group was founded in 1918 at the University of North Carolina by Frederick Koch, an advocate of folk drama who had a strong influence on numerous playwrights and movements in American theatre.

Historical dramas

PROSPECTS

North Carolina is changing rapidly as a result of expansion of urban areas, slum clearance, new housing, schools, recreational facilities, and industrial plants.

Economically, the state has two major problems. Tobacco is the state's main crop, and it is possible that its consumption may decline rapidly as a result of the national antismoking campaign. Textiles, also a very important segment of the economy, have been hurt by imports, and a number of mills have closed. That there is room for improvement in the performance of the administration is suggested by the fact that a national survey made in 1970 ranked the General Assembly 47th of the 50 states in performance.

BIBLIOGRAPHY

History: H.T. LEFLER and A.R. NEWSOME, *North Carolina: The History of a Southern State*, rev. ed. (1963); H.T. LEFLER, *A Guide to the Study and Reading of North Carolina History*, 3rd ed. rev. (1969).

Geography: R.E. LONSDALE, *Atlas of North Carolina* (1967), includes maps of the state showing points of local and historical interest.

Government publications: NORTH CAROLINA CROP REPORTING SERVICE, *North Carolina Agricultural Statistics* (annual), a report on crops, fruits, vegetables, nuts, livestock, and prices paid and received in 100 counties; EMPLOYMENT SECURITY COMMISSION OF NORTH CAROLINA, *General Economic Summary of North Carolina; Biennial Report* of the NORTH CAROLINA BOARD OF HIGHER EDUCATION; *North Carolina Manual*, issued annually by the Secretary of State, gives names, departments, and duties of various governmental agencies and officials; NORTH CAROLINA WILDLIFE RESOURCES COMMISSION, *Wildlife in North Carolina* (monthly).

(Pe.S.G.)

North Dakota

Officially classed as one of the seven west north central states of the United States, North Dakota is a land of clear skies, seemingly endless grain farms, and vast cattle ranches. North Dakota is even more rural, more agricultural, and more sparsely populated than the six other states of the region, and has less manufacturing. Its terrain rises through three regions from east to west, incorporating parts of the two major physiographic provinces that separate the Appalachian and the Rocky Mountain systems.

The state's 70,665 square miles (183,022 square kilometres) according to the 1970 census had only 617,761 inhabitants, showing a decrease of 2.3 percent from 1960. The largest city, Fargo, had fewer than 55,000 inhabitants, and Bismarck, the centrally located capital, some 35,000.

Among the last regions of the American frontier to be settled, the area admitted in 1889 as the state of North Dakota had experienced comparatively little of the fighting, lawlessness, and gold-rush excitement that give other frontier areas a colourful or lurid history. Instead the region had developed first as the home of hunting and farming Indian peoples, then as a trade hinterland for white fur traders and for steamboats working the upper Missouri from St. Louis, and last as a rich farming land for settlers. The cool, subhumid climate of its location made it ideal for spring wheat and for cattle ranching. With white settlement, the area inevitably developed a way of life dependent on outside centres of population, industry, and economic power. With adaptation to the environment, however, the people of North Dakota developed also constructive reactions to the circumstances that made their state dependent. (For information on related topics, see the articles UNITED STATES; UNITED STATES, HISTORY OF THE; NORTH AMERICA; and GREAT PLAINS.)

THE HISTORY OF NORTH DAKOTA

The recorded history of North Dakota falls into three periods: the period of Indian trade, from about 1738 to 1871; the period of white settlement, from 1871 to 1915; and the period of adaptation, since 1915.

Explorers and traders. Although European goods were traded among the Indian peoples before his arrival, the first-known white visitor to North Dakota was Pierre Gaultier de Varennes, sieur de La Vérendrye, a native of Canada who visited a cluster of earthen-lodge villages near present-day Bismarck in 1738. Traders from Hudson Bay and Montreal began to come on a regular basis in the 1790s. The most famous visitors of the early years were Meriwether Lewis and William Clark, whose expedition made winter camp in 1804–05 near present-day Stanton.

In the 1820s and 1830s American traders made the upper Missouri country a hinterland to St. Louis. They brought in guns, kettles, blankets, and axes, as well as liquor and disease. The white man's goods made the Indians dependent on the traders, his liquor demoralized them, and his diseases killed them. In 1837 smallpox, carried up the Missouri by passengers aboard a steamboat of the American Fur Company, reduced the Mandan population from about 1,800 to 125 in a few months. Indian hostility grew when steamboat traffic increased after the discovery of gold in Montana in 1862 and when the United States Army built forts along the rivers. In 1876 Col. George A. Custer and the 7th Cavalry set out from Ft. Abraham Lincoln, south of present-day Mandan, for their fateful encounter with the Sioux and Cheyenne on the Little Bighorn River.

Pioneering and statehood. The fur trade declined in the 1860s, and white settlement began in earnest in 1871, when railroads reached the Red River from St. Paul and Duluth. A flood of pioneers took up land under the Homestead Act and turned to wheat farming. During the period known as the Dakota Boom (from 1878 to 1886), the many giant farms advertised the new country, and North Dakota wheat made Minneapolis, Minnesota, the milling centre of the nation in the 1880s. The Northern Pacific and Great Northern railroads vied with one another to reach the richest grain centres. Dependence on wheat unified the farmers and strengthened the populist revolt against eastern monopolistic practices. The Dakota Territory was divided in 1889, and both North and South Dakota were admitted to the Union on November 2, 1889.

Populist strength

The modern state. Revolt against outside exploitation reached a climax soon after the period of pioneer settlement ended in 1915. Controlling the state government after the 1918 election, the Nonpartisan League enacted a socialistic program that included a state-owned bank and a state-owned flour mill and grain elevator. The league soon lost political control, but the North Dakota Farmers Union (founded in 1927) launched a strong cooperative movement to control the selling of grain and the purchase of farm supplies. Such radical farm movements made many North Dakotans oppose American intervention in both world wars, because they identified participation with war profits for Wall Street.

From 1915 on, North Dakota's history is marked by continuing adaptations to the cool, subhumid grassland environment. The most important of these have been the ever-increasing mechanization of agriculture, the enlargement of farms, the loss of rural population, and the widespread use of the automobile. After World War II came rural electrification, soil conservation, and highway construction. In the 1950s North Dakota became an oil-producing state, while the 1960s brought air bases, missile sites, and antiballistic-missile installations. National and international trends as well as internal adaptations began to influence the state as never before.

THE NATURAL AND HUMAN LANDSCAPE

Surface features. North Dakota is part of two major physiographic provinces. The eastern half belongs to the Central Lowland that stretches westward from the Appalachians, while the western half is part of the Great Plains that reach to the Rocky Mountains. The state is like three broad steps rising westward: the Red River Valley lies 800 to 1,000 feet above sea level, the Drift Prairie from 1,300 to 1,600 feet, and the Missouri Plateau from 1,800 to 2,500 feet. The highest point in the state is White Butte, at 3,506 feet (1,069 metres). The Central Lowland portion comprises the Red River Valley, a flat, glacial lake bed extending from ten to 40 miles on either side of the Red River of the North, and the Drift Prairie, a rolling country covered with glacial drift. On the west, the Missouri Escarpment separates the Drift Prairie from the Great Plains. The North Dakota portion of the Great Plains is known as the Missouri Plateau. East and north of the Missouri River, it is covered with a thick layer of glacial drift. The Altamont Moraine in this area, one of the principal flyways for migrating wildfowl, is full of potholes, lakes, and sloughs. Like the Drift Prairie, this region has a young drainage system, for rivers are few wherever the great ice sheets covered the land.

River systems

Forty-one percent of North Dakota is drained by the systems of the Red and Souris rivers, whose waters flow eventually into Hudson Bay. The Missouri Plateau and the James River system form a part of the drainage of the Missouri, which flows into the Mississippi and thence into the Gulf of Mexico. West of the Missouri River the landscape has been shaped by running water that has carried away as much as 1,000 feet of sedimentary deposits. In some places, especially along the Little Missouri River, it has carved spectacular cliffs, buttes, and valleys that form a spectacular landscape known as the North Dakota Badlands.

Climate. Its location at the centre of the North American continent gives the state a continental climate: hot summers and cold winters, warm days in summer and cool nights, low humidity and low precipitation, and much wind and sunshine. The western part of the state has lower humidity, lower precipitation, and milder winters than the eastern half. For the state as a whole, the average precipitation is about 17 inches (432 millimetres). The southwestern counties, the warmest, have an annual mean temperature of about 42° F (6° C); the

northeastern counties, the coldest, about 38° F (3° C). The growing season ranges from 134 days at Williston, in the northwest, to 104 days at Langdon, in the northeast.

Vegetation, soils, and animal life. Before settlement, 95 percent of the state was covered by grass, for low precipitation, drought, and grass fires discouraged trees. Long-lived perennial grasses begin to grow early in the spring, produce seed quickly, and go into a dormant state in drought. They protect the soil from erosion and provide food for grazing animals. The heavy grass cover of the Red River Valley and the Drift Prairie formed black soils, while the lighter grass cover of the Missouri Plateau formed lighter, thinner, dark-brown soils. The North Dakota grassland was a natural habitat for great herds of buffalo and antelope. Belts of timber and brush along the rivers provided homes for white-tailed deer, elk, and bear. Small buffalo herds today are protected in parks.

Human habitation. The regions are reflected to some degree in the character of the people. The inhabitants of the Missouri Plateau tend to be more informal and Western in their manners and dress, whereas those of the Red River Valley tend, perhaps, to be more reserved and Eastern. The Drift Prairie is a transition zone in this respect, as it is in climate and in plant and animal life.

The typical landscape of North Dakota

North Dakota is a land of large farms and ranches: a vast, open country with few fences. There is an awesome beauty in the great fields and pastures, the big sky, the endless view of flat or rolling prairie with the black earth of the plowed land, the green blanket of a new crop, or the yellow cover of ripened grain. The clean, dry air and the bright sun give a wholesome look to the land, but the large holdings, averaging more than 1,000 acres in 1970, make the countryside seem lonely and almost uninhabited. Outside of municipalities in 1970 there were only about three persons per square mile. Some 45,000 farms, more than one-half the number existing in 1933, have been absorbed since that year into neighbouring holdings.

With the loss of farm population many small towns have disappeared also, while in others businesses and houses stand empty. The larger cities and towns provide a sharp contrast, with their new stores, public buildings, and housing developments and their air of vigour and prosperity. The sparsity of population affects not only the state's economy but also the character of the people, who tend to be friendly, spontaneously helpful, and straightforward. Distances create isolation, but the electronic media now keep North Dakotans well informed about happenings elsewhere.

THE PEOPLE OF NORTH DAKOTA

Ethnic groups. When white traders reached what was to become North Dakota, several Indian peoples lived in the region: Mandans, Hidatsas, and Arikaras along the Missouri River, Chippewas and Crees in the northeast, Assiniboins in the north, Yanktonai and Wahpeton Dakotas in the southeast, and Teton Dakotas and Crows in the west. The fur trade brought Frenchmen, Scots, Englishmen, Canadians, and Americans, and, by 1800 the Métis, of mixed white and Indian ancestry were an established element in the population.

The earliest white settlers included many Norwegians, Canadians, and Germans whose people had migrated earlier to Russia. By 1890 the foreign-born constituted about 43 percent of the population, a higher percentage than in any other state; and by the census of 1920, when settlement had been completed, only 32 percent of the white population was of native-born American parentage.

Some Indians in North Dakota, like Indians in other states, form a submerged group with more than their share of poverty, ill health, and alcoholism. Others, however, have succeeded as farmers, ranchers, and professional men or in politics or sports.

Demography. After 1930, the population declined in every decade except the 1950s. The 1970 population included nearly 16,000 Indians and 2,500 blacks.

Population losses

North Dakota's birthrate is close to the national average, but for many years North Dakota had one of the lowest death rates in the nation. It would have experienced a rapid growth in population but for the heavy emigration that since 1930 has been greater than the excess of births over deaths. By 1960 nearly half of the persons born in North Dakota and still living resided in other states. Only Wyoming and Arkansas had lost a larger percentage of their natives. The lack of a diversified economy and, hence, of economic opportunities continues to account for most of this loss.

Within the state a steady migration from the farms and villages to the towns and cities has continued. The percentage of the population living in places of 2,500 or more, the U.S. minimum standard of the Bureau of the Census for "urban population," increased from about 17 in 1930 to about 45 in 1970. During the 1960s some counties in the central and western parts of the state lost one-fourth or one-fifth of their population. The four most populous counties, however, contained nearly 40 percent of the state's people. The density of the population declines to the west, as does the rainfall. Thus, in 1970 the eastern tier of counties in the Red River Valley, which contains the two largest cities, Fargo and Grand Forks, had a density of 23.9 persons per square mile; the central tier, 8.9 persons; and the western tier, 4 persons.

Religious affiliations. North Dakotans are a church-going people. Probably less than 10 percent of the population of an age for confirmation are not confirmed church members. About one-half of them are Lutherans, about one-third are Roman Catholics, and most of the rest are Methodists, Presbyterians, and members of the United Church of Christ.

THE STATE'S ECONOMY

North Dakota's cool, subhumid climate and its location far from the nation's markets have shaped its economy. Among the west north central group of states to which it belongs, North Dakota has the lowest farm income, the smallest cities, the lowest rainfall and temperature, the shortest growing season, and the least manufacturing.

Major crops

Agriculture. The state produces beef cattle, is second in the nation in production of wheat, rye, and oats, and is first in barley and flaxseed. It also sends dairy products, sugar beets, and potatoes to outside markets, from which it buys its automobiles and trucks, its farm machinery and equipment, its automotive fuels, its lumber and building materials, and its clothing and television sets and other consumer goods. In 1970 manufacturing accounted for only about 10 percent of the state's income, and its lignite, the largest supply of solid fuel in the United States, plays a minor role in its economic life. Wheat is the most important source of farm income, but the nation's annual per capita consumption of wheat flour dropped significantly during the 1960s.

Although agricultural production largely pays for the things the state buys in outside markets, it employs only about a fifth of the labour force—about 45,000 in 1970 compared to about 160,000 in nonfarm employment. Statistics on personal income show the relative unprofitability of agriculture, with governmental disbursements far exceeding farm income. In 1971 North Dakota ranked 43rd among the states in per capita income. Farming's economic disadvantages contrast sharply with its rapid increase in efficiency since World War II, with increased mechanization and the decline in number but increase in size of farms.

Other sources of income. The discovery of oil at Tioga in 1951 made North Dakota by 1970 the 14th largest producer of crude petroleum in the nation. The production of electrical power increased over 750 percent from 1950 to 1968. In the 1950s and 1960s the economy was stimulated by substantial investment of government funds in Garrison Dam, in highway construction, in rural electrification, in air bases, and in missile installations. Federal expenditures are of continuing importance in the state, which ranked sixth among all states in government payments to farmers in 1968, although the state was only 26th in cash receipts from marketings of crops and livestock.

Transportation. Intrastate and interstate traffic moves primarily over east–west and southeast–northeast routes and secondarily over north–south routes. It flows to and

from the principal trading centres within the state, the nearest metropolis—Minneapolis–St. Paul, in Minnesota—and the Pacific Northwest, with Fargo the chief centre for intrastate traffic. North Dakota's transportation network includes more than 5,000 miles of rail lines, nearly 9,000 miles of hard-surfaced and 57,000 miles of gravelled highways, and three airlines providing scheduled service to seven cities. In 1969 North Dakotans owned one motor vehicle for each 1.5 persons—only Wyoming and Nevada had a higher proportion nationally—and in 1971 the state ranked sixth in the amount spent on highways for each $1,000 of personal income.

ADMINISTRATION AND SOCIAL CONDITIONS

Structure of government. To many observers, North Dakota suffers from too much government. The state government has the usual structure: a governor who is elected for a four-year term, nine elected heads of executive departments, a bicameral legislature of 49 senators and 98 representatives, and several levels of state courts. In addition, it has nearly 150 departments, boards, and agencies, two state-owned industries, and 18 public institutions. In 1967 North Dakota ranked first among the states in number of elected officials per 10,000 population and second in local governments per 100,000 population. By 1971 almost 44,000 persons, including teachers, were employed by state and local government. In 1971 North Dakota ranked 26th in property taxes per capita, 31st in all taxes per capita, but only 43rd in personal income per capita. In the 1960s many of the state's leaders believed that the constitution of 1889, a document of detailed legislation rather than an effective framework for government, had become outmoded. In 1970 the voters authorized a constitutional convention; the new and simpler document was defeated at the polls in 1972.

Extensiveness of governmental machinery

Politics. Voting trends usually favour Republican candidates, but many voters are independent. In the 11 presidential elections from 1928 through 1968, the Republican candidate carried the state eight times, and in the 1960s the voters regularly elected a Republican majority to the legislature and a Democratic governor.

Education. In the late 1960s, 58 percent of North Dakota's high school graduates went on to college, and 17 percent to vocational school. Between 1959 and 1969 the number of school districts was reduced by nearly three-fourths, the number of one-room rural schools by nine-tenths, and the number of high schools by more than 100. The consolidation has been effective in providing better schools, but many observers feel even further reduction is necessary. Many high schools remained too small to provide adequate programs, since enrollments ranged from as few as 18 to almost 3,000.

Higher education has expanded spectacularly. Enrollments more than doubled during the 1960s in the ten state institutions. More than half were in the two universities, the University of North Dakota, in Grand Forks (founded in 1883), and North Dakota State University, in Fargo (1890), which offer a full range of undergraduate and graduate work.

Health and welfare. North Dakotans receive excellent medical care despite the small population scattered over a large area. Few people live more than a two-hour drive from one of the centres. Although some towns of less than 1,000 population have a doctor, medical practice is concentrated in the four larger cities—Fargo, Bismarck, Grand Forks, and Minot—often in group practice in well-equipped clinics. The state has 60 general hospitals, a rehabilitation centre, five regional mental-health centres, and a state hospital for the mentally ill. The state health department and smaller health districts provide public health services. Colleges of medicine and nursing at the University of North Dakota train health personnel.

Economic assistance and social services are provided by the state public-welfare board, county welfare boards, and private welfare agencies, especially denominational groups. The state welfare board gives aid to aged, blind, and disabled persons and to dependent children; it also provides eight regional social-service centres. County welfare boards administer general assistance and medical aid for the aged. Welfare assistance for able-bodied men is a minor expenditure. More than two-thirds of the funds paid to or for welfare recipients are from federal sources, while most of the remainder is from the state. Most of the money for dependent children, the second most expensive program, goes to families broken by divorce, separation, or desertion, although in the state only 36 children out of every 1,000 received such aid, compared with 75 nationally.

CULTURAL LIFE AND INSTITUTIONS

The arts. The traditional North Dakota spirit of self-reliance and voluntary cooperation, except where internal or external pressures have dictated institutionalization, is reflected in the cultural life of the state. Without a large metropolitan centre, North Dakota's cities and towns with universities or colleges provide the main cultural leadership. The three symphony orchestras have headquarters in Fargo, Minot, and Grand Forks, though they make appearances throughout the state. The North Dakota Ballet is located in Grand Forks, where in 1971 the University of North Dakota established the state's first College of Fine Arts. Most of the community art associations, public concert associations, and community theatre groups are also located in college or university towns. A summer School of Fine Arts is held at the unique International Peace Garden, a large and beautiful park that virtually eliminates the border between North Dakota and Manitoba near the Turtle Mountain area.

Interlocking educational and cultural activities

There is a fair amount of federal-assistance funding for arts projects in the state, but most other funds for the arts, apart from those expended by educational institutions, have had to come from public subscription. In 1971, however, the modest state appropriation made to the North Dakota Council on the Arts and Humanities, the agency through which federal funds for the arts are dispensed, was looked on as the beginning of a long-term state commitment to the arts.

Libraries. Among the weakest aspects of North Dakota's cultural life is library service. Because the larger part of the population lives in the country or in small villages, about half of the people have virtually no contact with library facilities. The libraries of the 13 towns with 5,000 or more population vary widely in their adequacy. Civic leaders have sought to meet the needs of the rural population with county and regional libraries. By 1968 there were eight county and four regional libraries, and 15 bookmobiles were operating. In 1969 the State Library Commission had nearly $1,000,000 to improve rural services, 80 percent of which was from federal funds.

Folk culture. Indigenous folk traditions continue within the state among the Sioux, or Dakota, peoples of Fort Totten and Standing Rock Indian Reservation, among the Plains Ojibwa (locally called Chippewa) people of the Turtle Mountain Reservation and area, and among the people of the Three Tribes—the Arikaras, the Hidatsas, and the Mandans—of Fort Berthold. Traditional music and dances, together with beadwork and other crafts, attract many art lovers. The strong, well-defined pottery of the Three Tribes is particularly sought after.

Indian arts and other ethnic activity

Scandinavian cultural traditions remain vigorous. Though none of the 50 Norwegian-language newspapers published, often briefly, between 1878 and 1955 survives, Norwegian language and literature are taught at the University of North Dakota and in several elementary schools. The Sons of Norway have more than 9,000 members in the state. Norwegian costumes, customs, and cookery are observed on many occasions but especially on Norwegian Independence Day, May 17. North Dakotans of Icelandic, Czech, and German ancestry also retain some ethnic customs, and in many families the ancestral languages are still spoken.

Recreations. The individualistic character of North Dakotans is reflected in their sports and pastimes. In 1970 the state issued more than 80,000 fishing and 200,000 hunting licenses, not counting those for taking fur-bearing animals. Snowmobiling was on the increase—18,000 new vehicles were registered in 1968–70—but ice skating, skiing, and ice hockey remained popular winter sports.

Communications. In 1971, 30 AM and 12 FM radio stations and 13 television channels operated within the state. Of the television studios, eight had local colour-production facilities. Cable television, some of it relaying Canadian and educational channels, was established in several communities. There are ten daily newspapers in the state and two semiweeklies, but the 90 weeklies are only a remnant of the nearly 350 that existed in 1915.

Prospects. In the future, North Dakota will probably see the continuation of the two most important trends of the recent past: the decline of the population in the countryside and small towns and the growth of larger geographical units to provide regional services. The first makes the second an essential consequence. As the population becomes sparser, it is necessary to enlarge administrative areas to incorporate enough people and enough potential income to insure the efficient provision of services.

The small population and relative lack of economic and political power sometimes leads North Dakotans to compare their state unfavourably with richer, more populous, and better-known parts of the world, and self-disparagement has almost become a native trait. If the winters are longer and the opportunities are more restricted than elsewhere, however, North Dakota by and large retains a cleaner environment and a closeness to and understanding of the land—qualities of life that in the 1970s were becoming increasingly longed for and sought after, especially in the crowded and polluted cities of America. How or in what direction North Dakota might contribute to satisfying these national needs and wants remains an unknown but intriguing question for the future.

BIBLIOGRAPHY. Two books serve as an introduction to the literature on North Dakota. The FEDERAL WRITERS' PROJECT, *North Dakota: A Guide to the Northern Prairie State*, 2nd ed. (1950), is a survey of all aspects of North Dakota life. Ten tours cover every part of the state and give short histories of every community. ELWYN B. ROBINSON, *History of North Dakota* (1966), is generally judged the definitive account; it covers all aspects of the state's history from the earliest times to about 1960, with more emphasis given to the trends of the 20th century. Both volumes have extensive annotated bibliographies and picture sections.

Three books deal with the flora and fauna of the state: JOHN E. WEAVER and F.W. ALBERTSON, *Grasslands of the Great Plains: Their Nature and Use* (1956); ORIN ALVA STEVENS, *Handbook of North Dakota Plants* (1950), a scientific work; and VERNON BAILEY, *A Biological Survey of North Dakota*, U.S. Department of Agriculture, *Biological Survey Bureau, North American Fauna 49* (1926).

Indian life in the early history of North Dakota is in ALFRED W. BOWERS, *Mandan Social and Ceremonial Organization* (1950). Much detail on the Indians and Métis at Pembina appears in the letters of missionaries found in *Documents Relating to Northwest Missions, 1815–1827*, ed. by GRACE LEE NUTE (1942). Two scholarly works throw some light on the Indians as they deal with related subjects: WILLIAM E. LASS, *A History of Steamboating on the Upper Missouri River* (1962); and JOHN E. SUNDER, *The Fur Trade on the Upper Missouri, 1840–1865* (1965). The conditions of tribes on the Missouri River in 1867–69 is detailed in the diary of an army officer: *Military Life in Dakota: The Journal of Philippe Régis de Trobriand*, trans. and ed. from the French original by LUCILLE M. KANE (1951). An outstanding volume on the Indians' Ghost Dance is ROBERT M. UTLEY, *The Last Days of the Sioux Nation* (1963).

A volume that carries the story from fur trade through settlement is VERA KELSEY, *Red River Runs North!* (1951).

The broadest picture of farming and farm life developments is STANLEY NORMAN MURRAY, *The Valley Comes of Age: A History of Agriculture in the Valley of the Red River of the North, 1812–1920* (1967). An intimate picture of the farming frontier is in AAGOT RAAEN, *Grass of the Earth: Immigrant Life in the Dakota Country* (1950). There are other books on social and cultural life. The story of the German-Russians is well told in ADOLPH SCHOCK, *In Quest of Free Land* (1964); while a vivid picture of farm life in the 1930s is found in *The Bones of Plenty: A Novel* (1962), and *Reapers of the Dust: A Prairie Chronicle* (1964), both by LOIS PHILLIPS HUDSON.

An excellent study of early 20th-century North Dakota political history is ROBERT L. MORLAN, *Political Prairie Fire: The Nonpartisan League, 1915–1922* (1955).

(E.B.R./B.O'K.)

Northern Ireland

The geographical position of Northern Ireland holds the key to much of its unique social, economic, and political development—a process beset by such deep-rooted antagonisms as to make the strife-torn nation globally significant by the early 1970s, perhaps because it seemed to mirror, in microcosm and in the Northern Hemisphere, many of the problems then afflicting evolving nation-states in other areas of the world. Geographic factors also help to account for the beauty of its landscapes, and its location is one of the keys to its culture and distinctive character of its peoples.

The country—created a self-governing state within the United Kingdom by the Government of Ireland Act of 1920—lies in the northeast of the island of Ireland, itself located on that western continental periphery often characterized as Atlantic Europe. Northern Ireland is often referred to as the province of Ulster (and its inhabitants as Ulstermen), though it includes only six of the nine counties which made up that historic Irish entity. It occupies 5,452 square miles (14,129 square kilometres) of land, about a sixth of the whole of the island, and is separated from Scotland, another constituent country of the United Kingdom, by the narrow North Channel, which is at one point only 13 miles wide. Historically, this channel has been a link rather than a barrier, and from the earliest times it has witnessed a constant coming and going of peoples. This interchange gave the northern part of the island a distinctive regional character confirmed during the Industrial Revolution when the province emerged as a linen-manufacturing region, and later strengthened when a shipbuilding and engineering industry, based on imported raw materials, made Belfast a major city. The intrusive influences that brought about this cultural and economic transformation were attenuated in the west and south, and the political border with the Republic of Ireland is more of a compromise conveniently based on long-existing county borders than a clear-cut regional boundary. The political separation of Northern Ireland has merely confirmed its economic position as a region existing within the framework of the United Kingdom as a whole, serving as an extension of the larger unit's industrial resources. Although the cultural links of some of the people of Northern Ireland with the republic of Ireland were still strong—and much in evidence politically—in the early 1970s, it seemed as though these ties were likely to suffer some erosion as state legislation follows economic trends, and progressively caters to educational and social needs along lines similar to those introduced in England, Scotland, and Wales (see also BRITAIN AND IRELAND, HISTORY OF; UNITED KINGDOM).

THE LAND

Relief and soils. Northern Ireland can be thought of topographically as a saucer centred on Lough (lake) Neagh, the upturned rim of which forms the province's highlands. Five of the six counties—Antrim, Down, Armagh, Tyrone, and Londonderry—meet at the Lough, and each has a highland core on the saucer's rim. To the north and east the Antrim Plateau tilts upward toward the coast. It reaches heights of between 1,000 and 1,500 feet, resulting in an impressive cliff coastline of basalts and chalk, broken by a series of the glaciated valleys known as glens, which face Scotland and are rather isolated from the remainder of the province. County Down's rounded landscape of drumlins—smooth mounds of clay left by the retreating ice of the final glaciation—is punctuated by Slieve Croob (1,755 feet [535 m]) and culminates in the Mournes, which rise to Slieve Donard (2,796 feet [852 m]) within two miles of the sea. This impressive landscape of granite peaks is bounded by Carlingford Lough to the south. Topography

Armagh's scenery is gentler, but the land rises to 1,893 feet in Slieve Gullion near the border. West of Lough Neagh the land rises gently to the more rounded Sperrin Mountains, shared by Londonderry and Tyrone: Sawel (2,240 feet [683 m]) is the highest of several hills over 2,000 feet. The sixth county, Fermanagh, is focussed geographically on its own lake basin, occupied by Lough

Erne, in a drumlin-strewn area ringed by hills more than 1,000 feet high.

Much of the Northern Ireland landscape is gentle and, in most low-lying areas, is covered with swarms of drumlins that have played havoc with the local drainage and are interspersed with marshy hollows. Glaciation also gave the land its main valleys, the Bann—draining Lough Neagh to the north—the Blackwater to the southwest, and the Lagan to the east. All of these valleys have been important routeways but none more important than the Lagan, penetrating from Belfast Lough to the very heart of Ulster.

Soils are varied. In spite of the fact that there is much glacially transported material covering the areas below 700 feet, the nature of the soil is predominantly influenced by the nature of the underlying parent rock. Brown earth soils, forming arable loams, are extensive, and are derived from the ancient Silurian rocks of Down and Armagh and from the basalts of Antrim. There are some ash-like podzols in the Sperrins, and the impeded drainage of much of Fermanagh gives acidic soil. Peat soils are common, particularly in the hollows lying between the drumlins, and hill peat is widespread over the province. Although it is of no great commercial value, peat traditionally has been a source of fuel for the peasant farmer and is still cut extensively.

Climate and ecology. Northern Ireland's climate is temperate and maritime: most of its weather comes from the southwest in a series of lows bringing the rain and cloud that often lend character to the landscape. As the country is near the central track of such lows, it often experiences high winds, and in north Londonderry and on the wild Antrim coast, particularly, severe westerly gales are common. Above the 800-foot level, distorted trees and windbreaks testify to the severity of the weather. Rainfall decreases from west to east, but the hills accentuate the amount to over 65 inches (1,650 millimetres) in the Sperrins and the Mournes, compared with 35 inches (900 millimetres) at Lough Neagh and east Down. A relatively dry spring gives way to a wet summer and a wetter winter. Conditions generally are very changeable, but there are no extremes of heat or cold. Normal temperatures vary between means of 41° and 50° F (5° and 10° C). These mild and humid climatic conditions have, in sum, made Northern Ireland a green country in all seasons.

Maritime influences

The general features of the plant and animal life of the province are similar to those in the northwest of Britain. The imprint of man is heavy on the landscape and is particularly evident in the absence of trees. Most of the land has been ploughed and drained and cultivated for centuries. Above the limit of cultivation, rough pastures have been grazed extensively and beyond them lies a zone of mountain vegetation. Only about 3 percent of the land is now under forest, and more than half of this has been planted by the state. Young trees in these plantations are economically unimportant, but locally they are helping to diversify the landscape.

Geographic regions. The geographic regions of Northern Ireland correspond closely to the main topographical elements, although they are also the outcome of the cultural evolution of the province. In Antrim and the Bann Vally, in north Down and in north Armagh, the influence of the Scots and English has been paramount. West of Lough Neagh and in the fastness of the Mournes and of Slieve Gullion or in the more distant Lough Erne region indigenous elements have maintained a distinctiveness that is still apparent. Such relatively isolated pockets as the Antrim glens and Kilkeel have retained a local consciousness that gives colour and interest to the human geography of the province.

Settlement patterns. The overwhelming impression of the province's landscape is of scattered and isolated farms. Occasional relics of tiny hamlets, or clachans, show that peasant crofts once were huddled together in kinship groups who worked their strips in an open-field system. Between the end of the 18th century and the middle of the 19th most of the land was enclosed and the scattered strips consolidated, partly as a policy of the landlords but finally because of the effect on the population of the potato famines of the 1840s. The end result was the orderly, small square fields of today dominating the contemporary landscape. Some landlords rearranged their tenants' land in narrow ribbons, from valley bottom to mountain pasture, giving a characteristic ladder of fields, with the farms strung along the road on the valley side. Drumlins also have had an effect on siting; houses are found away from the peaty bottom but below the windswept skyline. Most farmhouses are small and a few are still thatched. The occasional larger farm often has a Georgian house—simple and dignified and a reflection of the age of consolidation.

Small market towns rather than villages are frequent. Built by the English and Scots planters or by the landlords of the 18th century, they have a foreign touch of orderliness and urbanity. Many are grouped around a "diamond" (meeting place), or square, which is used as a marketplace. Some of them acquired a mill in the 19th century, but in few cases has this marred the essentially rural context.

Market towns

Few of the market centres have grown into substantial towns. In the western half of the province, regional services and administration have enlarged Omagh (12,000 population) and Enniskillen (6,500). Some towns have grown with the introduction of industry, notably Dungannon (7,500), specializing in fabrics, and Carrickfergus (15,000), noted for rayon. Armagh (12,000) is an ecclesiastical centre with two cathedrals, while Lisburn (29,000), Portadown (21,000), and Lurgan (24,000), all in the Lagan Valley, form an extension of the Belfast industrial complex, their size based on the textile industry. Bangor (35,000) is a resort and a residential outlier of Belfast. Londonderry (52,000) is a centre for shirtmaking and textiles. It was formerly the centre of the Lough Foyle lowlands until the hinterland that it served was split by the partition of Ireland, but it still remains the main focus of the west. The size of Belfast (359,000), at the head of Belfast Lough on the northeast coast, underlines its dominance of the region, as well as its significance as an industrial centre and major port. Its economic bases are linen and engineering and, notably, shipbuilding; textile machinery is also important, and the city is the centre of provincial government, finance, and education. Reflecting the city's 19th-century origin, most of the streets are inextricably and bleakly mixed with mills and factories, while the reclaimed land at the head of the Lough is given over entirely to industry.

THE PEOPLE

Cultural origins. The cultural differences that underlie many of Northern Ireland's contemporary social problems have a long and troubled history. The province has had lasting links with parts of western Scotland, strengthened by constant population movements. After the Tudor invasions and particularly after the forced settlements, or plantations, of the early 17th century, the English and Scots elements were further differentiated from the native Irish by virtue of their Protestant faith. Two distinct and often antagonistic elements—the indigenous Roman Catholic Irish and the intrusive Protestant English and Scots—date from that period, and have played a significant role in molding the province's development. The intrusive element dominated County Antrim and north Down, controlled the Lagan corridor toward Armagh, and also formed powerful minorities elsewhere. This situation contributed to the decline of Gaelic speaking, and it is reflected in the contemporary distribution of religions (see map.) Even a century ago Gaelic speaking was confined to the more remote mountain areas—the Sperrins, south of the Mournes, and the glens of Antrim. By the 1970s these areas had been reduced to the merest pockets where Gaelic may still occasionally be heard at a traditional entertainment. The accents given to English, however, are regionally distinctive. The northeastern dialect, dominating Antrim, Londonderry, and parts of Down, is an offshoot of central Scots dialect and reflects the latter in almost all of its features. The remainder of the province, including the Lagan Valley, has accents

Dialects

derived from England, more particularly from Cheshire and southern Lancashire, and the "west country" counties of Gloucester, Somerset, and Devon. The towns show more of a mixture and an overlay of standard English.

Distribution of Roman Catholics in Northern Ireland.

Some 35 percent of the population is Roman Catholic, 30 percent is Presbyterian, and some 25 percent Episcopalian (Church of Ireland), Methodists and members of other sects making up the remainder. The distribution of Roman Catholics and Protestants is, however, very uneven. In country districts, the latter are in a majority in Antrim, north Down, and the Lagan Valley, and in north Armagh. Elsewhere, they are in a minority, though fairly highly localized. Most towns have a Protestant majority: this is the case in Belfast, where Roman Catholics comprise only 30 percent of the population. Towns remote from Belfast—Newry and Londonderry—have higher percentages. In the towns there is a high degree of segregation of the sects and mixing is minimal. Industrial west Belfast is split into two sectors along two axial roads. The Falls Road is as exclusively Catholic as the Shankill Road is Protestant. In many streets adjoining the boundary line segregation is 100 percent. East Belfast has an exclusively Catholic core. In the middle class suburbs segregation is less apparent. Segregation increases as socioeconomic status decreases. Civil disturbances are almost confined to where segregation is highest. The proportions of the sects are changing slightly because of a differential in birthrate. In Belfast, for example, fertility ratios in Catholic districts are very much higher than in Protestant areas. The situation in the province is reflected in the relative decline of Protestants in the last half century, though more or less stable in absolute numbers, and the relative and absolute increase of Catholics. In 1926 the relative percentages were as follows: Roman Catholics 33.5, Presbyterians 31.3, Church of Ireland 27.0. In 1961 the corresponding figures were 34.9, 29.0, and 24.2. But this is unlikely to change the total predominance of Protestants in the province for some time to come.

Demographic trends. The population of Northern Ireland was 1,525,000 in 1971, having grown steadily over the previous half century: it has been estimated that it will probably increase to about 2,000,000 by the year 2000. In 1971, 30.1 percent were under 15 years of age, 54.5 percent were between 15 and 60, and 15.4 percent were 60 and over. The birthrate, which has been falling very slightly since 1965, was 21.1 per 1,000 in 1970, higher than in the remainder of the British Isles. The deathrate in 1970 was 10.9 per 1,000, lower than in the remainder of the British Isles. Marriage rates (8.1 per 1,000) were a little lower than in England, Wales, and Scotland, but fertility rates were considerably higher. In spite of a rate of natural increase running about 17,000 per annum between 1966 and 1970, there was a net outward migration of 6,901 per annum during the 1960s, a little less than the average in the previous decade. Inside the province, some movement from the countryside to the towns is having an effect on the distribution of population. The two eastern counties, Antrim and Down, contain the bulk of the population, with their residents totalling over 650,000 (excluding Belfast), as compared to 490,000 in the west (excluding Londonderry). These two counties are also gaining considerably in population, whereas the others are making only very modest gains, and Fermanagh actually lost population between 1961 and 1966. The countryside also is losing people to urban areas. The apparent slight increase in the rural population is due mainly to urban overspill, and the gross number in the countryside is still less than it was in 1901. On the other hand, some towns have grown considerably. Although there has been comparatively little change in some of the smaller and medium-sized towns—Newry and Newtownards, for example, have lost marginally—most of the industrial centres grew rapidly between 1961 and 1971: Lisburn from 17,700 to 28,904; Portadown from 18,609 to 20,577; Lurgan from 17,872 to 23,853; and Larne from 16,350 to 18,219. Belfast's apparent decline since 1951—from 443,671 to 358,991—reflects its growth outside the borough boundary, much of which is now accommodated in the contiguous new town of Newtownabbey (57,846). Increasing urbanization is most marked in the metropolitan zone existing around Belfast, for more than 500,000 people are now concentrated in the Lagan Valley. The key to this trend lies in the area's industrialization, and it is likely to continue to attract population. The projected city of Craigavon, when complete (it is expected to house 100,000 people by 1981), will further emphasize the importance of this corridor, of which the concentration of people dominates the economic, social, and political life of the whole province. Outside it, the second largest town, the county borough of Londonderry (52,000), is only a weak countermagnet, and there is little to balance the overwhelming dominance of Belfast. The future development of Coleraine (14,851) as a university town may well stimulate a greater range of urban services and a diversification of activities that will give a fillip to the western part of the province.

Urbanization

THE ECONOMY

Economically, Northern Ireland is an integral part of the United Kingdom. Its trade is dominated by imports from the United Kingdom (73 percent of the total) and exports to the United Kingdom (86 percent), though some of the latter are re-exported directly from England, or are the bases of manufactured goods that are subsequently exported. Trade with the Republic of Ireland accounts for about 8 percent of the total, and there are lesser links with parts of the Commonwealth and of Europe.

Compared with its southern neighbour, Northern Ireland is an industrialized country; and the mass of its exports is made up of manufactured goods. This situation, however, has placed a heavy reliance on the import of raw materials, and the Belfast region might well be thought of as an extension of the industrial regions of northwest England and the Clyde. Its own mineral resources are extremely meagre: the amount of coal mined locally is quite negligible, although local chalk, clays, and gravels are used to produce lime, bricks, and cement. Northern Ireland's power resources, too, depend on imported coal and oil.

Agriculture. Northern Ireland does possess considerable agricultural resources, which it is developing as a

major part of its economy. Fundamentally, it is a country of peasant farmers; the great majority of farms are still under 30 acres, and only about 5 percent exceed 100 acres. By the early 1970s, however, this situation was changing. The number of farms—approaching 37,000—was decreasing by some 3 percent per annum, with a substantial increase in average size, which now exceeds 50 acres. Consolidation has meant a better livelihood for fewer farmers, whose total numbers dropped from over 100,000 to under 83,000 over the 1960s. Even so, only half the farms were economically viable, and many were operated on a part-time basis. Almost all farms now have electricity, and there has been a great increase in the numbers of tractors and combine harvesters in use. The frequent rain, the high humidity, and the prospect of wet harvests discourage arable farming, but local conditions produce good grass and rich pasture: nearly all grassland is ploughed and there is very little "rough grazing." Mixed farming predominates, with a seven-year rotation —four years of grass or hay, two of oats, with a root crop, such as potatoes, intervening. About 85 percent of farming income comes from livestock and their produce. By the 1970s, there were well over 1,000,000 cattle, nearly 1,000,000 sheep, 1,000,000 pigs, and 12,000,000 poultry in the country. The production of grass seed and seed potatoes for export is also important. To the south of Lough Neagh lies a rich orchard country, and apples and market gardening are constant features of the landscape of north Armagh.

Consolidation of farms

Table 1: Value of Imports and Exports, 1970
(£000,000)

	imports	percent	exports	percent
Live animals and food	159	19	146	20
Beverages and tobacco	57	7	1	0
Crude materials	28	3	41	6
Mineral fuels	48	6	2	0
Animal and vegetable oils	1	0	1	0
Chemicals	42	5	4	1
Manufactured goods	217	26	194	26
Machinery and transport equipment	222	27	132	18
Miscellaneous manufactured goods	55	7	224	30
Total	829	100	745	101

Industry. Although farming dominates the landscape of the country, it provides a livelihood for only one in ten of Northern Ireland's inhabitants, and the real wealth lies in manufacturing. More than 58,000 people work in the engineering, shipbuilding, and vehicle trades, and nearly 50,000 in the textile industry. Locally grown flax and a plentiful supply of soft water originally stimulated the growth of the textile industry, and spinning still flourishes in many inland towns. The other industrial group is more a product of foresight, as raw materials, capital, and even skills were drawn into Belfast to make it possible.

Flax is now imported, and linen is still synonymous with Northern Ireland, though cotton and artificial fibres, as well as woollen fabrics and carpets, are now important. The clothing industry employs a further 28,000 people. Although all these activities are dominated by Belfast, Londonderry is well known for shirt and pajama manufacturing. Belfast's shipyards are world-famed, and some of the skills developed there have been diverted into the aircraft industry. Engineering grew largely to meet the demands of the textile and shipbuilding industries. Belfast also has chemical plants, rope-making factories, and a large percentage of the food-processing work of the province; tobacco also plays a leading role.

The industrial structure of Northern Ireland is a vulnerable one, and unemployment in the province is considerably higher than in the remainder of the United Kingdom, fluctuating between 6 and 9 percent of the labour force, with a much higher rate among men than women. For over two decades the government has actively encouraged the growth of manufacturing and such economic diversification as broadening the base of the textile industry by introducing man-made fibres. A series of govern-

Table 2: Distribution of Manpower, March 1971
(000)

	number	percent
Agriculture, forestry and fishing	54	9.7
Mining and quarrying	3	0.5
Manufacturing industries	185	33.1
Construction	54	9.7
Gas, electricity and water	8	1.4
Transport and communications	24	4.3
Distributive trades	64	11.5
Professional services	76	13.6
Financial and miscellaneous services	52	9.3
Public administration	39	7.0
Total	559	100.1*

*Figures do not add to one hundred because of rounding.

ment acts has tempted new enterprises into the country by providing factory space and making grants toward capital expenditure. In 1964 an Economic Council, not unlike the regional economic planning boards of the United Kingdom, was set up, and a second development program was initiated for 1970–75. In relation to the remainder of the United Kingdom, Northern Ireland is peripheral and partially isolated. This adds to the difficulty of maintaining local growth. The organization of labour is strongly oriented to Britain: local trade unions are affiliated to the Trades Union Congress, but, in fact, a vast majority of the 228,000 members belong to English- or Scottish-based unions. About 15,000 workers are members of unions controlled from the Irish Republic.

Development programs

Northern Ireland is unified fiscally with the United Kingdom. Its public revenue comes from two sources. Death duties, stamp duties, motor vehicle taxes, and betting duties are collected locally, but customs and excise duties, income tax, and surtax are collected by the United Kingdom with a Joint Exchequer Board deciding the province's share, as well as its contribution to certain United Kingdom services, such as the armed forces, and its share toward paying off the national debt.

TRANSPORTATION

One of the more noteworthy features of the countryside of Northern Ireland is a close network of well-maintained roads, totalling in all some 14,200 miles. The majority are very minor, with over 8,000 "unclassified," but, even so, all parts of the province are accessible. Over 1,000 miles of the roads are first class and 344 miles are trunk roads. There are some 40 miles of motorway linking Belfast with northern Armagh. Public-road transport outside the Belfast municipal service has been nationalized since 1935, and since 1948 the Ulster Transport Authority (since 1967 the Northern Ireland Transport Holding Company) has also controlled the railways. The latter have diminished rapidly—from 824 miles to 200 miles—in the economic reorganization following railway nationalization; a link with Londonderry through County Antrim, and the line to Dublin, via Newry, were all that remained by the early 1970s. Inland waterways have almost disappeared, although a little commercial traffic still uses the lower Bann navigation to Coleraine, and there is some recreational sailing.

The province's links with Britain by sea and air are of paramount importance. Belfast is one of the major ports of the British Isles, handling between 6,000,000 and 7,000,000 tons of shipping and freight annually by the 1970s. Its eight miles of quays now contain modern container-handling facilities. Londonderry and Larne, the only other ports of significance, both handle over 1,000,000 tons of freight annually. Newry and Coleraine handle cargo, but Warrenpoint is likely to have improved facilities, which will replace those of Newry. Larne and Belfast handle passenger transport. The former has services to Stranraer, Ardrossan, and Preston, the latter, to Liverpool, Ardrossan, Preston, Heysham, and the Isle of Man. All services handle car ferries, an important factor in the province's increasing tourist trade; and the Stranraer and Heysham routes also carry mail. The main

civil airport for Northern Ireland is at Aldergrove, which has regular daily air service to major cities in Britain, and, since 1968, scheduled flights to North America. Passengers total over 1,100,000 a year.

ADMINISTRATION AND SOCIAL CONDITIONS

Constitutional framework. Under the Government of Ireland Act (1920) separate parliaments were planned for both northern and southern Ireland, but the plan became effective only for the north, where a Parliament was formally opened in 1921. When the Irish Free State emerged in 1922, Northern Ireland opted to continue its form of government, which was marked by strong ties with the central government of the United Kingdom at Westminster; a commission eventually confirmed a boundary with the south that would contain the six counties. When the Irish Free State seceded from the Commonwealth in 1949, the union of Northern Ireland with the United Kingdom was further confirmed. Paradoxically, its measure of self-government is the result of its determination to preserve union with Britain, and its links with London are safeguarded by the election of 12 members to Westminster.

Political links with the United Kingdom

Until 1972, the crown exercised all legislative power in Northern Ireland through a governor who held office for six years. The Parliament consisted of a Senate and a House of Commons. The former had 24 members elected by the Commons for eight years, and the lord mayor of Belfast was an ex officio member. The House of Commons had 52 members and a maximum life of five years, and in all ways was closely modelled on Westminster, which was responsible for matters relating to the crown, war and peace, the armed forces, and foreign powers, as well as trade, navigation, coinage, and many others. Northern Ireland could make no laws interfering with religious equality. In almost all internal matters it followed Westminster policy. There is universal franchise above the age of 18.

Local government is also derived from the pattern in England and Wales. There are six county councils besides Belfast County Borough Council—all responsible, among other things, for education, health, and welfare. The nine borough and 24 urban councils are responsible for sanitation, housing, planning and development, and care of towns in general, while the 26 rural councils are mainly sanitary and housing authorities. Craigavon has an independent Development Commission, as has Londonderry, but local councils are elected. From 1971, voter eligibility in local elections was to be determined in the same way as for parliamentary elections.

Political developments. Politics since the country's inception has been dominated by the issue of union or separation, and this split has followed religious lines. Protestants have been Unionists, and have always had a majority in Parliament. Roman Catholics have supported the Nationalist Party, advocating reunion with the Republic of Ireland, and have always been in opposition. The Labour Party has made some impact in representing labour, irrespective of religious affiliation. Northern Ireland entered the 1970s with 36 Unionist members of Parliament, six Nationalists, three Independent and unofficial Unionists, three Independents, two Northern Ireland Labour, and two Republican Labour. Electoral geography follows almost exactly a map of the distribution of religious majorities (see map), the west being a stronghold of Nationalism, the east of Unionism: the concentration of people in and around Belfast has assured an overall Unionist majority. Periodically, political and religious antagonisms have flared into civil disorder, most severely in 1921–22, but, since 1968, they have flared again. The main issues were discrimination in housing and employment, and the fact that franchise for local government, based on property ownership, favoured Protestants. Differences rose in the Unionist Party on the speed with which reforms should be pressed, giving rise to protests against the alleged dilatoriness of social reform against a background of economic insecurity. The early stages of recent civil disobedience partly ignored party lines, but continuing violence tended to restore the established religious split, which, accompanied by increased terror, led the British Parliament to suspend Northern Ireland's Parliament and government (March 1972), and place the country under direct rule from London. In March 1973 the British government outlined future proposals to be implemented in a new constitution. They included: the pledge of continued union with the United Kingdom, as long as this was the majority wish; a Northern Ireland Assembly of 80 members to be elected by proportional representation; departmental committees reflecting the religious balance of a province; and a guarantee of no religious discrimination. Earlier, 57.4% of the electorate had voted for a retention of the border and 0.63% against: 41.4% had not registered a vote.

Religious antagonism

The judiciary and civil order. The Supreme Court is under the Parliament of the United Kingdom. Under the chairmanship of the lord chief justice a High Court of Judiciary and a Court of Appeal hold assize courts in the county towns and also sit as a Court of Criminal Appeal. There are also locally administered county and magistrate courts.

Civil order, increasingly controversial in the 1970s, is maintained by the Royal Ulster Constabulary (RUC), which replaced the Royal Irish Constabulary in 1922. In 1971 the RUC consisted of approximately 4,000 officers. Because of the sporadic outbursts of unrest since Northern Ireland came into being, the RUC has been assisted by a volunteer part-time Ulster Special Constabulary. Permanent special constables known as category A disappeared early, and the reserves (category C) were no more than a list: but the "B Specials" number about 8,000. A 1970 act authorized the reorganization of the entire police force under a police authority and the formation of a Royal Ulster Constabulary Reserve to replace the B Specials, whose use had engendered considerable controversy. By 1972, regular British Army troops had entered the country to quell the civil strife.

Services. The social services in Northern Ireland are very closely patterned on those of the United Kingdom, although there is sometimes delay in implementing policy, partly because of historic inertia, and partly because of the dual nature of the society. Nowhere is this more apparent than in education. The 1947 Education Act parallels the 1944 act of England and Wales, but its implementation is hampered by the traditional tenacious denominational control of education and the general reluctance to abandon it. By 1970, a little fewer than half the 1,240 or so primary schools were still under voluntary management, a situation that reflected the religious division. Only 21 of the 81 grammar schools were under the control of county authority. The 80 voluntary county secondary schools are rapidly becoming maintained schools; that is, obtaining most of their financial support from a local authority, together with a measure of public representation on their managing bodies. Northern Ireland has three teachers' training colleges—one of which is nondenominational—and two universities. The Queen's University, established in 1845 as one of three in Ireland, has had a charter since 1908, and in the early 1970s had over 6,500 students; the New University of Ulster at Coleraine (1968) anticipated a student enrollment exceeding 2,000 by 1973.

Denominational education

In health services, as in education, Northern Ireland follows the United Kingdom. The Queen's University has a large medical faculty that supports the health service's 93 hospitals and 18,000 beds. Ulster is also known for its export of doctors and nurses.

As in other regions of Britain, one of the major social problems in Northern Ireland is housing. There is a legacy of much poor rural housing in the countryside, exacerbated by the isolation of many of the small dispersed farms, while in Belfast the province suffers from a grim legacy of 19th-century industrial housing. In the quarter century following World War II, some 40 percent of the houses of Northern Ireland were replaced, rehousing about 600,000 people. A third were privately built, and the remainder were constructed by local authorities and the Northern Ireland Housing Trust. This replacement and growth has not been subject to planning legislation in

the same way as in England and Wales, although some controls have recently been introduced. In 1963 a regional plan proposed a "stop line" for Belfast's growth, and in 1965 a New Town Act was passed. This resulted in the establishment of the town of Craigavon, linking and incorporating Lurgan and Portadown in a linear design. In 1966 Antrim was designated a new town, with its population scheduled to increase from 8,000 to 30,000, and in 1967 Balleymena was similarly designated for a growth from 15,000 to 50,000. In 1967, too, the Londonderry Development Commission took over the future of that city. The central coordination of all planning, though proposed in 1964, had not materialized by the early 1970s.

Living standards. Social conditions and standards of living in the province have not yet entirely overcome a rural tradition of subsistence farming and low cost of living. Nor has this been offset by conditions among industrial workers, whose employment, as a result of the province's peripheral status in the United Kingdom economy, has felt every depression early and has always been slow to recover. High unemployment and more marked differentials between skilled and unskilled workers' wages have aggravated social problems. Generally speaking, although the standard of incomes and living in Northern Ireland and in England and Wales is roughly comparable, there are still many in Northern Ireland who do not achieve those standards.

Unemployment

CULTURAL LIFE

In the arts and in cultural life generally it is difficult—as is the case with most aspects of life in Northern Ireland—to distinguish between native and imported. Few traces remain of any culture predating the Protestant Ascendancy: the occasional *caley*, or traditional Irish group entertainment, is only the faintest echo of a past that faded with the language. Gone, too, are the strolling players and rhyming weavers. Folk participation and recreation are periodically focussed on religious ceremonies and processions—colourful, noisy, and (in recent years) tragically violent demonstrations of sectarian feelings. This fundamental division has also given rise to the nearest approach to folk art, the painting of William of Orange crossing the Boyne, a deeply symbolic representation, adorning the gable end of many a Protestant terrace. In the name of religion, paint, bunting, flags, and arches briefly splash colour over drab grey streets.

In other respects, the cultural milieu of the province is one shared with the remainder of the British Isles and has few distinctive regional characteristics. With the exception of a few country mansions—a legacy of 18th-century English architects—the buildings of Northern Ireland are undistinguished, though this same influence has lent a dignity of proportion and urbanity to some larger farmhouses and to the occasional terrace of houses in the towns. For the most part, however, the industrial growth of the 19th century has completely overshadowed the previous planned phase of growth without contributing any buildings of distinction. It remains to be seen how far larger projects—university buildings and hospitals, for example—have improved the situation, but on the domestic level the Northern Ireland Housing Trust, responsible for rehousing outside urban limits, has made a considerable contribution to both the design and the grouping of houses.

Arts and music

Northern Ireland has an Arts Council, which successfully encourages all aspects of the arts. Its activities tend to be concentrated in Belfast because this city alone can provide for ballet and opera. Belfast also has two theatres; and drama, much of it local, plays a large part in the recreational life of the city. Music is mainly imported, but the city has a symphony orchestra and a youth orchestra, and in the last decade it has fostered one of the largest festivals in Britain. The council also sponsors art exhibitions. Belfast has a permanent gallery, and so does Londonderry, though some exhibitions tour the entire province. The 19th century saw little development in the visual arts, but a new interest in landscape emerged briefly at the beginning of the 20th. A provincial school of painting may now be emerging for the first time and some recognize a school of poetry. But the province is only slowly shaking off the utilitarian, no-nonsense approach to life that underlay its Victorian growth.

The British Broadcasting Corporation has a monopoly in sound broadcasting, but Ulster Television, Ltd., representing the Independent Television Authority, vies with the British Broadcasting Corporation. Northern Ireland also shares the British press: all the national papers are distributed in the province, which also has two morning, one evening, and one Sunday paper. More local news is handled by more than 40 weekly papers.

The Ulster Museum provides an interesting link with peasant origins in Northern Ireland, for one of its most promising aspects is an open-air folk museum, first opened in 1964. This is another reflection of an appreciation of the past that is also shown by the state's care of ancient monuments, both historic and prehistoric.

Of other cultural institutions perhaps the most notable is the observatory at Armagh. Founded by Archbishop Robinson in 1790, it has remained an independently governed institution, though now considerably state-aided. It has links with observatories in South Africa and also has one of the few astronomy libraries in the British Isles.

PROSPECTS

One of the attractions of Northern Ireland is that it is rich in tradition: its tragedy is that it cannot forget its history. Its social problems are an inheritance from its past, aggravated for many by uncertainty of employment and by insecurity. Its stability depends as much on economic prosperity as on religious equality, both necessary conditions for the continuing development of those qualities that have given the province a distinctive personality and of which the Ulsterman is justly proud.

BIBLIOGRAPHY. E. JONES (ed.), *Belfast in Its Regional Setting* (1952), is a general historical, geographical, and social introduction to the province. L. SYMONS (ed.), *Land Use in Northern Ireland* (1964), and R. COMMON (ed.), *Northern Ireland from the Air* (1964), are more strictly geographical. A brief but masterly summary is E. ESTYN EVANS, "The Personality of Ulster," *Trans. Inst. Br. Geogr.*, 51:1–20 (1970), and the character of one part of the province is admirably portrayed in the same author's *Mourne Country* (1951). The best background to the cultural and political history of Ulster is M.W. HESLINGA, *The Irish Border as a Cultural Divide* (1962). D.P. BARRITT and C.F. CARTER, *The Northern Ireland Problem* (1962), deals with the social and political prelude to the current situation; and T. WILSON, *Economic Development in Northern Ireland* (1965), provides the economic background. R. MATTHEWS, *Belfast Regional Survey and Plan, 1962* (1964), looks at possible future developments. The details of the capital city are dealt with in E. JONES, *A Social Geography of Belfast* (1960), which discusses the historical and environmental background to the social situation. Useful factual material is found in the *Ulster Year Book* (annual), as well as in the Census reports for 1961 and 1966.

(E.J.)

Northern Territory

The central section of northern Australia, the Northern Territory is bounded on the north by the Timor and Arafura seas, and by Western Australia to the west, Queensland to the east, and South Australia to the south. It is approximately 1,000 miles (about 1,600 kilometres) from north to south and 600 miles (about 1,000 kilometres) from east to west. Its area is 520,280 square miles (1,347,519 square kilometres)—17½ percent of the Australian Commonwealth. It is largely tropical. The population in 1971 was estimated at 85,500—only 0.67 percent of the Australian population.

Constitutionally, the territory is inferior in status to the states, and it has very limited legislative powers. Its development since 1911, when it was transferred from South Australia, has been a major item of expenditure in terms of works, services, and inducements to producers to accept the risks of an uncertain physical and economic environment. The nature of the climate, the poor soils, distance from assured markets, and problems of recruiting labour have been considerable handicaps.

The landscape. The unspectacular coastline is flat with low headlands and is mostly fringed with mangrove swamps. There are many offshore islands, of which Melville and Bathurst islands and Groote Eylandt are the largest. Inland from the coastal belt there is a gradual rise to the town of Tennant Creek (1,229 feet, or 375 metres) on the vast Precambrian plateau (1,000–2,000 feet) that extends south and west into the neighbouring states. Farther south, Alice Springs (1,790 feet, or 545 metres) is situated on an alluvial plain in the Macdonnell Ranges, of which Mt. Zeil (4,955 feet, or 1,510 metres) is the highest point in the territory. There are some remarkable tors 200 miles southwest of Alice Springs: Mt. Olga (3,507 feet, or 1,069 metres) with 30 separate domes, and Ayers Rock (2,845 feet, or 867 metres), a red, ovoid monolith rising 1,100 feet.

A number of rivers, of which the Finke and the Todd are the largest, flow from the central ranges after rainstorms. Areas north of the plateau are drained by some substantial rivers: the Victoria, 350 miles long, and the Daly, 225 miles long, flow to the Timor Sea; the Katherine flows southwest from Arnhem Land to join the Daly; the Adelaide, Mary, and South and East Alligator rivers enter Van Diemen Gulf; and the Roper and the McArthur flow east and northeast into the Gulf of Carpentaria.

Vegetation and animal life. Some forests with Indo-Malaysian vegetation elements exist in Arnhem Land, but otherwise the northern vegetation is open woodland with low eucalypts and tall grasses of low nutritive value. In the main cattle areas of the Victoria River Downs and the Barkly Tableland, an open-tussock grassland on heavy, gray-brown cracking soils is dominated by Mitchell grass with subdominant Flinders grass and herbs. Between the Barkly Tableland and the ranges a wide belt of pervious sand supports only spinifex. In the Alice Springs district the rocky hills carry spinifex, but on the light-textured soils of the lower slopes and valleys an association of mulga (an acacia tree) with short grasses is a valuable fodder resource. Mixed mulga-spinifex scrub with seasonal herbs and ephemerals occupies the nearby red plains of desert loam, and farther to the west is a desert of hummock grassland composed of widely spaced clumps of spinifex and *Plechtrachne*.

Higher plants exceeding 5,000 species are represented. Species endemic to the Macdonnell Ranges include a cycad, *Macrozamia macdonnellii*; a palm, *Livistona mariae* —a surviving representative of the vegetation that once covered central Australia; and several acacias and eremophilas. A broad-leaved species of mulga, the witchety bush, harbours a grub much sought by Aborigines as food. The baobab tree, with a girth of 30 feet, occurs near the Victoria River.

Exotic animal life

Principal birds are the princess parrot of the central, rocky spinifex country, the flock pigeon of the grasslands of the Barkly Tableland, and lorikeets, parrots, and rock pigeons in rugged areas of the north. The black-banded pigeon is known only in the western escarpment of Arnhem Land. Mammals include one of the egg-laying species—the echidna—and a variety of marsupials. The kangaroo is widely distributed, but some species have restricted habitats: the rat kangaroo is adapted to the arid regions; the rock wallaby and antilopine wallaroo inhabit the rocky ridges of the northwest; and the black wallaroo is restricted to the granitic ranges of Arnhem Land. The unique rock-haunting ring-tailed opossum is an interesting evolutionary development. Formerly domestic animals now existing as large wild populations include camels, buffalos, cattle, pigs, goats, horses, and donkeys. With the exception of the buffalos, which are hunted for hides and meat, these are serious pests. The cattle tick also causes great economic problems.

Climate. There are two seasons: a wet summer, from November to April, and a dry winter. Rainfall is extremely variable and of marked summer incidence. The change from the north, where the annual rainfall is 60 inches, to the southeast, where it is only 5 inches, reflects the diminishing influence of the Australian-Asian monsoon and an increasing dependence on tropical thunderstorms. Only 30 percent of the territory has an annual rainfall of between 15 and 40 inches—the effective limits of agricultural development. The climate is hot and, in the north, uncomfortably humid for eight months of the year. Mean temperatures at Darwin are 84° F (29° C) in January and 76° F (24° C) in July, and at Alice Springs 81° F (27° C) in January and 53° F (12° C) in July. The north is free from frosts.

Patterns of settlement. The pastoral potential of the territory was first recognized during construction of the Overland Telegraph Line, and on its completion, in 1872, areas near Alice Springs and Darwin were stocked with cattle. The establishment of cattle stations on the Barkly Tableland during the 1880s completed the foundations of the pastoral industry. The land is rented from the government, and in 1969–70 there were 322 such holdings. Tenure of pastoral leases is normally 50 years, and the covenants stipulate conditions of maintenance and improvement. These have been minimal, and inadequate safeguards against ill-advised practices of absentee leaseholders have caused great damage to vegetation and extensive soil erosion. Annual rentals are low; on a typical property it was 30 Australian cents per square mile in 1960; this was reappraised at 76 cents in 1970.

The two traditional regions

The territory is traditionally viewed as two regions, roughly delineated by latitude 20° S and commonly referred to locally as the North and the Centre. The two areas differ in topography and climate and represent the spheres of interest of Darwin and Alice Springs. As natural outlets for cattle and mineral concentrates from their respective regions, these centres have developed rapidly, and in June 1970 accommodated 60 percent of the total population. Little urban growth has occurred elsewhere: Tennant Creek and Katherine have urban populations, but there are only four other groupings of more than 20 dwellings.

The people. *Demography.* During the gold boom at Pine Creek in 1872, labour problems led to the engagement of Chinese from Singapore, and by 1881 the population comprised 670 Europeans and 2,781 Chinese. By 1888, when immigration restrictions were imposed, Chinese numbered about 6,000. Subsequent immigration was mostly European, and at the 1966 census the population of 56,504 was 60 percent European, 37 percent Aboriginal, with 3 percent of some 20 other groups. (The racial statistics, based on self-assessment, are only of approximate accuracy.) The increase of 29,000 to June 1971 was predominantly European. Nearly 80 percent of the whites are of Australian birth. Age structure and other population characteristics are typical of Australia. Of the 80 percent claiming to be Christian, one-third is of Catholic and two-thirds of Protestant faith. In 1966, 39 percent of the population was under 21 and 77 percent under 40 years of age. There were 122 males to every 100 females. Whereas most of the white population is concentrated in the towns, about 86 percent of the Aborigines live in rural areas or in the 15 Aboriginal reserves, which total 94,196 square miles.

In 1970 the birth rate was 36.5 and the death rate 8.5 per 1,000 persons. Infant mortality was 48.4 per 1,000 live births, with greatest incidence among Aborigines. The rate of growth of population was 4.9 percent in 1970; most of this increase was at Darwin. Internal movements are small but might be expected to increase as new centres develop. The new town of Nhulumby, for example, was planned to accommodate 5,000 persons associated with a bauxite project upon its completion by 1974.

The Aborigines. Few Aboriginal tribes are wholly nomad. Many have retained their tribal structure and their customs and religious rituals that govern their social relationships. Their religious and magical rites draw on a rich repertoire of songs and dances in which sacred objects and personal adornment play an important part. These are accompanied by a variety of musical instruments. Tribes and dialects are very numerous. Languages are of the agglutinating type; *i.e.*, they combine into single words two or more elements of distinct and separate meaning. Though many of the languages have characteristics in common, the prefixing languages of the north are basically different in structure and vo-

Aboriginal languages

cabulary from the Aranda languages of the south. Many tribes have permitted in the main their languages and religious rites to be placed on permanent record. Government policy on Aboriginal welfare is one of assimilation, promoting their participation in the privileges and responsibilities of citizenship so that they may become an integral part of the Australian way of life.

The economy. *Agriculture and mining.* The traditional dependence of the economy on cattle raising has been substantially changed since 1966 by developments in mining and other primary industries. Agricultural production was small and related to local needs, until, in 1969, 3,000 tons of grain sorghum was exported to Japan. Seven companies commenced prawn fishing and processing in 1969, and in the first year of operations 6,000 tons valued at A$4,900,000 was exported. Pastoral production is confined to beef cattle, grazed largely under open range conditions. In 1969–70 cattle numbered 1,185,000—6 percent of the Australian total. Of the 247,000 cattle marketed, 177,000 were distributed to three states for fattening; the rest were processed by meatworks and slaughterhouses at Katherine and Darwin for export to the United States for use as hamburger and sausage meat.

Income from mining was A$5,500,000 in 1964 and close to A$39,000,000 in 1970. Mining of iron ore at Mt. Bundey and Frances Creek (near Darwin), and of manganese ores at Groote Eylandt (an island off the east coast), commenced in 1965–66. During 1969–70 more than 1,000,000 tons of iron ore was exported to Japan, and most of the manganese ore, valued at $10,800,000, was exported to Europe, Japan, and the United States. Copper and gold ores are mined chiefly at Tennant Creek. A plant with an annual mining capacity of 400,000 tons commenced production from a newly discovered copper–gold ore body in 1971. On the Gove Peninsula (northeast) a large alumina–bauxite project with a planned output of 1,000,000 tons annually is well advanced and should be operational by 1974. Extensive lead–zinc deposits near the McArthur River and at Woodcutters, near Darwin, are being evaluated by several companies.

Manufacturing and services. Secondary industries are small service industries meeting local needs; in 1967–68 production from 188 factories was valued at A$9,000,000. Tourism is increasing, and many visitors are attracted by the climate of Alice Springs and by the stark, colourful scenery of the Macdonnell Ranges and the Mt. Olga-Ayers Rock National Park.

Transport. The Northern Territory has two single-track railways: from Alice Springs to Port Augusta in South Australia, and from Darwin to Birdum. The Stuart Highway (954 miles, or 1,538 kilometres), from Alice Springs via Tennant Creek and Katherine to Darwin, and the Barkly Highway (403 miles, or 649 kilometres), connecting Tennant Creek with Mount Isa, Queensland, are the main road links. Until 1962 these were the only surfaced roads, but in the following decade more than 2,000 miles of all-weather roads were constructed. Air services are well developed: many overseas airlines operate through the international airport at Darwin, and there are services between state capitals and Darwin and intermediate towns, a substantial internal network, and charter services at Darwin and Alice Springs. There are 124 small airports, and most homesteads have airstrips. Shipping of exports from Darwin increased forty-fold between 1965–66 and 1968–69 and necessitated a large extension of port facilities.

Development of air services

Administration and social conditions. *Governmental structure.* The Northern Territory (Administration) Act of 1910–1969 provides for an administrator, appointed by the governor general of the Commonwealth of Australia, to administer the territory on behalf of the commonwealth. There is also provision for a Legislative Council consisting of 11 elected members—of whom one is elected president—and six official members. The council makes ordinances for the peace, order, and good government of the territory but may not propose money votes except upon recommendation by message from the administrator. The Department of the Interior in Canberra is responsible for general administration. Other departments in Canberra administer defense, civil aviation, public health, justice, education (excepting Aborigines), and the provision of basic services. Municipal services are health, justice, education (excepting Aborigines), and the provision of basic services. Municipal services are provided by the administration, except at Darwin, where they are the responsibility of the Corporation of the City of Darwin. The main political groups are the Labor, Liberal, and Country parties. The territory elects one member by adult franchise (including Aborigines) to the 125-member House of Representatives of the Commonwealth Parliament. The legal system and the jurisdiction of the courts are similar to those of the states, but there is no resident judge of the Supreme Court. Law and order are maintained by 235 police, with 30 trackers and other staff. Darwin, as a defensive port, has extensive army, air force, and naval installations.

Health. The Department of Health provides medical and dental services and maintains hospitals at Darwin, Alice Springs, Katherine, and Tennant Creek, and a leprosarium near Darwin. Its aerial services to remote areas are based at Darwin; it also provides the medical personnel for the Royal Flying Doctor Service of Australia, which operates within 500 miles of Alice Springs. Two-way radio consultation usually suffices for minor ailments. Health inspectors visit all settlements, and medical and dental officers make periodical examinations of all schoolchildren.

The Flying Doctor Service

Education. Education is the responsibility of the Department of Education and Science, but there is an agreement whereby curricula and teachers are provided by the South Australian Department of Education. Of the five secondary schools, those at Alice Springs and Darwin provide tuition to leaving and matriculation levels, respectively. Higher education is not provided. Children in remote areas are served by a correspondence school supplemented by School of the Air broadcasts in which pupils may participate by two-way radio. Aborigines attend special schools, a responsibility of the Welfare Branch of the administration, that take account of their cultural background to ensure a smooth transition to their ultimate education in community schools.

Cultural life and institutions. Cultural institutions are naturally limited in a region of small population of very low density. Because of its isolation—Darwin, for example, is nearer to Hong Kong than to Sydney—the territory is largely deprived of opportunities afforded by visiting artists. Partial destruction of Darwin during World War II led to replanning on modern lines, but even though the decade following its proclamation as a city in 1959 saw the population increase from 13,000 to 32,000, Darwin lacked museums and art galleries and a seat of higher learning. There are, however, active art, musical, and dramatic societies, and the performing and fine arts are occasionally brought to the larger towns through the sponsorship of the Arts Council of Australia.

Aboriginal arts

Indigenous arts are those of the Aborigines, which symbolize the ritual of their religion. Throughout the territory there are examples of sculpture, rock carvings, bark and rock paintings, and "X-ray" drawings (which show not only the outlines but the bone structure and internal organs of creatures of esoteric significance). The best of the decorative arts are produced in Arnhem Land and range from gravestones and ceremonial posts to personal ornaments in attractive, polychromatic designs. Aboriginal art has received wide recognition by artists and designers, and Aboriginal motifs figure prominently in the decor of many public buildings in the state capitals.

Regional national broadcasting stations are situated at Darwin, Alice Springs, Tennant Creek, and Katherine, and there is one commercial station at Darwin. The first television station came into operation at Darwin in August 1971. Library services are provided from the five principal centres. Newspapers are published weekly at Alice Springs and Tennant Creek, and the *Northern Territory News* is published daily at Darwin.

Prospects. Since 1946, the Commonwealth Scientific and Industrial Research Organization (CSIRO) has con-

ducted research at Katherine on matters of basic importance to the future of rural industries. A major problem of the cattle industry is the loss of production during winter caused by deterioration of native grasses, but introduction by CSIRO of a legume—Townsville stylo (*Stylosanthes humilis*)—which is either oversown on native pasture or grown as a pure stand to provide high-protein standing hay, has greatly increased potential production. And major changes in the pattern of development of the territory that have occurred since the middle 1960s, chiefly in the mining industry, suggest that its economic prospects are assured and that it will make a growing contribution to the national economy.

BIBLIOGRAPHY. Accounts of investigations of dry-land and irrigation farming, establishment of pastures, and husbandry are published in the COMMONWEALTH SCIENTIFIC AND INDUSTRIAL RESEARCH ORGANIZATION (CSIRO), *Division of Land Research and Regional Survey* (annual). Other references are CSIRO, *Katherine Research Station Report 1946–56* (1959); the DEPARTMENT OF TERRITORIES, *Prospects of Agriculture in the Northern Territory* (1960); B.R. DAVIDSON, *The Northern Myth: A Study of the Physical and Economic Limits to Agricultural and Pastoral Development in Tropical Australia* (1965); J.H. KELLY, *Struggle for the North* (1966); G.W. LEEPER (ed.), *The Australian Environment*, 4th ed. (1970); and the annual *Official Year Book of the Commonwealth of Australia and Northern Territory Statistical Summary*.

(A.E.Sc.)

North Mexican Indian Cultures

The generally accepted ethnographic definition of northern Mexico includes that portion of the country roughly north of a convex line extending from the Río Grande de Santiago on the Pacific coast to the Río Soto la Marina on the Gulf of Mexico. This southern boundary coincides in a general way with the northern margins of pre-Columbian Meso-America. Northern Mexico is more arid and less favourable for human habitation than central Mexico, and its native Indian peoples have always been fewer in numbers and far simpler in culture than those of Meso-America. Today, the native peoples are extinct over all of northeastern Mexico; the only Indians present in that area are a group of Kickapoo who immigrated to Coahuila from the United States in the 19th century. In the west the Sierra Madre Occidental, a region of high plateaus that break off toward the Pacific into a series of rugged *barrancas,* or gorges, has served as a refuge area for the Indian groups of the northwest, as have the deserts of Sonora. At present only the northwestern states of Baja California, Sonora, Sinaloa, Nayarit, Jalisco, Chihuahua, and Durango have Indian populations.

Although accurate population data is lacking in parts of this region, estimates place the total population that is still Indian in language and culture at approximately 130,000, a tiny minority among the several million non-Indians of northwest Mexico.

MAJOR INDIAN GROUPS

Surviving Indian peoples of northern Mexico today fall easily into two divisions. By far the greater number are members of the first type, the ten groups which speak some language of the Uto-Aztecan linguistic stock and are traditionally agriculturists. The second type is now reduced to four groups and less than a thousand individuals—the descendants of nomadic bands who resided in Baja California and coastal Sonora and lived by hunting and gathering wild foods. The second type spoke various languages not related to Uto-Aztecan.

Uto-Aztecan peoples

Uto-Aztecan peoples of northern Mexico have been divided into three branches—Taracahitian, Piman, and Aztecoidan. The Taracahitian branch consists of the Tarahumara of the southwestern Chihuahua numbering 50,000; the Varohío, a small people of 1,500 who border the Tarahumara on the northwest and are closely related to them; the Yaqui, with 18,000 in the Río Yaqui valley of Sonora and in scattered colonies in towns of that state and in Arizona; and finally the Mayo of southern Sonora and northern Sinaloa, who still number about 30,000. Another Taracahitian group, the once prominent Opata, have lost their own language and no longer maintain a separate identity. The Piman branch consists of three groups. One of these, the Pima Bajo of the Sierra Madre border of Sonora–Chihuahua, has a population of some 1,500; a few score other Pima are scattered over Sonora. The second group, the 300 Papago of northwest Sonora, are identical with a much larger portion of the same tribe in Arizona. The third Piman tribe is the Tepehuán, one enclave of which is located in southern Chihuahua and another in the sierras of southern Durango and Nayarit; the total Tepehuán population is estimated at about 8,000. The third branch of Uto-Aztecan is the Aztecoidan family, including the Cora located on the plateau and gorges of the Sierra Madre of Nayarit and the Huichol in similar country of northern Jalisco and Nayarit. Cora and Huichol number about 10,000 each. A final member of this branch, locally called the Mexicanero, includes speakers of Nahuatl remnants of central Mexican Indians introduced into the area by the Spaniards. The Mexicanero number only a few hundred and live in the mountains of Nayarit and southern Durango.

Baja Californian and coastal Sonoran peoples

The remnants of the Baja California Indians—the Tipai (Diegueño), Akwa'ala (Paipai), and Kiliwa—live in ranch clusters and other tiny settlements in the mountains near the American border and number fewer than 500 in all. Speaking Yuman languages of Hokan stock, they are little different today from their relatives in American California. Some 300 Cocopa in the Colorado Delta in like manner represent a southward extension of Colorado River Yumans from the American Southwest. Three hundred Seri are found along the desert coast of north central Sonora. This famous group also speaks a language of Hokan origin and is probably related to now extinct peoples who lived across the gulf in Baja California two hundred years ago.

Missions and isolation helped to preserve the several surviving Indian groups of northwest Mexico through the colonial period (1530–1810), but all underwent considerable alteration under the influence of European patterns. Nearly all the agricultural tribes adopted some form of Roman Catholicism and much Spanish material culture. It was at this time that the traditional cultures of northern Mexico were formed, the basic patterns continuing until the present. Many groups faded away—gradually losing their languages and identities and becoming a component of the emerging mestizo, or mixed-blooded European and Indian population, the predominant people of present-day Mexico. Only the Huichol, Seri, and Tarahumara remained primarily aboriginal cultures, but even these groups adopted many items and ideas from the Spanish invaders.

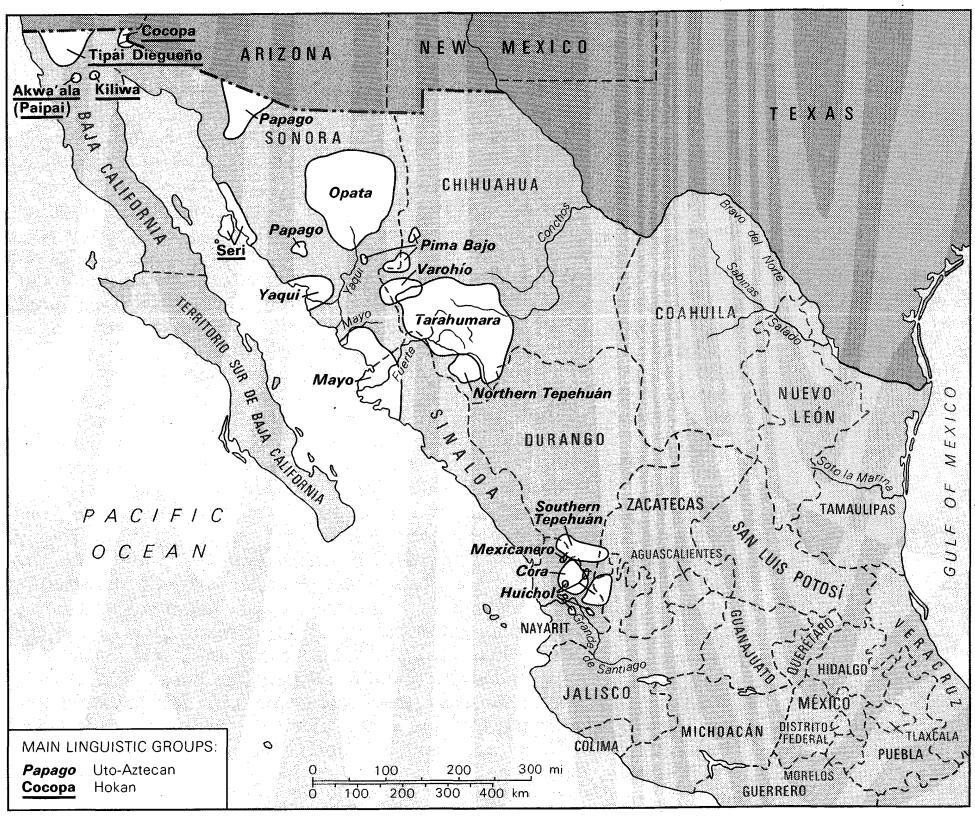
Today, all these peoples exist as ethnic enclaves surrounded by, and in most cases sharing their lands with, non-Indians and manifesting some of the characteristics of ethnic minorities everywhere. There is competition for lands with mestizo ranchers and, in most groups, a conscious desire for survival as distinct cultural entities.

CULTURE PATTERNS

Basic orientation in life

Although in some aspects of culture the Indian groups share much with other rural Mexicans, psychologically they tend to be very distinct. The basic orientation of life of the traditional Indian cultures is primarily religious in that they look to the supernatural to solve their problems and tend to see their own lives as requiring continuous service to the deities. This is in strong contrast to the practical and materialistic orientation of other northern Mexicans. The average Uto-Aztecan Indian in this area is reserved toward outsiders, especially non-Indians, and prefers not to be too much in evidence. The ideal person is industrious, carries out his religious obligations, and does not seek change in his traditional lifeways. In contrast, the Seri and Yuman groups, being of another tradition, are less inclined to secrecy and more aggressive toward outsiders.

Social structures. The social structure of the Uto-Aztecan peoples of northwest Mexico are variations of one basic type, while those of the Seri and Baja California remnants follow other forms.



Distribution of north Mexican Indians in 1970.

The Uto-Aztecan agriculturalists of modern times possess only two real social units, the family and the Indian village community. No real tribal organizations exist, and only among the Cora and Yaqui are there strong feelings of tribal solidarity. For all these peoples the community is the society, and there is little movement in or out—or even interaction with other communities of the same "tribe."

The typical Indian community

The traditional Indian community is largely the result of colonial missionary efforts to concentrate the scattered *rancherías,* or hamlets, of aboriginal times into Spanish-type villages in which the natives could be more easily administrated. Usually furnished with land grants, these communities survived the close of the mission period and remained viable social units. The modern Indian community is built around a political and religious structure having its origin in the village organization set up by early missionaries to carry out church fiestas. It consists of a series of *cargos,* or civil and religious offices, in which most males of the village participate, the higher offices being achieved with age and experience. Some version of this structure or copy of it is or was present among all these groups. Today, among most groups it is still the backbone of the community, serving to produce the village fiestas as well as settle disputes among the people. While serious crimes or major issues are now handled by the Mexican government, in most situations the Indian *gobernador* (governor) and elders are consulted as well.

Kinship patterns. The family is of one type for all the agricultural peoples. Kinship is based largely on descent from both parents, though with some orientation toward the authority of the father. The judgment of the elder males is highly respected, and in most groups the advice and permission of family elders is sought on all important occasions—to produce a ceremony, sell a cow, contract a marriage. Everywhere, however, the role of women is more nearly equal to that of men than is the case among the heavily male-dominated rural mestizo families. Needless to say, there is little of the famous Mexican *machismo* (cult of the male) among the aboriginal peoples of northern Mexico. The normal system of Spanish Christian and family names is present in all these groups.

Marriage customs

Marriage is primarily monogamous, though some informal polygyny occurs. Polygyny is most common among the Huichol (constituting about 5 percent of the marriages), which suggests that it was more common in northern Mexico in aboriginal times. Marriage tends to be brittle, and many unions are not permanent. Parents commonly arrange a match, though the custom is not uniform over the area. Among the Cora and Huichol, the boy's parents must make a series of ritual requests before the match is accepted. In most cases there is little or no ceremony; a few seek church marriages and have wedding fiestas. Among the Cora, the young couple are given formal advice by family elders in the manner of ancient Mexico.

There is a strong tendency for marriages to take place within the community. In most areas intermarriage with the mestizo population is not approved but occurs to a small degree. Among the Mayo and Tepehuán, however, there has been so much outside marriage that the ethnic status of many individuals is uncertain. In most of northern Mexico the offspring of mixed marriages pass easily into mestizo society and tend to ignore their Indian origin.

The Seri have an elaborate system of gift exchange for marriage in which the groom is expected to furnish the family of the bride with certain valuable gifts, such as a rifle, fishing canoe, or, more recently, a pickup truck. His services to the bride's family may continue for several years.

The institution of *compadrazgo,* or coparenthood—that complex of ritual relationships between parents, godparents, and child set up at the baptism of a child—has been deeply integrated into the traditional cultures. Indeed, a copy of it occurs among such non-Christians as the pagan Tarahumara and Huichol. Many minor variations of *compadrazgo* occur among the Indians. It serves everywhere to extend the kinship structure and provide "spiritual" kin on which one can call in times of need. Some Indians may choose mestizos to serve at the baptism of their child; others, such as the Cora, do not favour this.

Egalitarianism

None of the Indian peoples of northern Mexico show even minimal evidence of social stratification. Egalitarian societies exist everywhere, and historical accounts suggest

that this has always been the case, Only among certain extinct coastal peoples of southern Sinaloa and Nayarit is there found some mention of stratification on the Meso-American pattern. Prestige accrues to wealth and ceremonial knowledge, but this seldom passes to the children.

Socialization and education. Socialization and native education follow traditional patterns. A child is given little formal training, but is often admonished by his elders as to proper behaviour. Training in household tasks for the girls and men's work for the boys is through observation and gradually increasing participation until the techniques of maintaining oneself and family are mastered. Both boys and girls are expected to contribute their work as soon as possible. At an early age, children haul water, gather wood, or herd sheep. Small Seri boys spear crabs with miniature turtle harpoons as training for their adult pursuit of the sea turtle. Among the farming peoples all ages take part in the work during critical periods, such as periods of weeding and harvesting.

Rites of passage survive for children among several of the least Hispanicized groups. Cora, Huichol, and Tepehuán have rites in which a newborn is introduced to the gods. The Cora have a ceremony in which children are symbolically introduced to the use of alcoholic drinks.

Today, all Mexican Indian groups have access to schools, usually federal rural schools, and most attend to some degree, though the remote areas in which most Indians live and their lack of interest in education designed for the mestizo world does not make for the most effective program. In some areas native Indians have been trained as teachers with somewhat better results. The Kickapoo have steadfastly refused schools, seeing in them a strong threat to their cherished way of life.

Economic systems. *Settlement patterns and housing.* The aboriginal settlement pattern of the agricultural Indians centred on the *ranchería*, which consisted of a number of household units clustered in spots convenient to cultivation sites or water. When colonial missionaries incorporated these into larger villages, the lowland peoples, such as the Yaqui, Mayo, and Opata, accepted this arrangement. They continue to live in concentrated settlements today. In the Sierra Madre, subsistence patterns did not lend themselves to such towns, and the Indians returned to their *rancherías* when the missionaries withdrew.

Villages and *rancherías*

As a result, all contemporary Indians of the mountain regions live in scattered ranch clusters using the village itself as a religious and political centre. Most Indian villages have a church (often the original mission structure), a school, government buildings used for courts or meetinghouses, and small mestizo-operated stores as well as a few houses owned by Indians who spend most of the year at their ranches. In many areas the once exclusively Indian village has acquired a non-Indian population that resides there permanently. Communities of the Tarahumara, Pima, Tepehuán, Cora, and Huichol follow this pattern. Yaqui and Mayo settlements are more like towns. Many Mayo settlements are now on the outskirts of Mexican towns.

The originally nomadic Hokan peoples now tend to live in small permanent settlements. The Seri have coalesced into two fairly permanent settlements, though they still move about to fish, hunt sea turtles, or sell their crafts.

Housing utilizes the available local material, be it stone, adobe, wattle and mud, planks, bamboo, or even caves, which are occupied seasonally by some of the Tarahumara and Pima. Dwellings primarily consist of one room with a dirt floor and no chimney. The kitchen tends to be an auxiliary structure as does the corn crib. Many Indians sleep on the floor on mats and blankets, but crude bamboo or rawhide beds are also used. Furniture is seldom more than a stool or two, although a few may possess tables. In most areas Indian housing is more primitive than that of other poor rural people. Water, lights, and sanitary facilities are nonexistent.

Subsistence agriculture

Patterns of production. All the Uto-Aztecan peoples of northern Mexico are subsistence agriculturalists raising maize, beans, squash, a few other plants, and some livestock. Maize remains the basis of life and everywhere is a sacred substance, considered, for instance, as a deity by the Cora and Huichol. Cultivation methods range from the primitive digging stick used in slash-and-burn plots on hillsides and ox-plow agriculture in level fields, to some mechanized agriculture among the Yaqui. Characteristically, farm technology is primitive and of low yield. Few of the Indians have any considerable amount of good productive land, and there is competition with mestizos for even the poor mountainside plots of the Sierra Madre Occidental. The corn supply seldom lasts out the year, and in many areas it is significantly supplemented by gathering wild plants, cactus fruits, wild greens, maguey, and the usually abundant seedpods of mesquite, guamuchil, and other treelike legumes. The cash needed for outside items comes from the occasional sale of a young bull or from sporadic wage work in which many Indians engage. Deer and other game once abundant both in the sierra and in the desert are now rare. All the Uto-Aztecan tribes still hunt, but scarcity of game makes hunting relatively unimportant. The rivers yield a few fish and crayfish, which are much esteemed. Among the Seri, hunting of deer and sea turtles as well as fishing is still common but, even so, is of decreasing importance.

Wage labour

For centuries individuals from all these groups have worked for wages first in Spanish and then in Mexican mines and fields. The great silver mines at Hildago de Parral, Chihuahua, in colonial times made use of Tarahumara, Pima, Opata, and Yaqui labour. Yaqui and Mayo have long laboured on ranches and railroads and in mines away from their own country, even in the United States. Today, wage work is a factor in every Indian community in northern Mexico with some members of every group working for wages in the new agricultural areas of the Mexican west coast. The typical pattern is for young men or whole families to go to the coast for a few weeks to pick cotton or harvest corn or tobacco, returning to their own homesteads to plant and care for their maize. Some Tarahumara and Tepehuán work in lumber camps. No matter how important wages are as supplementary income, these peoples all prefer their own agriculture in their own communities, and few are drawn away permanently. It is rare indeed for sierra Indians to live and work in cities. This is not as true for the Yaqui, who have permanent settlements in several large Sonora towns.

About 1930 the Seri adopted commercial fishing and developed a mixed economy that included traditional hunting and gathering. Since 1965 a new industry has grown up in which the Indians have learned to carve animal figurines from the dense wood of the palo fierro, a desert tree. In the 1970s nearly all Seri families were producing figurines for sale to tourists and appeared to be abandoning gathering and fishing as primary means of livelihood.

Property and personal customs. The Indian market system of central Mexico does not exist in northern Mexico. All Indian areas are served by small rural stores almost entirely owned by non-Indians. Here the few but important necessities such as cloth, metal tools, soap, salt, tin cups, and matches are purchased. Money, in use everywhere, is completely a part of modern Indian culture.

Clothing and grooming

Clothing combines the older styles of rural Mexico with modern lower class dress. Only the Tarahumara, some communities of whom still wear a type of loincloth, and the Huichol, with a colourful embroidered costume, have retained forms that stand out as distinct. Some, like the Cora and Tepehuán men, favour the pajama-like muslin garments of two generations ago and today consider them Indian dress. All others wear modern clothing with few reminders of earlier attire. Huaraches (sandals) are generally worn, as are homemade or commercial hats, usually made of palm; people near the United States border, however, prefer modern shoes and cowboy hats. Women's dress throughout tends toward a skirt and blouse with a rebozo, or head scarf.

Long hair is worn by males in some Tarahumara and Huichol communities and by many adult Seri. Elsewhere, short hair is the custom. Women wear the hair loose or, among the more Mexicanized, in braids.

Food is largely vegetable and consists of local varieties of the rural Mexican staples—tortillas, tamales, beans, and cheese. Indians, however, make much use of *atole* (corn mush) and pinole (ground parched corn) both of which were aboriginal favourites and are not as popular with the mestizos.

Crafts, nearly everywhere disappearing, are very largely limited to household necessities and are seldom made primarily for sale. Wool blankets are produced by the Mayo, Tarahumara, and Cora and woven shoulder bags by the Cora and Huichol. Utilitarian pottery and twilled containers and mats are still made in most groups, using native Indian techniques. The only objects produced for the tourist market are copies and elaborations of colourful ceremonial material by some Huichol and figurines and shell beads by the Seri.

Other technology differs but little from that of other rural Mexicans of the northwest. Ranch tools and the paraphernalia used in handling livestock are identical. None of the sierra Indians possess automobiles. Occasional individuals in the other groups, including the Kickapoo, have acquired motor vehicles.

Religion. The northern Mexican tribes, like all Mexican Indians, have had contact with Christian missionaries for centuries, and all the agricultural Indians of northern Mexico are nominal Roman Catholics except for a few communities of pagan Tarahumaras, called "gentiles," and the majority of the Huichol. Even pagan groups, however, have incorporated Christian ideas and ritual practices. It can be generally stated that all the Uto-Aztecan speakers of northern Mexico today practice some form of Roman Catholicism blended with native religion. The extent of aboriginal retentions varies from group to group, with the Huichol approximating pre-Columbian patterns to the greatest degree, keeping their gods, pilgrimages to sacred places, and native religious concepts almost intact. Others like the Cora have fused these, retaining native gods but equating them with Christian personages. The Yaqui and Mayo, both with a strong religious orientation to their culture, have an even more homogenous mixture. It is doubtful, though, whether any of the groups have absorbed Christian philosophy or belief systems to any great degree. Religion retains its aboriginal functions of protecting one's health, bringing the rains, and insuring the abundance of agriculture, rather than as a means to a glorious afterlife.

Blend of Roman Catholicism and native religion

The relationship of many of these peoples to the modern Roman Catholic Church is tenuous. Modern priests tend to discourage folk observances such as the fiestas; the Indians, by and large, produce them without help of a priest. In the southern Sierra Madre, among the Cora, Huichol, and Tepehuán, there are modern Franciscan missionaries, while in the Tarahumara area there are Jesuits. Among all the Uto-Aztecans, religion remains central to the traditional culture, and it is an area in which there is great resistance to outside pressure for change.

Protestants have been active among the Seri and Yumans, achieving their greatest success with the Seri, who in the 1950s were all converted by an evangelical sect and largely abandoned their non-Christian practices. The Kickapoo, though also contacted by Protestants, remain, for the most part, followers of their aboriginal religion.

Shamans, or *curanderos*

The shaman, or medicine man, still exists in most of these groups. Called a *curandero* in Spanish, he uses supernatural means to cure illnesses, to insure the success of crops, or to assist in other situations requiring divine aid. He is very distinct from the mestizo *curandero,* who utilizes European folk medicine. Huichol medicine men, or *marakame,* are especially famous among the tribes of the Sierra Madre for their knowledge and power. Witchcraft still exists among all these peoples. A shaman himself may be accused of being a witch.

EVOLUTION OF THE CULTURES TODAY

The traditional Indian cultures of northern Mexico that emerged from the Spanish colonial world remained remarkably stable throughout the 19th century. Combinations of isolation, poverty, and conservatism resulted in what were essentially static societies. There were sporadically some uprisings, but it was not until the Mexican Revolution of 1911 and after that most Indian cultures were significantly affected by changes taking place elsewhere in Mexico. All these peoples took part in the revolutionary conflicts, and some disorganization took place everywhere. Thousands of individual Indians followed the armies, many never to return. The aftermath of the revolution marked the beginning of governmental concern with the Indians. The Instituto Nacional Indigenista was established and took upon itself the task of raising living standards and gradually integrating the Indians into the national life. Some groups benefitted significantly by the revolution and its results. Many Yaqui and Mayo escaped from hacienda peonage, and the Yaqui had a large portion of their ancestral lands restored to them.

Developments since World War II

It has been only since World War II that real changes have been taking place in the cultures themselves. There are multiple reasons for increased pressures from the outside world, but the primary cause is simply the development of modern transportation and a corresponding loss of centuries-long isolation. A major factor has been the construction of dams in the rivers flowing out of the Sierra Madre and the development of major irrigation projects and large modern cities on the coastal plain. There is now greater interest in the resources of the mountain hinterlands on the part of non-Indians, and there are opportunities for the mountain and desert peoples to engage in migrant farm labour.

Opening of the Mexican west coast highway has brought major changes in the whole area. The desert lands of the Papago, the Sonoran seacoast of the Seri, and the thorn forests of the Yaqui and Mayo are now penetrated by paved roads. A new railroad across the Sierra Madre from the city of Chihuahua to Los Mochis on the Pacific bisects the Tarahumara country, and roads have made numerous villages of this tribe accessible to truck traffic. Truck roads are approaching even the gorges of the Cora and the Huichol. Many parts of the Sierra Madre previously accessible only by animals or by foot are now served by local airlines that fly small planes hauling freight and passengers to most of the mountain communities. It is a common sight at the airport in Tepic, Nayarit, to see rural mestizos, Cora, and Huichol Indians waiting to board a plane for airstrips near their remote homes. Although there are still vast areas in these mountain regions not reached by modern transportation, the outside world has become more accessible, and all of these peoples have been affected to some extent.

Development of modern transportation has greatly increased the possibility that many Indian communities in time will cease to exist. There is growing competition for lands in all areas as non-Indian cattlemen, lumbermen, and farmers exploit these regions more intensely. Almost everywhere Indians find themselves being pushed or crowded out of ancestral lands by more sophisticated forces who have use for their resources. The Instituto Nacional Indigenista has come to the aid of some, such as the Tarahumara and Huichol and given them legal assistance in securing land titles. In most areas, though, Indians having no legal knowledge and, lacking sophisticated leadership, are severely handicapped in dealing with these problems. Without a land base it is doubtful that communally oriented societies such as these can long survive.

The basic strength of the Indian groups today is found in their village organization and associated ceremonial structures that effectively preserve ethnic boundaries. Those retaining the strongest organized communities—the Tarahumara, Yaqui, Cora, and Huichol—while threatened, appear to be in no immediate danger of cultural disintegration. Others, few in number (such as the Pima) or lacking well preserved independent village structures (such as the Mayo), appear less likely to survive the impact of modernization.

Fear of alienation of lands and ultimate loss of ethnic identity has led to avoidance of innovations in many cases. Most still continue to resist social and religious change and avoid too close contact with non-Indians. There is, it is true, less reluctance to adopt material ob-

jects. The last decade has seen the spread of the transistor radio even to the depths of the Sierra Madre, with the result, among other things, that more are learning Spanish. The cast-iron corn mill is now used to grind *masa* (tortilla dough), saving hours of kneeling at a *metate*, or millstone, grinding by hand. Pedal sewing machines and metal containers are common. Nevertheless, most northern Mexican Indians express concern that their lifeways are dying. They fear that loss of community lands, and the cultural seduction of their youth will mean that their days as separate peoples are indeed numbered.

BIBLIOGRAPHY. CARL LUMHOLTZ, *Unknown Mexico*, 2 vol. (1902), gives the first modern account of North Mexican groups and is still a major source. The best modern source is EVON Z. VOGT (ed.), *Handbook of Middle American Indians*, vol. 8 (1969), which contains articles covering each of the groups. A basic source on the history, cultural geography, and ethnography are the following bulletins of the *Ibero-Americana* series that cover western Mexico in depth: RALPH L. BEALS, *Comparative Ethnology of Northern Mexico Before 1750* (1932) *Acaxee: A Mountain Tribe of Durango and Sinaloa* (1933) and *Aboriginal Culture of the Cáhita Indians* (1943); CARL O. SAUER, *Distribution of Aboriginal Tribes and Languages in Northwestern Mexico* (1934) and *Aboriginal Population of Northwestern Mexico* (1935); ALFRED L. KROEBER, *Uto-Aztecan Languages of Mexico* (1934). EDWARD SPICER, *Cycles of Conquest: The Impact of Spain, Mexico, and the United States on the Indians of the Southwest, 1533–1960* (1962), presents the only overview of change since colonial times. CAMPBELL W. PENNINGTON, *The Tarahumara of Mexico* (1963) and *Tepehuan of Chihuahua* (1969), are very good for material culture and ethnogeography of the northern sierra. CARL LUMHOLTZ, *Symbolism of the Huichol Indians* (1900), furnishes much on Huichol gods and belief, as does R.M. ZINGG, *The Huichol: Primitive Artists* (1938). A definitive work on Cora religion is found in K.T. PREUSS, *Die Nayarit-Expedition* (1912), in German.

(T.B.H.)

North Sea

The North Sea is a shallow, northeastern arm of the Atlantic Ocean located between the British Isles and the mainland of northwestern Europe and covering an area of 220,000 square miles (570,000 square kilometres). It is bordered by the United Kingdom and the Orkney Islands to the west; the Shetland Islands to the north; Norway and Denmark to the east; and West Germany, The Netherlands, Belgium, and France to the south. It is connected to the Atlantic by the Strait of Dover and the English Channel and opens directly onto the ocean between the Orkney and Shetland islands and between the Shetland Islands and Norway. The Skagerrak, Kattegat, and Danish sounds provide the connection between the North Sea and the Baltic Sea.

The North Sea has long been important as one of Europe's most productive fisheries. Its grounds alone normally account for over 5 percent of the total world commercial catch. It also serves as a prominent shipping zone among European countries and between Europe and the Middle East. These functions may soon be overshadowed, however, by the enormous reserves of petroleum and natural gas recently discovered beneath the sea floor. Estimates of their size hold out the promise that the North Sea's deposits may end western Europe's dependence upon the Middle East for oil and gas by the 1980s.

The North Sea in history

The North Sea has had a strong influence on European history. Because of its long coastline, and the rivers emptying into it, it has been readily accessible to many areas, providing highways of commerce and of conquest. It was the scene of early development of maritime trade, dating back to before the days of the Hanseatic League. Its waters have protected the British Isles from invasion from the Continent for over one thousand years, yet the North Sea also has served as a springboard for the growth of the overseas empires of the countries bordering on it. Without the interchange of people, goods, and ideas made possible by the existence of the North Sea, the cultural development of northwestern Europe after the Middle Ages might have been greatly retarded.

Geologic history. The extent of the North Sea and the level of its water surface have varied considerably over geologic time. At the end of the Pliocene Epoch (about 2,500,000 years ago) the North Sea Basin south of Dogger Bank was part of the European mainland, and the Rhine River—joined on its left bank by the Thames—emptied into the sea about 250 miles north of modern-day London. During the next 2,000,000 years of the Pleistocene Epoch, the ice sheet advanced and retreated several times and deposited a thick layer of clay on the sea floor. At the time of the greatest advance, the ice covered all of the North Sea from a line joining the Thames estuary with the Dutch coast. The final retreat took place about 6000 BC; and, some centuries later, the expanding sea area broke across the land bridge that linked Britain with France, and the waters joined with the English Channel. The present coastlines of the North Sea probably were not established until some 3,000 years ago.

The physical environment. *Relief features.* Few parts of the North Sea are more than 300 feet in depth. The floor dips to the north and is generally irregular. In the south depths measure less than 120 feet; many shallow, shifting banks, presumably of glacial origin, have been reworked by tidal currents. These present serious navigational hazards. Off northern England the vast moraine (glacial deposit of earth and stones) known as Dogger Bank is covered to depths of only 50 to 100 feet. This is the location of one of the finest fishing areas in the sea. In contrast, the waters deepen in the Norwegian Trench, an unusual trench that runs parallel to the coast of southern Norway from north of Bergen around to Oslo. It is between 15 and 20 miles wide and is some 600 feet deep in the vicinity of Bergen and over 2,400 feet deep in Skagerrak at the entrance to the Baltic Sea. There are also some deep trenches in the western part of the North Sea, including Devil's Hole off Edinburgh, where depths exceed 1,500 feet, and Silver Pit, 318 feet deep, off the Wash of England. These trenches may have been formed at the time of the last glaciation when parts of the North Sea were free of ice, and rivers coming off the mainland could have eroded deep channels in the basin floor.

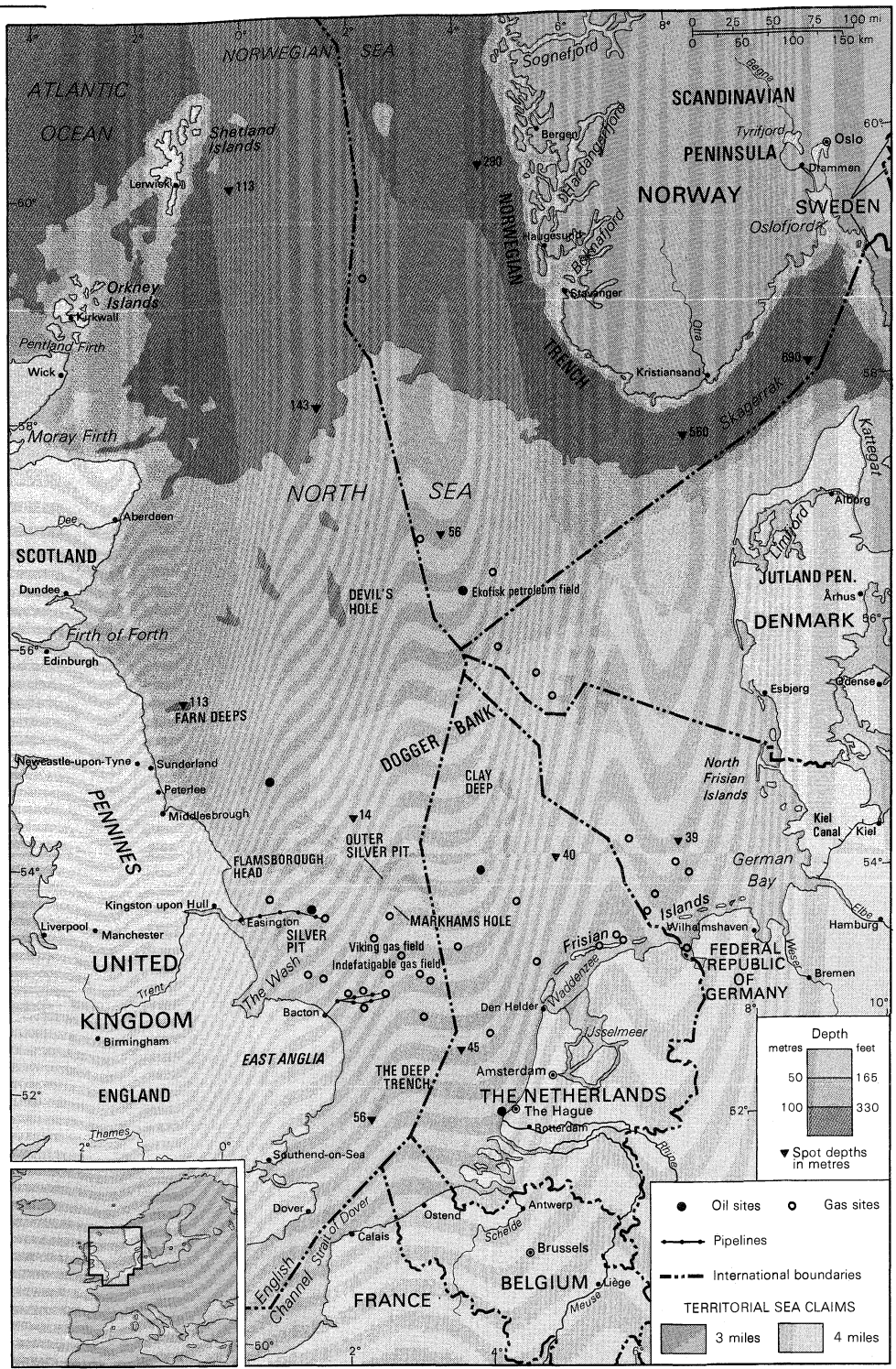
Dogger Bank and the Norway Trench

Coasts. There is a marked difference between the rugged upland coasts of the north and the regular lowland coasts of the south. The glaciated mountain coastline of Norway north of Stavanger is broken by fjords (narrow inlets between steep cliffs), headlands, and an offshore fringe of thousands of rocks and islands. Below Stavanger the coast is less precipitous and there are fewer islands. Scotland's east coast is also composed of uplands—although it is less broken—and the resistant rocks continue south into England. In the vicinity of Flamborough Head the cliffs are lower and their less resistant clays are subject to extensive erosion. In the fens area of East Anglia the coast is low and marshy, as it is in the delta region of The Netherlands. Most of the southern and southeastern coast is straight and sandy. The low, offshore Frisian Islands stretch from IJsselmeer (Zuiderzee) in The Netherlands to southwestern Denmark.

Currents and tides. The North Sea waters are affected by the warm current of the North Atlantic Drift, which moves northward along the western side of the British Isles and enters the Norwegian Sea. Atlantic waters with salinities exceeding 35 parts per thousand enter the North Sea through the English Channel and between the Shetland Islands and Norway. Colder, less saline waters come from the Baltic Sea through Skagerrak, creating a counterclockwise circulation in the basin. Salinities generally range between 34 and 35 parts per thousand, with higher readings occurring off the British coast and lower readings off Norway. Large quantities of freshwater also enter through the Rhine, Thames, and other rivers.

Climate

Average air temperatures vary in January from 32° F (0° C) to 40° F (4° C) and in July from 55° F to 64° F (13° to 18° C). Winters are stormy and gales are frequent. Tidal ranges average between 13 and 20 feet along the British coasts and in the southern estuaries, while the range to the north and east is less than ten feet. Because of the low-lying nature of much of the southern coast, abnormal tides can be disastrous. In 1953 a storm surge of nearly 11 feet above the mean high-water level inundated large areas of the delta region of The Netherlands and caused over 2,000 deaths.



Submarine relief and gas and oil resources of the North Sea.

Resources and their exploitation. *Fisheries.* The constant mixing of waters in the shallow sea basin provides a rich supply of nutrient salts upon which the lower forms of marine organisms—the basis of the sea's food chain—are dependent. The resulting abundance of plant and animal plankton supports a varied and rich supply of commercially valuable fish. The annual North Sea catch exceeds 3,300,000 tons. Herring and mackerel account for almost half of the total catch, and the rest consists of pout, cod, sand eels, whiting, haddock, and plaice, a type of marine flatfish. The major fishing countries and their approximate percentages of the annual catch are Norway 33 percent, Denmark 30 percent, the United Kingdom 11 percent, Sweden 7 percent, The Netherlands 5 percent, the Soviet Union 4 percent, France 3 percent, and West Germany 3 percent.

Oil and gas. Important discoveries of oil and natural gas beneath the sea floor have been made since 1959, when a major natural-gas field was located in the northeastern portion of The Netherlands. Subsequent exploration extended to the shallow waters off the Dutch coast, and in 1966 discoveries were made in the deeper waters off the southeast coast of England. One year later gas from offshore wells was commercially produced by the British. Since then, gas development has also been carried out by Norway, Denmark, and West Germany, in some cases at depths of nearly 200 feet. Although estimated potentials of the offshore wells frequently change, it is

clear that these reserves will have a major impact on the economies of the North Sea countries. The giant Ekofisk petroleum field that was discovered off of southern Norway in 1970 has an estimated capacity of 300,000 barrels per day, and other fields have since been located off Scotland, Norway, and The Netherlands.

Man's influence. The resources of the North Sea are increasingly exploited as the level of the western European economy rises. The volume of shipping to bordering countries and to the Baltic is steadily growing, thereby generating problems of navigation that are exacerbated by the more than 400 gas and oil wells located offshore. Shipping and industrial exploitation in coastal areas contribute to pollution problems, particularly in the southern portions of the North Sea. In 1971 the danger of oil pollution from crippled oil tankers, pipelines, or exploration for petroleum prompted the United Kingdom to legislate against such polluters.

The development of gas and oil industries has resulted in the leasing to major oil companies of large areas of the seabed, and also in offshore boundary disputes, particularly between West Germany, The Netherlands, and Denmark. The fisheries have long been threatened from overfishing, a trend accentuated in the 1970s by the use of more efficient vessels and harvesting techniques.

BIBLIOGRAPHY. L. DUDLEY STAMP, *Britain's Structure and Scenery*, 2nd ed. (1946), a geological description of Britain with considerable discussion and maps of the North Sea area; C.A.M. KING, *Oceanography for Geographers* (1962), an oceanographic text with emphasis on features of the world ocean, including some interesting discussions of the North Sea; *Bulletin Statistique des Pêches Maritimes* (annual), contains extremely detailed statistics of North Sea fisheries; LEWIS M. ALEXANDER, *Offshore Geography of Northwestern Europe* (1963), includes considerable discussions on fisheries and on coastal types in the North Sea area; M.N. HILL (ed.), *The Sea*, vol. 3, *The Earth Beneath the Sea* (1963), a very thorough treatment of geological history of marine areas; contemporary reports from *World Oil* and the *Oil and Gas Journal*, two periodicals that carry concise, current materials on oil and gas developments in the North Sea.

(L.M.A.)

Northwest Coast Indians

The most sharply delimited culture area of native North America was the Northwest Coast. It covered a long narrow arc of Pacific coast and offshore islands from Yakutat Bay in the northeast Gulf of Alaska south to Cape Mendocino in modern California. Its eastern limits were the crest of the Coast Ranges from the north down to Puget Sound, the Cascades south to the Columbia, and the coastal hills of Oregon and northwest California. The Kuroshio (Pacific Ocean current) offshore warms the coast and deluges it with rain. The northern Coast Range, cresting at heights of 5,000 feet and more, rises steeply from the sea and is cut by a myriad of narrow channels and fjords. The shores of Puget Sound, southwest Washington, and the Oregon coast hills are lower and less rugged.

Physical aspects of the Northwest Coast

Coastal forests are dense, predominantly coniferous: spruces, Douglas fir, hemlock, red and yellow cedar, and, in the south, coast redwood. These forests support an abundant fauna. Most important from the cultural point of view was the aquatic fauna, for it was on this that the areal culture depended primarily. Five species of salmon; herring; oil-rich "candlefish," or eulachon; smelt; cod; halibut; and mollusks abounded.

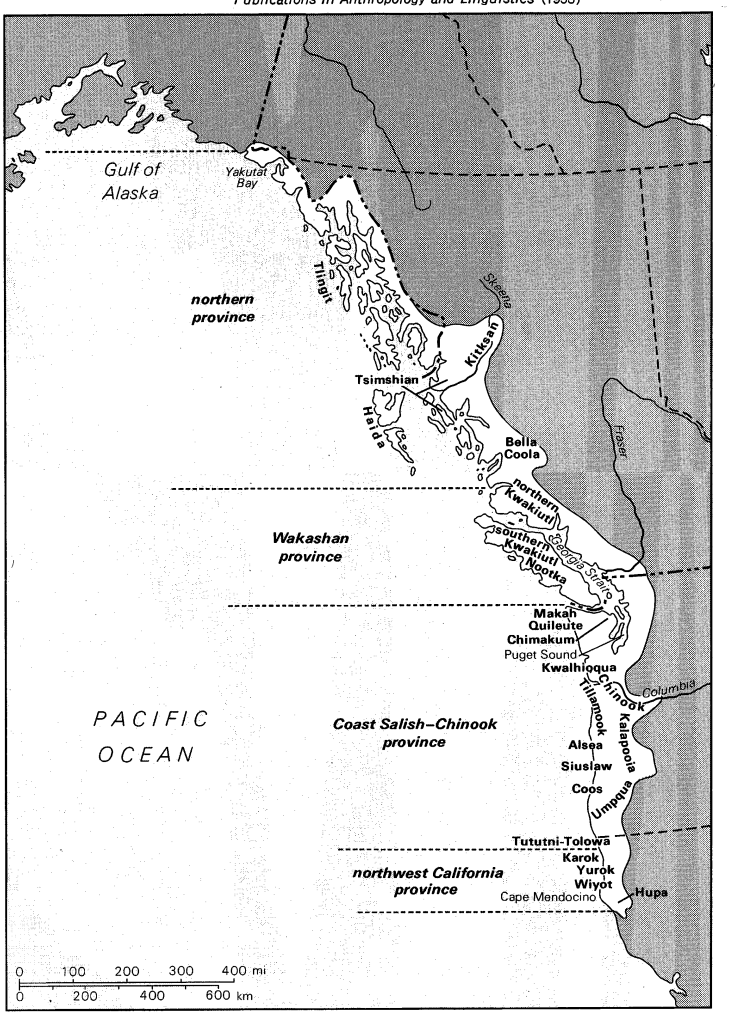
The peoples of the Northwest Coast linguistically consisted of a series of units related to widespread "stocks" of native North America (see NORTH AMERICAN INDIAN LANGUAGES). From north to south the following linguistic divisions occurred: Tlingit; Haida; Tsimshian; northern Kwakiutl, or Heiltsuq; Bella Coola; southern Kwakiutl; Nootka; Coast Salish; Quileute-Chimakum; Kwalhioqua; Chinook. Then along the Oregon Coast and northwest California a series of small divisions occurred: Tillamook, Alsea, Siuslaw, Umpqua, Coos, Tututni-Tolowa, Yurok, Wiyot, Karok, and Hupa.

Major Indian groups

Culturally, Northwest Coast groups can be classified into four subareal units, or "provinces": the northern one, including speakers of Tlingit, Haida, Tsimshian, and the Tsimshian-influenced Haisla (northernmost Heiltsuq or Kwakiutl); the Wakashan province, including all other Kwakiutl, the Bella Coola, and the Nootka; the Coast Salish–Chinook province, which included various enclaves of other speech down to the central coast of Oregon; and the northwest California province plus the Athabascan-speaking Tututni-Tolowa.

The Northwest Coast was densely populated. Estimates of density in terms of persons per square mile mean little in a region where long stretches of coast were uninhabitable cliffs rising from the sea. But early historic sources indicate that many villages had hundreds of inhabitants. One conservative population estimate of 129,000 persons on the coast at the dawn of the historic period must have represented nearly the maximum that the area could support without improvement of the already complex technology.

From H. Driver *et al.*, *Indiana University Publications in Anthropology and Linguistics* (1953)



Distribution of Northwest Coast Indians.

TRADITIONAL CULTURE PATTERNS

Social structure. The Northwest Coast was the outstanding exception to the anthropological truism that "hunting-and-gathering" (in this case "fishing-and-gathering") cultures are characterized by simple technologies, sparse possessions, and egalitarian, loosely organized societies consisting of small bands comprising small total populations. In this area, complex patterns of culture were the rule.

Familial and group organization

The nuclear family—a man, his wife or wives, their children or, in the northern provinces, the man's sisters' sons—was the basic production unit. In the native view, the significant social unit was the local group—that is, a group of men who considered themselves related and who formed a corporate entity holding title to fishing

places, berry-picking and hunting grounds, habitation sites, and a host of incorporeal rights, such as names, songs, dances, and, especially in the north, totemic representations called "crests." Members of each group were graded in an integrated sequence from high to low, according to closeness to the direct line of descent from the group ancestor. The highest in rank, invariably holder of a special title that in each language translated into English as "chief," was administrator of the group's properties. It was he who set the time for the move to the salmon-fishing station, decided when the weirs and traps should be built, when the first catch should be made and the rite propitiating the first salmon of the season celebrated, when other groups should be invited to feasts, and so on. He had many prerogatives and sumptuary privileges, but he was expected to administer efficiently and to tend to social and ritual affairs for the general welfare of his group.

From Tlingit country in the north at least as far south as Puget Sound and perhaps farther, several such local groups assembled at a site in some sheltered cove protected from winter winds to pass the winter. Food-gathering activities were limited by weather but were unimportant; the stores of salmon dried in the fall were adequate. Practice of arts and crafts, jollity, and feasts and ceremonials were the order of the day. These assemblages of several local groups at winter village sites are often called "tribes," but it must be noted that such units were not politically integrated, for each of the component local groups retained its economic and political autonomy. For ceremonial purposes, though, the local groups were ranked in series from highest to next highest, and so on.

Stratification. A signal feature of Northwest Coast society was the emphasis on each individual's hereditary social rank. His position within his local group depended on his genealogical closeness to the legendary group ancestor. When several groups assembled at a common winter site to form a "tribe," the relative rank of the individual's group also was another factor important for ranking purposes.

Individual social ranking

Indian informants tend to oversimplify the situation in casual conversation, describing their former society as a class-structured one with a class of "chiefs," a class of commoners, and below these two a class of slaves. It is true that slaves (*i.e.*, war captives) formed a special social division, but the division into two great classes of "chiefs" and "commoners" is an inadequate explanation. The fact is that each person had his particular hereditarily acquired status, which placed him within his group as though he stood on a step of a long staircase of statuses, with the eldest of the senior line on the highest step, the most remotely related at the bottom. Strictly speaking, each person was in a class by himself.

Nominally, those of high rank were said to have had vast authoritarian powers, and group members of low degree have been described as serflike. In actual fact, mature persons of the latter sort voiced their opinions on group affairs, for they held interest in group properties. The chief refrained from abusing them because they were his kin and also because he was aware that he needed their assistance. Many strong arms and sturdy backs were needed to obtain, assemble, and position the heavy materials needed to build or repair a house, to construct fish weirs and traps, and to launch and paddle the chief's huge dugout canoe. Many singers, dancers, and attendants were necessary to stage an important ceremonial properly. Many bold warriors were needed to defend the group's wealth against foemen. There was enough flexibility in the social structure so that those of low rank could abandon an abusive chief and reside with kindred elsewhere who would welcome them.

Slaves usually were persons captured in childhood and taken or traded so far from their original homes that they had little hope of finding their way back. They were mere chattels, who might be treated well or ill, traded off, slain, or freed at their owner's whim.

The statuses of group members were hereditary, but they were not automatically assumed at birth. They had to be formally and publicly assumed at a "potlatch," a performance given by all coast groups north of the Columbia River. The term comes from a widespread trade jargon and means "to give." A potlatch always involved invitation of another group or groups, who were received as guests with great formality; they served as witnesses to the announcements by the host chief at the assumption or bestowal of prerogatives, such as noble titles, crests, or ceremonial rights, and then were given gifts. The previously assumed statuses of the guests were recognized in the gift-giving, for distribution was made in the order of their rank sequence, and the more splendid gifts were given to guests of highest status. Not only were titles and other honours announced for the host chief or his heir and for his close kin of high position but children of low-rank group members were awarded names from the group stock and, at times, minor prerogatives. Participation of all members in major or minor roles in the proceedings also served to identify them with the social unit. There were some regional variations: in the northern province, a major potlatch was part of the cycle of mortuary observances after the death of a chief, at which his heir formally assumed his rights; in the Wakashan and Salish regions a chief gave such affairs to bestow rights on an heir apparent before his own demise.

Potlatch

Some early anthropologists, and a few modern ones, considered potlatch to be an economic enterprise in which the giver expected to recover a profit on the goods he distributed when his various guests potlatched in their turn. This actually was an impossibility because only a few guests of highest rank would ever stage such affairs and invite their former hosts; those of intermediate and low rank never did so, yet the total amount of gifts bestowed on them was considerable. Indeed, before white traders came and made great quantities of trade goods available, potlatches were few, whereas feasts, though also formal but not occasions for bestowing titles and gifts, were very frequent.

Socialization and education. An interesting aspect of Northwest Coast culture was the emphasis on teaching children etiquette, moral standards, and other traditions of social import. Every society has processes by which children are taught the behaviour proper to their future roles, but often such teaching is not overtly a deliberate process. On the Northwest Coast, particularly northward of the Columbia, children were instructed formally. This instruction began at an age when modern educators would consider children too young to learn effectively—while they were still in their cradles. Children born to high statuses were given formal instruction throughout childhood and adolescence. They had to learn not only routine etiquette but also the lengthy traditions by which the rank and privileges of their particular group were validated and many rituals including songs and formulaic prayers. It was not only the parents (or, in some groups, the mother's brother) who taught children. All elder relatives, particularly grandparents, participated. The educative procedure did not consist of dry, barebones moral lectures. Some of it was given in the form of folktales, amusing and entertaining but with pointed morals: the troubles of the anti-hero, Raven, in the tales were obviously the result of his dissolute way of life, his laziness, his gormandizing, and his lechery.

Early training of youth

Other socialization processes, those that involved public recognition of the attainment of new status, were usual among Northwest Coast groups. One variety included ritual observances considered necessary at each critical stage in a person's lifetime. At the birth of an infant, at a girl's attainment of puberty (there were no boys' puberty rites in the area), and at death, it was considered that persons involved might be in danger or might be dangerous to the society at large. A newborn infant was believed to be in danger from supernatural causes; the infant's parents were both in danger and dangerous. A girl at puberty was similarly viewed, as were the close kin of a deceased person and those who actually participated in preparing and disposing of the body. Such perils were avoided by isolating the persons involved, either within a boarded-off cubicle in the house or in a makeshift hut out in the woods and by limiting his or her diet to old dried fish and water. At the conclusion of the isolation period some sort of

formal ritual purification was carried out, such as ceremonial bathing. The intensity of the restrictions varied considerably, not only in different parts of the coast but within individual groups. Often the pubescent daughter of a chief, for example, was secluded for many months, whereas her low-ranking kinswoman might have to suffer only a few days of confinement.

Over most of the coast there was a very great fear of the dead. A body was removed from the house through some makeshift aperture other than the door, to be disposed of as rapidly as possible. Only in the northern region were bodies of chiefs set up in state for several days while the clan dirges were sung. Disposal of the dead varied. In the northern province cremation was practiced (anciently, interment was customary). In the Wakashan and part of the Coast Salish areas large wooden boxes suspended from branches of tall trees or placed in rock shelters served as coffins. Other Coast Salish deposited their dead in canoes set up on stakes. In southwest Oregon and northwest California interment in the ground was preferred.

Economic systems. Of the various distinctive attributes of Northwest Coast culture, one of the most important was the highly efficient exploitation of natural resources. The resources, particularly the fisheries, were very bountiful; but they were scattered and not equally easy to exploit. Certain species of salmon, for example, ran in certain rivers at various times of the year. The important species for preservation for winter stores were the pink and the chum salmon. Because these species ceased to feed for some time before entering freshwater and their flesh thus had less fat content, they could be smoked and dried and kept for a long period of time. Other species, such as sockeye, coho, and the flavorsome chinook or king salmon, could be utilized immediately or dried and kept for a short period but could not be preserved the whole winter through. Therefore, the principal fishing sites were those along rivers and streams in which pink or chum salmon ran in the fall. In the spring of the year other sorts of fish became available in tremendous schools: herring, which came in to spawn in coves; "candlefish" (eulachon), which entered certain rivers; and, farther south, "smelt," which spawned on sandy beaches in summer. Elsewhere, in the summer months, bottom fishing for such species as halibut was a profitable enterprise on shallow offshore banks.

Reliance on fishing

To exploit the total resources of the region it was most efficient for the people to have various bases of operation. In the winter months when storm winds blew and heavy seas slammed against the coasts, shelter behind some point of land that broke the force of the sea was highly desirable. The Northwest Coast adaptation to this pattern was shifting residence. The people moved from one site to another, according to the season and according to the resources about to become available. This was not nomadism. The Indians moved systematically in certain seasons from one fixed site to another for convenient access to seasonal resources. Typically at these seasonal stations there were either permanent houses or permanent house-frames that could be covered over with planks brought along for the purpose. Occasionally, when the weather permitted or only a short stay was planned, makeshift shelters were erected.

Patterns of production and technology

Aboriginal Northwest Coast economy may be viewed as a system comprised of several mutually supporting subsystems. The first of these subsystems is the efficient techniques for taking fish and other marine resources and for preserving them. The second subsystem consisted in the construction of large rectangular plank houses that made possible the smoking and drying of fish during the torrential rains of autumn. Such houses also provided storage space for the bulky preserved foods. The third subsystem consisted of the water-transport complex. Canoes, both large and small, provided access to fishing grounds and the means of transporting preserved foodstuffs to other habitation areas. It was the combination of these subsystems that made the exploitation of local resources so efficient that the people were able to live with a wide margin of plenty.

Northwest Coast houses shared a few significant traits. All were rectangular in floor plan with plank walls and plank roof, and all but those of northwestern California were large structures designed for multi-family use. At the northern and southern extremes of the area deep central pits were dug within the house. In the north, houses were built on a nearly square plan—averaging about 50 feet wide by 55 feet long—and had gabled roofs; walls and framework were intermeshed to form a permanent structural unit. In the Wakashan province, on the other hand, the houses were rectangular—40 feet by 60 to 100 feet. Huge cedar posts with side beams and ridgepoles comprised the permanent framework, and to these were attached wall planks and roof planks that could be taken down, loaded onto canoes, and transported from one site to another. Some Coast Salish similarly built houses of permanent frameworks with detachable siding and roofing; their houses, however, had "shed-roofs," that is, a roof with only one slope instead of the two of the Wakashan house. Some Coast Salish houses were tremendously long, housing many people. Along the lower Columbia the typical house had a deep large rectangular pit, lined with planks, capped with a gabled roof. Only roof and gable ends showed above ground. The northwestern California house type was designed for single family only. Each house had low side walls of redwood planks and a three-pitch roof. Living space was the plank-lined central pit. Northwestern California also added a specialized structure: the men's combined clubhouse and sweat house, a common native Californian institution.

Water transport was highly important in the area. All groups made efficient dugout canoes. Northern groups, as well as the Kwakiutl and Salish down to Puget Sound, made dugouts with projecting bow and stern pieces, vertical "cutwaters," rounded sterns, and rounded hulls. The Nootka and some of their neighbours made vessels with projecting bow pieces, curving cutwaters, vertical sterns, and angular flat bottoms. Northwest Californian dugouts had upturned rounded ends, rounded hull, a carved seat and foot braces for the steersman. All types were made in different proportions for different purposes: large, beamy ones for moving people and cargo; shorter, narrow ones with racy lines for sea mammal hunting, and so on.

Crafts

Northwest Coast woodworking was facilitated by the natural abundance of easily worked timbers, especially the red cedar and the redwood. Trunks of these trees could be split into planks or they could be hollowed out into canoes, containers, or other useful objects. Along the Northwest Coast as far south as the Columbia, wooden boxes were made of red-cedar boards that were "kerffed"—cut nearly through transversely. The wood was steamed at these points until it was flexible enough to bend into the form of a box. Dishes often were hollowed out of pieces of wood, sometimes plain, sometimes in the form of animals and monsters. Also of wood were spoons and ladles (some were of horn), canoe bailers, trinket boxes, chamber pots, and masks used in ceremonials and rattles for musical accompaniment. A special character of Northwest Coast woodworking was the emphasis on symmetry, neatness of finish, and frequent decoration of the surfaces, with relief carving or with relief carving and painting. All of this woodworking was accomplished with rather limited tools, the principal ones being the adz, mauls and wedges, chisels, drills, curved knives, abrasive stones and sharkskin for polishing. Mountain-goat horn, mountain-sheep horn traded from the interior, and, in the south, elk horn were carved by essentially the same methods as wood.

Another highly developed craft was weaving. The inner bark of red cedar was stripped long and ribbonlike to be woven into mats and baskets in a checkerwork technique. The same bark was shredded into finely divided flexible hanks, which were twined together to make a slip-on rain cape shaped like a truncated cone. The softer inner bark of yellow cedar was made into robes. Persons of high status wore such robes edged with strips of sea-otter fur and a few strands of yarn made of mountain-goat wool. Salish of the Georgia Strait wove robes of mountain-goat wool and also of wool from a special breed of

shaggy little dog. The Chilkat, a Tlingit group, wove robes of mountain-goat wool that was twilled like basketry. Chilkat blankets bore representations of crests in blue, yellow, black, and white.

Twined basketry made from long, flexible splints split from spruce roots was made with great technical skill in several regions. Baskets so tightly woven as to be waterproof were made for cooking in the northern and northwestern Californian regions. Storage containers, receptacles for valuables large and small, and rainhats (Californians' hats were snug and caplike, worn only by women) also were woven. The Coast Salish specialty was coiled basketry.

Clothes

Dress patterns of the area were simple. Only the northernmost Tlingit and the Kitksan of the upper Skeena wore tailored buckskin clothing: breechclouts, leggings, and shirts for men; long gowns for women. Elsewhere, men wore robes of yellow cedar bark or of crudely tanned pelts in cold weather, rain capes in downpours, and nothing but ornaments on the infrequent bright sunny days. Ornaments, such as necklaces, earrings, bracelets, and anklets, were made of various materials, mostly shells, copper, wood, and fur. Some groups practiced tattooing.

Head-flattening was considered a beautifying process from the northern Kwakiutl region to the central Oregon coast. A newborn infant in its cradle had its head bound in such a way as to produce a long subconical form, a strong slope from the eyebrows back, or a distinctive wedge shape in which the back of the skull was flattened.

Belief and the aesthetic systems. *Religion.* Among no group or groups on the Northwest Coast did religion consist of an organized coherent body of beliefs in and attitudes toward the supernatural. Rather, there were several quite unrelated concepts that provided the widespread bases for various kinds of religious activity.

Fish spirits

One concept was that salmon were supernatural beings who voluntarily assumed piscine form to sacrifice themselves annually for the benefit of mankind. On being taken, the spirits of the fish returned to their home beneath the sea, where they were reincarnated if their bones were returned to the water. If offended, however, the salmon-beings would refuse to return to the river. Hence, there were numerous specific prohibitions on acts believed to offend them and observances designed to propitiate them, chief of which was the first salmon ceremony. This rite varied in detail along the coast but invariably involved honouring the first salmon of the main fishing season by sprinkling them with eagle down, red ocher, or some other sacred substance, welcoming them in a formal speech, cooking them and distributing their flesh, or morsels of it, communion-fashion, to all the members of the local group and any guests. The maximal elaboration of this rite was found in northwestern California, combined there with first fruits observances, and dances in which lineage wealth was displayed, in what have been called world-renewal ceremonies. Elsewhere the first salmon rituals were less elaborate but still important (except among the Tlingit, who did not perform them).

Another concept was that of acquiring personal power by seeking contact with a spirit—as through prayer and a vision. Among Coast Salish all success in life—whether hunting, woodworking, accumulating wealth, military ventures, or magic—was bestowed by spirits encountered in the spirit quest. From his spirit or spirits each person acquired songs, special regalia, and dances. Collectively, the dances comprised the major ceremonials of these people; known as the Spirit Dances, they were performed during the winter months. In the Wakashan province and the northern one it was believed that remote ancestors on spirit quests had been rewarded with totemic symbols called "crests." Displaying these hereditary crests and recounting their traditional acquisition formed an important part of the potlatches. In the Wakashan area certain ceremonial cycles called for the dramatization of the whole tale of the supernatural encounter, including the spirit's possession of the seeker and the eventual exorcism of the spirit; such dramas were performed by what were called "dancing societies."

Shamanism differed from other acquisitions of supernatural power only in the nature of the power obtained; that is, power to heal the sick through extraction of disease objects or recovery of a strayed soul. It was commonly believed that some shamans, or medicine men, had power to cause these infirmities as well as to cure them. Witchcraft, to kill or make ill, was also believed to be carried out by malicious persons who knew secret rituals for that purpose.

Art. The Northwest Coast is noted for its art styles. In the northern province low-relief carving accented by painting was essentially an applied art. The motifs were the hereditary crests of the clans or parts of the crests, applied to the magnificent memorial poles and interior house-posts, painted on house-fronts and screens, wrapped around boxes and dishes, painted on basketry hats, woven into Chilkat robes, and carved on the handles of ladles and spoons, on halibut hooks, and even on the triggers of animal traps. There were differences within this style. Haida art tended to be massive, of highly conventionalized balanced elements, and slightly static. In Tsimshian carving and painting there was an effort to leave no open space in or between the conventionalized motifs: filler elements such as eye-designs and miniature figures were used intensively. Tlingit art was slightly less conventionalized, more vigorous by modern standards, with relatively little use of filler elements.

Wakashan representative art was more frankly sculptural than applied, and it was impressionistic and bold. There was a limited amount of simple geometric design on such things as whalebone clubs and whaling harpoon barbs. Their Coast Salish neighbours used some, but less, representative art, similar if cruder in style. On Puget Sound there was little if any representative art; the formless painted designs on the canoe boards were unlike anything else on the Coast. All that is known of Chinook art is represented by a few angular figures incised on mountain sheephorn bowls. In northwestern California art was limited to geometric patterns incised on elkhorn objects and shells.

MODERN CULTURAL DEVELOPMENTS

The impact of white man's culture on that of the Northwest Coast varied at different periods and in different regions. Maritime traders, searching for precious sea-otter pelts, purveyed Euro-American manufactures to the Indians; but the material objects affected native culture only slightly. The Indians picked and chose the articles that had meaning to them—those that could fit well into their existing culture patterns. They acquired steel blades, for example, that could be fitted to their adzes to cut more efficiently than the aboriginal stone or shell blades; they spurned axe and hatchet blades that required a drastic change in motor habits and coordination patterns. In other words, the Indians accepted what they wanted; they were under no compulsion to change their way of life. Contagious diseases—smallpox, venereal infections, and the rest, introduced incidentally—had more effect on native society. The abnormal rate of deaths forced unusual distributions of roles and status positions, involving frequent adoptions, allocation of various titles to the same person, and other makeshift compromises to maintain the social system despite rapid population decline.

The establishment of white trading posts had somewhat more effect; and the great pressures started when white settlers began streaming into western Washington, Oregon, Vancouver Island, and the lower Fraser River Valley about the middle of the 19th century. This foreign occupation was accompanied by removal of the Indians to small reservations in Washington and Oregon under the provisions of formal treaties. In the area that became British Columbia there were no treaties extinguishing Indian title to the land; land transfers were private affairs.

Economic problems

In the closing decades of the 19th century, the Indians were in dire straits. Divested of most of their lands and at the same time more and more dependent upon white American and British-Canadian manufactured goods, the Indians had to develop new economic patterns. The fur trade was inconsequential; logging and mining were still

underdeveloped, requiring skills that the Indians did not have. Northwest Coast concepts of wealth differed from the Euro-American ones, but there were enough general similarities so that the Indians could perceive and accept certain equivalences. Northwest Coast Indians thus came to be more disposed to enter the new economic system, working for wages in a dull day-after-day routine, something that most other North American Indians refused to do. There was at first, however, little hired work available —guiding prospectors, back-packing cargo over mountain passes, cutting cordwood for coastal steamers—until the canned salmon industry developed, principally from the Fraser River northward. It was this industry that most effectively deprived the Indians of their fishing economy by monopolizing salmon streams, while at the same time offering them entrée into the new economy. Indian labour was cheaper than Oriental because it did not have to be imported. The Indians also knew more about the habits of the salmon than anyone else. Of great importance was the fact that the commercial salmon fishery began with a very simple technology. As motive power changed from paddles and oars to two-cycle gasoline engines, to high-speed gasoline engines, to diesel engines; as harvesting changed from gill nets and crude "beach seines" to huge purse seines handled with power gear; and as navigation changed from eyeball piloting to navigation using tide tables, compasses, and charts, the Indians could learn the new skills along with white and Oriental fishermen. The problem now is that the Indians are largely committed to a short-season industry, which ties up capital in expensive boats and nets. When the salmon run is over, there is no significant source of income until the next year.

Effective missionary activity began in various parts of the coast at about the same time that administrative controls were established. Missionaries on the Northwest Coast as elsewhere have been very competent at directed culture change, teaching not only Christian precepts but also etiquette, values of work and sobriety, household hygiene, and a host of other things that the native needed to know in order to participate in modern culture. Formal schooling of Indian children was in the hands of missionaries on much of the coast for many decades.

The aggressive and warlike Northwest Coast Indians never mounted a major war against the whites. There were only a few isolated local conflicts. This pacific pattern was not due to cowardice but to a realistic appraisal of the vulnerability of their coastal villages to naval gunfire.

There were some nativistic movements—that is, attempts to preserve or resuscitate former valued concepts or activities: for instance, the southern Kwakiutl tried to revive potlatch on a splendid scale, despite administrative prohibitions; and the Canadian Coast Salish tried to continue their Spirit Dances. But there is a distinction between these limited efforts of the Northwest Coast Indians and the "revitalization" movements of the Great Plains Indians, who, with their Ghost Dance, sought by supernatural means to evict the oppressor culture and return completely to the "good old days."

In southeast Alaska and later on in coastal British Columbia a different type of organization was created, the "Native Brotherhoods," whose purpose is to foster cultural change. The accomplishments of the Alaska Native Brotherhood have surpassed those of the Canadian organization, but both have been effective in creating a sense of Indian unity and community of interest. The organizations also provide valuable training in modern political processes and negotiations. The Alaskan organization has been partially superseded by the Tlingit-Haida Association, which handles matters related to the successfully prosecuted land claims of those people. Leaders of the British Columbia Brotherhood also have been effective in legal suits for compensation for lands.

BIBLIOGRAPHY. There are only two books that treat Northwest Coast culture on an area-wide basis: P. DRUCKER, *Indians of the Northwest Coast* (1963), and *Cultures of the North Pacific Coast* (1965). The former emphasizes material culture, technology, and art; the second, social and ceremonial organization. There are many good descriptions of individual Northwest Coast divisions: A. KRAUSE, *Die Tlinkit-Indianer* (1885; Eng. trans., *The Tlingit Indians*, 1956); V.E. GARFIELD and PAUL S. WINGERT, *The Tsimshian Indians and Their Arts* (1967); T.F. MCILWRAITH, *The Bella Coola Indians*, 2 vol. (1948); FRANZ BOAS, *Kwakiutl Ethnography*, ed. by H. CODERE (1966); P. DRUCKER, *The Northern and Central Nootkan Tribes* (1951); H.G. BARNETT, *The Coast Salish of British Columbia* (1955); A.L. KROEBER, chapters on Yurok, Karok, and Hupa in *Handbook of the Indians of California*, 2nd ed. (1953).

(P.Dr.)

Northwest Frontier Province

The Northwest Frontier Province, the northernmost province of Pakistan, covers an area of 39,283 square miles (101,743 square kilometres) and is bounded by Afghanistan to the west and north, Jammu and Kashmir to the northeast, Punjab Province to the southeast and Baluchistan Province to the southwest. The population at the start of the 1970s numbered almost 7,600,000. Peshāwar is the provincial capital.

This rugged, mountainous land is noted for the fierce independence and ethnic pride of its Pathan population; those who live within the tribal territories are not subject to Pakistani law. It is a hard land, poor in natural resources, and unsuited to extensive agriculture. The region is strategically important because it contains the Khyber Pass, a gateway through which several times in the course of history invading armies from the north have marched on their way to conquer the Indian subcontinent. (For related physical features see HINDU KUSH [MOUNTAINS]; PAMIR MOUNTAIN AREA; KARAKORAM RANGE; and INDUS RIVER.)

History. *The early period.* The early history of the province relates to the ancient state of Gandhāra, which was comprised of the Vale of Peshāwar and adjoining areas. The kingdom was important because of its strategic location at the end of the Khyber Pass, the most direct and easily negotiable route to Afghanistan, Iran, and Central Asia. Gandhāra was annexed by the Persian Achaemenid Empire in the early 6th century BC, and remained a Persian satrapy until 327 BC. The region then passed successively under Greek, Indian, Indo-Bactrian, Śakan, Parthian, and Kushan rule.

Foreign conquests

Muslim administration was first brought to the region by the Turks, whose ruler, Sebüktigin, gained control of Peshāwar by AD 988. His son, Maḥmūd of Ghazna, invaded northern India several times between 1001 and 1027, bringing a large area of the present province, excluding Hazāra, into his Ghaznavid empire. In 1779 Muʿizz-ud-Dīn Muḥammad of Ghūr captured Peshāwar which remained under his rule until his death in 1206. After the decline of the Ghūrids, the region was held first by the Muslim Afghan dynasties and then by the Mughals. After the invasion of the Persian ruler Nāder Shāh in 1738, the territory remained under a loose form of Afghan Durrānī rule.

The Sikh invasions from the Punjab region of India began in 1818, after which the Sikhs increasingly made themselves the masters of the frontier territory until the coming of the British in 1849.

The British period. The northwestern frontier areas were annexed by the British after the Second Sikh War of 1849. They formed a part of the Punjab (including what is now the Indian state of Punjab) until the Northwest Frontier Province was created in 1901. The new province was administered from Peshāwar by a chief commissioner and agent to the governor general. The territory was not granted the measure of self-rule given to the other Indian provinces in 1919. In 1935, however, the Government of India Act raised its status to that of a governor's province and allowed provincial autonomy.

Independence. After independence in 1947, the region continued to exist as a separate province. The feudal states of Phūlra and Amb were merged with the province in 1950. In 1955, Northwest Frontier Province, along with other provinces, was amalgamated into West Pakistan province. After the dissolution of West Pakistan province in 1970, the region regained its former provincial status, with the states of Swāt, Dīr, and Chitrāl being added to its territory.

Landscape. *Relief and drainage.* The terrain is comprised of mountain ranges, undulating dissected submontane areas, and plains surrounded by hills. In the north, the general orientation of the ranges is longitudinal, while it is transverse south of the Kābul River, which runs from west to east. The Hindu Kush region in the northern part of the province is divided by the Kunar River into two distinct ranges—northern Hindu Kush and the Hindu Rāj. Tirich Mīr rises to 25,263 feet (7,700 metres) and is the highest peak of the northern Hindu Kush. The highest peak of the Hindu Rāj is Shāh Dok at 20,737 feet (6,321 metres). East of Tirich Mīr, the Hindu Kush is extensively covered with snow and lofty glaciers.

To the south of the Hindu Rāj is the rugged country of the Pānjkora, Swāt, and Kandia river basins. Hazāra District in the east contains the Lesser Himalayas and the sub-Himalayas, which form definite ranges broken by hilly country and small plains. The transverse arrangement of the ranges south of the Kābul is markedly represented by the Safed Koh, whose highest peak, Sakaram, rises to 15,620 feet (4,761 metres). This southern area is traversed by the Kurram, Tochi, and Gumal rivers.

The Vale of Peshāwar and Bannu Plain

The fertile Vale of Peshāwar is covered with alluvial sediments that have been subjected to erosion. The tableland in the Kohāt area south of Peshāwar is rugged and cut by streams; the borders of the Bannu Plain southwest of Peshāwar are hilly. Topography has resulted in the development in some places of fascinating examples of the trellis pattern of drainage. In the Dera Ismāīl Khān area, the land is mostly a dry plain intersected by torrents and bordered on the east by the *kachī*, or the narrow Indus riverine tract.

Soils. The northern ridges are generally associated with the Himalayan territory; they contain gneiss (a coarse-grained rock containing bands of minerals) and granite in upper Chitrāl. The transverse hills are mostly formed of nummulitic limestone (limestone formed during the Eocene Epoch between 54,000,000 and 38,000,000 years ago) and sandstone. Much of the Vale of Peshāwar is covered with surface gravels and alluvium, while the rocks of Kohāt are mostly sandstones surrounding outcrops of nummulitic limestones. The Bannu Plain is mostly composed of soft sandstones and conglomerates; the greater part of the Dera Ismāīl Khān plain is covered with gravels, alluvium, and limestone.

Climate. The climate is highly diversified, according to altitude. While the northern mountains have snowy winters and cool summers, the mean annual temperature increases markedly toward the south, rising from 72.9° F (22.7° C) at Peshāwar to 76.6° F (24.8° C) at Dera Ismāīl Khān. Precipitation falls in winter and spring. The mean annual rainfall is 18 inches at Drosh, 14 inches at Peshāwar, and ten inches at Dera Ismāīl Khān. While the rainiest season is usually from January to April, precipitation is generally variable, and droughts may occur in either summer or winter.

Vegetation and animal life. In the far north, where extremes of altitude and climate are experienced, the higher mountains are bare and rocky. Elsewhere in the north, the mountain slopes bear stands of evergreen oak and pine; and broad-leaved deciduous trees such as the plane and poplar grow on the warmer, sunnier slopes. There are also extensive mountain grasslands. The hills to the south are sparsely covered with bushes, acacia, and grasses. About 9 percent of the uncultivated area of the province is forested.

Population. *Demography.* In 1961 the province had a total population of about 7,600,000; there were slightly more males than females. The overall density of population was about 180 persons per square mile. Peshāwar and Mardān (northeast of Peshāwar) are the most densely populated districts—Peshāwar with a density of about 500 persons per square mile—while in 1961 Dera Ismāīl Khān Division had a density as low as about 108 persons per square mile.

Population groups. Muslims constitute 99 percent of the population. Pashto is the predominant language, except in Hazāra and Dera Ismāīl Khān districts, where Punjabi predominates. The province is inhabited by numerous Pathan tribes and *khēl*s, or clans, each taking great pride in its genealogy. The major tribes are the Yūsufzai, Utmān Khēl, Mohmand, Afrīdī, Orakzai, and Wazīrī. Others include the Moḥammadzai, Shinwārī, Zaimukht, Bangash, Khaṭak, Bannūchī, and Maḥsūd.

The Pathans

Patterns of settlement. Only 10 percent of the overall population is urban. This percentage varies from 32 percent in Peshāwar District to only 5 percent in Hazāra District northeast of Peshāwar. Urban growth is generally slower than that of Pakistan as a whole. The province has only one city, Peshāwar; there are seven large towns with populations of between 25,000 and 100,000. There is a generally low percentage of migrants, except in Peshāwar District where the immigrant population numbered 34 percent in 1961.

Administration. *Government structure.* The province is divided into Malakand, Peshāwar, and Dera Ismāīl Khān divisions, which are subdivided into a number of districts. Each district is broken down into several *tahsils* (smaller administrative units). Between the settled districts to the south and the Afghan border is the tribal territory, whose inhabitants enjoy a large measure of independence. The tribal territory is divided into the Mohmand, Khyber, Kurram, North Wazīristān, and South Wazīristān agencies, each headed by a political agent.

The governor is the chief executive. In the directly administered areas, the provincial secretariat is headed by a chief secretary. A commissioner is in charge of each division, and the districts are headed by deputy commissioners. At the district level, judicial functions are performed by district and sessions judges, civil judges, and magistrates. Magistrates also perform some executive functions. In each *tahsil*, an official called the *tahsildār* collects revenue and discharges judicial and executive functions.

The tribal territory. In the tribal territory tribesmen are free to rule themselves according to their own customs. Political agents have power to award or withhold subsidies and to control entry into and departure from the tribal territory. The agent—aptly described as "half ambassador and half governor"—reports to the divisional commissioner and the Ministry of Home and Kashmir Affairs of the central government.

Counselling, judicial functions, intertribal liaisons, and communal dealings with the political agent are performed by the *jirgah* or council of elders. The selection of the *jirgah*'s members is based on tradition; their decisions are unanimously arrived at by consensus.

Social conditions. *Education.* The tribal order is inherently resistant to social change, and educational progress is quite slow. The literacy rate among the total population is lower than that of Pakistan.

There were almost 4,900 primary and middle schools with an enrollment of 480,000 students. High schools and colleges numbered 280 and had an enrollment of about 160,000 students. The university at Peshāwar, together with its constituent professional colleges, has a total enrollment of more than 8,000 students. There are far fewer female than male students.

Health services. Health facilities are inadequate, to the extent of being meagre. In the early 1970s the authorities assigned a high priority to both preventive and curative medical facilities and to the establishment of rural health centres. Among preventive schemes, the greatest emphasis was placed on malaria eradication. A family-planning program was started on a large scale in 1965, after which the birth rate was estimated to have declined to 45 per 1,000.

The government social welfare development program is implemented by 75 urban and 30 rural community development centres. The program aims at inculcating among the people a spirit of self-help, at developing leadership qualities, and at promoting the acquisition of skills.

The economy. The province is poor in natural resources but receives developmental allocations from the central government. The development strategy of the provincial government is to concentrate on productive schemes of a short gestation period. The main emphasis is

placed on agriculture, education, health, and transport and communications. In the early 1970s, emphasis was placed on establishing the development of cooperative and rural credit, on food storage, and on land reform.

Agriculture. The economy is essentially agricultural. Agriculture contributes 40 percent of the total gross provincial product and employs 80 percent of the population either directly or indirectly. Out of a total of 10,700,000 acres, 34 percent is cultivated. The net sown area is 81 percent and the current fallow area 19 percent of the cultivated area. Of the uncultivated land, 40 percent could be cultivated if irrigation or other facilities were made available. Irrigation is carried out on 38 percent of the cultivated land.

Principal crops

Wheat, maize (corn), sugarcane, and tobacco are the major crops. Other crops include millet, barley, rice, and cotton. Wheat production is highest in Peshāwar, Bannu, and Mardān districts, and Hazāra, Mardān, and Peshāwar districts are important maize-producing areas. Peshāwar and Mardān are also important regions of production of sugarcane and tobacco.

Industry. There is little industry; the province contains only about 3 percent of the manufacturing industries located in Pakistan. Industries in the province, which provide employment for more than 27,000 workers, include the manufacture and refining of sugar, the canning and preservation of fruits and vegetables, tobacco processing, and the manufacture of arms and accessories. Other products are cotton textiles, cement, vegetable ghee (margarine), nonmetallic mineral products, furniture, and milled grains. It is believed that the province may contain valuable mineral ores, but prospecting was still in progress in the early 1970s.

Transport and communications. In view of the strategic importance of the region, the British government constructed railways and roads across it, despite tribal opposition and the difficult nature of the terrain. Since independence, these facilities have been improved to help to promote economic growth. The province has 179 miles of broad-gauge and 196 miles of metre-gauge railway track. The length of paved roads totalled almost 4,300 miles in 1970.

Cultural life. The Pathans have a rich cultural heritage in which they take great pride. *Pakhtūnwalī*, or *pashtūnwalī* (the way of the Pathans) constitutes a code of honour, which imposes three chief obligations—*badal*, or revenge, the most important and binding aspect of *pakhtūnwalī*; *nanawātai*, is the right to seek asylum; and *maelmastyā*, which entitles a stranger or even an enemy to hospitality.

The code of honour

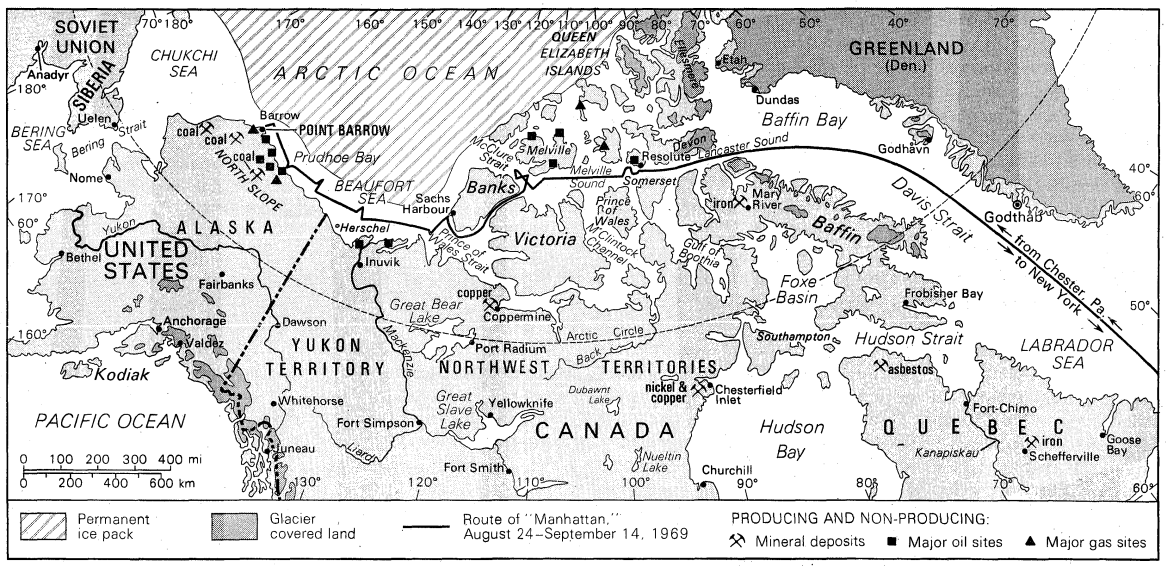
The Pathans are an intensely religious people who uphold a long tradition of preserving their freedom. Their council of elders, the *jirgah*, settles disputes and maintains high standards of justice. In the tribal order the position of the *malik* or *khān*, the head of the tribe, is most distinguished. Each village has one or more guest house, *ḥujrah*, which also serves as a club or town hall. Their festivals are mostly of a religious nature, but entertainments such as tent pegging and the whirling Khaṭak dance demonstrate the people's martial spirit. Folklore is generally related to the theme of Pathan bravery or to other traditional characteristics and often strikes an ethical note. The everyday dress of the Pathans consists of a long shirt, waistcoat, *shalwār* (trousers), and turban. Firearms are often carried. Women usually remain in *pardah* (*i.e.*, secluded from public view).

BIBLIOGRAPHY. Numerous works, principally by British writers, emphasizing strategic aspects include: W.P. ANDREW, *Our Scientific Frontier* (1880); T.H. HOLDICH, *The Indian Borderland, 1880–1900*, 2nd ed. (1909) and *The Gates of India* (1910); A. VINCENT, *The Defence of India* (1922); and C.C. DAVIES, *The Problem of the North-West Frontier, 1890–1908* (1932). General treatment of the area, as a part of the Indo-Pakistan subcontinent, occurs in E.J. RAPSON (ed.), *The Cambridge History of India*, vol. 1 (1922, reprinted 1935); and O.H.K. SPATE and A.T.A. LEARMONTH, *India and Pakistan*, 3rd ed. (1967). The Imperial Gazetteer of India, Provincial Series, North-West Frontier Province (1908), is a reference work on the region. DAVID DICHTER, *The North-West Frontier of West Pakistan* (1967), provides useful geographical coverage. An informative work on landforms and land use in Pakistan is the COLOMBO PLAN COOPERATIVE PROJECT, *Landforms, Soil and Land Use of the Indus Plain* (1958). Basic statistical data is contained in the provincial series of the Northwest Frontier Province of the *Census of India* up to 1941, and in the *Census of Pakistan*, 1951 and 1961; in *Provincial Season and Crop Reports*; and in the *Provisional Census of Manufacturing Industries of West Pakistan 1967–68* (1970). The *Economic Survey and Outlook: North West Frontier Province* (1971), is a useful government publication.

Northwest Passage

Since the end of the 15th century, it has been the elusive goal of Western man to establish a commercial sea route north and west around the American land barrier encountered by Christopher Columbus. This search, by many European nations, for a short water route to the Orient wrought momentous changes in world history. It brought about European colonization of the New World, shifted the centre of world trade from the Mediterranean to the Atlantic, and changed man's perspective of the world in which he lives. Over the centuries, the quest for a Northwest Passage evolved into scientific expeditions by intrepid Arctic explorers, who eventually found a hazardous route through Canada's Arctic Islands. Supply convoys began making limited summer use of it after World War II, as the U.S. and Canadian governments established Arctic weather stations for transatlantic and polar airline flights and built radar stations, air fields, and other defense installations in the frozen north.

With discovery of vast oil deposits in the Alaskan Arctic in 1968, the Northwest Passage again became a goal of global commercial significance. In the fall of 1969, the giant experimental icebreaking oil tanker ss "Manhattan" became the first vessel to batter straight through the pack ice covering the passage, proving that especially designed ships could operate the year round in the Arctic. The economic feasibility of such projects is less certain, due as much to complex governmental policy issues as to the harsh environment and tender ecology of the Arctic. Deployment of a fleet of mammoth icebreaking tankers to haul Alaskan oil through the Northwest Passage would create a new world-trade route, nearly halving the distance between Europe and Asia. Such an accomplishment would realize an objective that has eluded man since King Henry VII of England sent John Cabot in search of a northwest route to the Orient in 1497. Five years earlier, Columbus had set out in search of a westward route after conquest of the Middle East by the Ottoman Turks in the mid-15th century disrupted Europe's overland routes to the East. Vasco da Gama sailed south around Africa to India in 1498; Ferdinand Magellan sailed southwest around South America to the East Indies in 1521; and Dutch explorers vainly sought a northeast passage around Russia. But it was the Northwest Passage that captured the imagination of many of the world's famed explorers, including Jacques Cartier, Sir Francis Drake, Sir Martin Frobisher, and Capt. James Cook. All met with failure, and many with disaster. Sir Humphrey Gilbert, whose treatise on the passage inspired many voyages by others, drowned on his own attempt in 1583. Henry Hudson, his young son, and seven others were cast adrift by a mutinous crew in 1611, when his discovery of Hudson Bay proved to be an icy trap instead of the passage he sought. Knowledge of an Arctic passage came slowly, over hundreds of years, from information gathered during voyages by such explorers as John Davis, William Baffin, Sir John Ross, Sir William Parry, Frederick William Beechey, and Sir George Back, augmented by overland expeditions by Henry Kelsey, Samuel Hearne, and Sir Alexander Mackenzie. The worst tragedy came when Sir John Franklin and 129 men aboard HMS "Erebus" and HMS "Terror" vanished in 1845. One searcher for the lost expedition, Robert McClure, entered the passage from the west, became locked in the ice for two winters, then sledged overland to another rescue ship coming from the east to complete the first Northwest Passage in 1854. Adolf Erik Nordenskiöld led a Swedish–Russian voyage



The Northwest Passage.

First conquests of the passage

through the Northeast Passage in 1878–79, and Soviet polar icebreakers have opened this route to limited use in modern times. But the Northwest Passage was not finally conquered by sea until 1906, when the Norwegian explorer Roald Amundsen, who had sailed secretly to escape creditors seeking to stop the expedition, completed an arduous three-year voyage in the converted 47-ton herring boat "Gjöa." The first single-season transit came only in 1944, when Sgt. Henry A. Larsen, of the Royal Canadian Mounted Police, made it through in the schooner "St. Roch."

THE ENVIRONMENT

The hostile Arctic makes the Northwest Passage one of the world's severest maritime challenges. It is 500 miles (800 kilometres) north of the Arctic Circle and less than 1,200 miles (1,900 kilometres) from the North Pole. It consists of a series of deep channels through Canada's Arctic Islands, extending about 900 miles (1,450 kilometres) east to west, from north of Baffin Island to the Beaufort Sea, above Alaska. Thick pack ice, moving at speeds up to ten miles a day, completely covers its waters about nine months of the year and chokes nearly half the passage all the year round. Arctic water can freeze a man to death in two minutes. Frigid polar northeasterly winds blow almost constantly and can howl to hurricane force in winter. Temperatures plunge to —50° F (—46° C) and rise above freezing only in July and August. Exposed human flesh is subject to frostbite during most of the year, and breathing in open winter air is difficult and can be dangerous. Visibility is often obscured by whiteouts of blowing snow, the area is plunged into four months of 24-hour, midwinter darkness, and thick fog usually shrouds open water during the brief summer. The main channels are more than 1,000 feet deep, but there are uncharted shoals, and little is known about Arctic currents and tides. Navigation is difficult even with the most modern devices: the magnetic compass is useless because the magnetic north pole lies within the passage; the gyro compass can be unreliable at high latitudes; underwater ice distorts sonar depth soundings; and the bleak, featureless Arctic islands provide few distinguishable landmarks for visual or radar checks. Even navigation by space satellite suffers from electronic interference, and Arctic blackouts can block all communications for periods from a few hours to nearly a month. To reach the Northwest Passage from the Atlantic requires a hazardous voyage through a stream of about 50,000 giant icebergs, up to 300 feet in height, constantly drifting south between Greenland and Baffin Island. The bergs are often hidden in dense fog created by collision of the warm Gulf Stream with icy Arctic waters, and small "growlers" up to several thousand tons can be invisible to radar and lookouts. The exit to the Pacific is equally formidable, as the polar ice cap presses down on Alaska's shallow north coast much of the year and funnels masses of ice into the Bering Strait, between Alaska and Siberia.

Navigational hazards

The greatest challenge of the Northwest Passage to shipbuilding technology is the ice pack. It grows only about seven feet thick in an average season, but much of the ice survives from year to year, increasing to 14 feet or more in thickness and becoming much harder as it loses salinity. The polar cap also constantly feeds hard old ice into the passage from the west, through McClure Strait. The twisting, grinding action of the wind and currents on the pack creates pressure ridges, huge cone-shaped masses of ice believed to extend as much as 150 feet below the waterline, and for miles laterally. Early explorers skirted the pack close to shore in shallow draft vessels, and modern supply convoys stick mainly to open summer waters. Commercial use of the passage would require vessels with the power to batter through the ice pack year round.

The ice pack

THE "MANHATTAN" VOYAGE

To evaluate the economic feasibility of such ships, and to gather the data needed to build them, a $50,000,000 project was launched in 1969 by the three international or North American oil companies with the largest holdings of Alaskan Arctic oil. They leased the largest and most powerful commercial ship ever built in the U.S., the "Manhattan," and sliced it into four pieces so that 10,000 workers, from Maine to Alabama, could simultaneously convert it to the world's largest icebreaker and an Arctic research vessel. She was fitted with a massive 725-ton, shark-nosed icebreaking bow, girdled with a 16-foot, 2,200-ton steel ice belt, and her rudders and 22-foot-diameter propellers encased in ice shields. She was equipped with computer-controlled navigational equipment, the latest electronic devices, and charts made by U.S. Navy nuclear submarines. A landing pad for two helicopters was added to her stern. The reassembled giant was nearly as long as the Empire State Building tipped on its side and, at 151,000 tons, displaced about twice the weight of the "Queen Elizabeth." Sailing from Chester, Pennsylvania, August 24, 1969, she plowed into the Baffin Island ice pack September 2, entered the Northwest Passage three days later, and completed a historic transit on September 14, after smashing more than 650 miles of ice. She sailed as far west as Point Barrow, Alaska's northwestern tip, then returned to the passage for extensive ice tests, completing the voyage in New York on November 12. The "Manhattan" broke ice up to 14 feet thick for extended periods and smashed ridges as deep as 40 feet. But she often got stuck in hard polar ice, proved underpowered and underprotected, and was seriously hampered by possessing only one-third as much power to go astern

as to go forward. The most harrowing moments came the night of September 10–11. Attempting to become the first vessel to cross McClure Strait east-to-west, the "Manhattan" was halfway through when she became locked in polar ice grinding north of Banks Island. She escaped only with assistance from the Canadian icebreaker "John A. MacDonald," her constant companion and frequent rescuer on the voyage, and by diverting steam from living spaces to squeeze an extra 7,000 horsepower from her 43,000-horsepower turbines. She completed the Northwest Passage through the narrow Prince of Wales Strait to the south, the likely route for commercial vessels. But the chief mission of the "Manhattan" on the Northwest Passage, and on a second voyage to Baffin Island in the spring of 1970, was to gather mountains of data from which vessels more than twice her size could be designed. More than 400 gauges measured strains on her hull and acceleration through the ice; closed-circuit television recorded action of the ice on her bow and stern; and scientific parties were sent onto the ice pack to drill and analyze the ice. The data were fed into computers for months after her return. Ice smashed a 15- by 30-foot hole in a bow tank unprotected by her ice belt during the first voyage. This incident increased concern for protecting the fragile Arctic ecology by governments and conservation groups mindful of the 1967 disaster when the grounded supertanker "Torrey Canyon" spilled oil over wide coastal areas of Cornwall. Bacterial action that fairly rapidly consumes oil in warmer waters is far slower in the Arctic, and cleanup techniques effective in open waters might not work on oil trapped under the ice pack. Arctic waters teem with seals, walrus, and whales that could be endangered by an oil spill. Canada has claimed jurisdiction over Arctic waters up to 100 miles from its shores for purposes of setting tough pollution control regulations and required a $6,500,000 bond against pollution damage before approving the "Manhattan's" second voyage. It appears certain that vessels using the Northwest Passage will be subject to stringent antipollution measures.

The pollution threat

ECONOMIC SIGNIFICANCE AND POTENTIAL

Opening the Northwest Passage to regular commercial ocean traffic would have worldwide economic significance in natural resources, transportation, and trade relations among nations. The greatest impact would be on the U.S. and Canada, but effects could be felt from the Persian Gulf to Panama, and from Chile to Scandinavia. But competitive developments, governmental policies, and many complex economic issues, are likely to determine how soon, and how much, such a route would be used. The most likely key to opening the passage is oil. Alaska's Arctic North Slope is estimated to contain as much oil as proven 1970 reserves for all the rest of the United States, where steadily rising demand began to exceed productive capacity by 1972. Many believe the lowest cost method of moving Alaskan oil to the major markets in the eastern U.S. will be through the Northwest Passage. Potential advantage of this route is estimated at 35 to 55 cents a barrel under the use of pipelines. This would mean annual transportation savings of $190,000,000 to $300,000,000 on expected 1980 shipments of 1,500,000 barrels a day. The Alaskan oil could significantly reduce the need for East Coast refineries to begin importing major amounts of oil from the Middle East and North Africa to supplement traditional supplies from the Gulf Coast and from Venezuela. The Alaskan oil could cost less than that from Texas and Louisiana. To move it would require a fleet of about 30 mammoth icebreaking tankers, each capable of carrying 250,000 tons or about 1,800,000 barrels of oil, more than twice the "Manhattan's" 115,000 tons. They would cost a minimum of $1,500,000,000 ($50,000,000 each), and the price tag could escalate as high as $3,000,000,000. Building them would require the greatest U.S. shipbuilding program since World War II. The new vessels would modernize, and expand by 250 percent, capacity of a U.S. flag tanker fleet made up largely of small and nearly obsolete ships. Unloading them would mean building, off the shores of the U.S., those terminals, built to handle mammoth vessels, that have been sharply reducing oil transportation costs for many other nations since the late 1960s. Loading them could require an investment of $150,000,000 to $500,000,000, including such options as building a man-made harbour on Alaska's shallow north coast, perhaps using nuclear explosives; a pipeline to a natural harbour on Canada's Herschel Island, off the Yukon Territory; or an underwater pipeline and ice-resistant mooring facility in the Beaufort Sea. Polar icebreakers to assist the tankers would cost about $50,000,000 each, and navigational aids and additional shore installations would be needed in the Arctic.

Oil-transport costs

Such facilities and regular use of the passage would accelerate exploration for minerals in Canada's vast Arctic, and sea transportation could bring commercial development of known major deposits. These include an estimated 1,000,000,000 tons of high-grade iron ore at Mary River on northern Baffin Island; large copper reserves in the Coppermine area of the mainland near the western end of the Northwest Passage; and major amounts of lead, zinc, and sulfur in the Arctic Islands. Nearness to markets and relative security of Arctic ores could make them stiff competitors for those from Africa and South America. Geologists have predicted that drilling will locate oil fields in the Canadian Arctic to rival those in Alaska. Exploratory drilling was spurred in 1970 by the discovery of oil in the Arctic Mackenzie River Delta, and of natural gas in the Arctic islands. The high value of crude oil in the U.S., due to controls on lower cost foreign imports, makes it unlikely much Alaskan oil will move into world markets, except in supply emergencies. But Canadian Arctic oil could become a competitor for that from the Middle East. Through the Northwest Passage, Arctic oil would be only about 4,000 miles from Japan and Europe, compared with the 8,000- and 11,000-mile distances, respectively, that Middle East oil travels to those markets. Canada's Arctic oil would also move to its own east coast refineries, which at the start of the 1970s were supplied almost entirely from Venezuela. Opening the Northwest Passage could bring new settlements to Canada's vast empty northlands, which in 1970 had only about 22,000 people spread over 1,300,000 square miles. Arctic natural gas could make it feasible to process minerals in the Arctic and might even make domed, climate-controlled Arctic cities a reality. Wheat could move from Canada's prairies to world markets through Hudson Bay and the Northwest Passage on a year-round basis, providing competition for the St. Lawrence Seaway. Easier access to the Arctic would aid U.S. and Canadian military defense of North America. The cost of ice-strengthening ships, and the probable high insurance rates for vessels used in Arctic service, however, could inhibit use of the Northwest Passage as a trade route. But it would cut the distance between London and Tokyo, for example, to less than 8,000 miles, from the 14,670 mile route around Africa made necessary by the shutdown of the Suez Canal in 1967. The Northwest Passage would permit use of far larger vessels than allowed by dimensions of the Panama and Suez canals. Icebreaking techniques learned in the Northwest Passage could be applied in other ice-locked waters from the Great Lakes to the Baltic, including the Soviet Union's Northeast Passage with its vast Siberian oil fields. Canada has held sovereignty over the Arctic Islands since 1880, but some authorities contend that much of the Northwest Passage is in international waters. Canada has indicated it would welcome, and permit, international commerce over the route, subject to pollution control regulations.

Mineral deposits

Regional implications

THE OUTLOOK

Modern technology can overcome the hostile Arctic environment of the Northwest Passage, but economic considerations may prove a greater obstacle. Crews of Arctic vessels can be protected by completely enclosed upper decks, Arctic clothing, and life-support capsules in place of lifeboats. Polar icebreakers would be stationed along the route and might convoy groups of vessels through the passage. The "Manhattan" was guided by infrared photographs, laser-beam measurements, and "side-looking" ra-

dar maps made of the ice pack by U.S. and Canadian ice reconnaissance planes. Space satellites fixed the location of the "Manhattan" every 22 minutes to an accuracy of about one-third ship length. The greatest barrier to building icebreaking tankers to haul Alaskan oil may prove to be the Jones Act, which requires vessels used in domestic trade to be built in U.S. shipyards and operated by U.S. crews. This adds about 50 percent to the cost of building and operating such vessels. Any major change in U.S. restrictions on foreign oil imports could make Alaskan oil uncompetitive in east coast markets. Pollution controversies could delay construction of offshore tanker terminals, or use of the Arctic by giant tankers. In late 1970 the chief financial backer of the "Manhattan" voyages stated that evaluation of data collected indicated that the Northwest Passage is a commercially feasible route. But the company announced it was suspending studies of icebreaking tankers to concentrate on pipeline alternatives which appeared to have an economic edge. Proposed construction of the world's largest pipeline from Prudhoe Bay across Canada to the U.S. Midwest could result in a decision to supply the east coast through a connecting pipeline link. Other proposed pipeline competitors for the Northwest Passage include a transcontinental line from Puget Sound, to be filled by conventional tankers loading at the ice-free terminus of a pipeline across Alaska, and a line across Central America, linking Pacific and Atlantic supertankers. The Northwest Passage may be opened by vessels going under, rather than through the ice. A group of five oil companies were studying in the '70s whether to undertake a $3,000,000,000 program to build nuclear-powered submarine tankers, each capable of carrying 255,000 tons of oil, and the underwater terminal to load them. The submarines could travel three times faster under the ice than an icebreaking tanker smashing across the surface, and they would be protected from the harsh Arctic environment. Regular voyages under the polar cap by U.S. Navy submarines, beginning with USS "Nautilus" in 1958, indicate that the idea is operationally feasible. The need for the Arctic's natural resources and the obvious geographical advantages of the Northwest Passage as a trade route are likely to make it an eventual commercial reality.

BIBLIOGRAPHY. W.D. SMITH, *Northwest Passage: The Historic Voyage of the S.S. Manhattan* (1970); and B. KEATING, *Northwest Passage* (1970), are well-written accounts by two men who were on the "Manhattan." Two books by H.A. LARSEN on his 1940s voyages are *The Big Ship: An Autobiography* (1967), and *The North-West Passage 1940–42 and 1944* (1958). Other accounts include: J.A. MIERTSCHING, *Frozen Ships: [his] Arctic Diary, 1830–1854* (1967); F. MOWAT (ed.), *Ordeal by Ice* (1960); L.H. NEATBY, *In Quest of the North West Passage* (1958); N.M. CROUSE, *The Search for the Northwest Passage* (1934); R. HAKLUYT, *Voyages in Search of the North-West Passage* (1886); T. RUNDALL, *Narratives of Voyages Towards the North-West* (1849); V. STEFANSSON, *Northwest to Fortune: The Search of Western Man for a Commercially Practical Route to the Far East* (1958); G. WILLIAMS, *The British Search for a Northwest Passage in the Eighteenth Century* (1962); E.S. DODGE, *Northwest by Sea* (1961); L.S. MCDONALD, *Search for the Northwest Passage* (1958); and S.E. MORRISON, *The European Discovery of America: The Northern Voyages* (1971).

(R.W.B.)

Northwest Territories

The Northwest Territories comprise over a third of Canada, itself the world's second largest nation in area, and they reach almost from the eastern to the western extremities of the country across the roof of the North American continent. The territories are among the most sparsely populated but habitable regions of the world, their 1,304,903 square miles (3,379,683 square kilometres) containing nearly 35,000 inhabitants, over half of whom are native Eskimos and Indians. In the north, the territory extends far above the Arctic Circle to incorporate thousands of islands, the largest of which are Victoria Island in the west and Baffin Island in the east, as well as the islands within bays and straits of Hudson Bay and Ungava Bay. The land is one of high plateaus separating a virtually Arctic lowland on the east from a western depression of more moderate climate. Thousands of lakes dot its heavily glaciated surface.

There are three administrative districts in the territories. The District of Franklin takes in the Arctic islands of the north and Boothia and Melville peninsulas. The District of Keewatin lies north of the province of Manitoba and includes the islands of Hudson Bay. The most developed region, the District of Mackenzie, borders the Yukon Territory to the west and the provinces of Saskatchewan, Alberta, and British Columbia to the south. The mineral wealth beneath the land of the territories is thought to be immense, but the climate and distance from major population centres have kept the territories in much of a frontier condition. (For information on related subjects, see the articles CANADA, HISTORY OF; and NORTH AMERICA.)

HISTORY

Vikings probably visited parts of the Canadian Arctic during the Middle Ages, but there are no records of exploration until the voyage in 1576 of the English mariner Martin Frobisher in search of a northwest passage to the Orient. Other expeditions in the 17th century also failed to find the route, but they added to knowledge of the Arctic regions. Interest in finding the route waned in the 18th century, but whaling ships became commonplace in the Arctic waters. The first recorded exploration of the mainland was by Samuel Hearne, who in 1770–72 journeyed from the west coast of Hudson Bay to the mouth of the Coppermine River on the northern coast. Other inland explorations were mainly the work of Montreal-based fur traders. In 1789 Alexander Mackenzie of the North West Company travelled down the river that bears his name to reach the Arctic Ocean. In the 19th century there was renewed interest in a northwest passage. Sir John Franklin and others explored much of Mackenzie District and mapped parts of the northern coastline during the 1820s, work that Thomas Simpson continued in 1838–39. Searchers for the lost Franklin expedition of 1845–48 explored and mapped other parts of the eastern Arctic in the following decade. Later a series of expeditions attempted to reach the North Pole; such exploits continued into the 20th century but by then were overshadowed by more practical activities directed at identifying the resource potential of the Canadian north.

Early explorers and map makers

Settlements were first established to serve the whaling fleets and fur traders. Missionaries became active in the Mackenzie valley in 1852 and in the eastern Arctic toward the end of the century. No resident administrative authorities were established within the present limits of the Northwest Territories until the 20th century. Responsibility for the mainland territories that drain into Hudson Bay, known as Rupert's Land, was vested in the Hudson's Bay Company. The remaining part of the mainland, the North-Western Territory, was under nominal British rule until 1870, at which time both it and Rupert's Land were ceded to Canada. In 1880 the Arctic islands claimed by Britain were also placed under Canadian jurisdiction. Separation of the Yukon Territory, creation of new provinces, and enlargement of other provinces reduced the Northwest Territories to its present limits by 1912. The Royal Canadian Mounted Police were made responsible for maintaining law and order and for providing whatever governmental administration was required in the area.

Fur traders, missionaries, and the police directed the life of the Northwest Territories until the 1920s, when discovery of oil near Fort Norman on the Mackenzie River prompted the Canadian government to establish a territorial administration for the area. Mining replaced the fur trade as the most important industry in Mackenzie District in the 1930s. World War II brought much government-financed construction activity to the territories. In the southern Mackenzie area the Canol pipeline, linking the oil field at Norman Wells to a refinery at Whitehorse in the Yukon, and construction of several large airfields in the eastern Arctic did much to open the Canadian north to further exploration and development. After the war, construction of the Mackenzie Highway to Great Slave Lake and the building of the Distant Early

Warning radar network, the DEW line, continued this process. A great expansion of government-sponsored health, education, and welfare services transformed living and social conditions throughout the north.

THE PHYSICAL ENVIRONMENT AND NATURAL LIFE

Two main regions

Two main types of landscape blend into one another along the tree line, which runs southeast from near the Mackenzie Delta on the Arctic Ocean to near Churchill, Manitoba, on Hudson Bay. Southwest of this line lies the northernmost part of the Canadian boreal forest, extending westward to the mountain ranges that separate the Northwest Territories from Yukon Territory. North and east of the tree line stretch the relatively barren grounds of the Arctic, vast reaches of flat, often poorly drained lowlands and of Precambrian rock more than 570,000,000 years old. Within each of these two main regions, the surface vegetation and the animal life it supports vary with soil and climatic conditions.

The Mackenzie Lowlands. The most favourable conditions are found in the Mackenzie Lowlands in the west, where forests of black and white spruce mixed with deciduous species extend north to the Mackenzie Delta. The climate here is relatively mild, with warm and dry summers during which mean July temperatures of 60° F (about 16° C) are recorded at most of the settlements along the Mackenzie River. The winters are long and cold, with an average mean January reading at Yellowknife on the northern shore of Great Slave Lake of −18° F (−28° C). With only about 70 frost-free days, the growing season is short, which limits plant growth. While it lasts, however, wildflowers and grasses flourish and root and cereal crops can be cultivated. Many species of valuable furbearing animals are found, notably muskrat and beaver. Moose, wolves, black and grizzly bears, and mountain sheep and goats are also native to the area. Pickerel, northern pike, and whitefish are among the fish found in the rivers and lakes.

The coolness of the climate limits navigation on the Mackenzie River system to about four months a year and gives rise to a permanently frozen subsoil, or permafrost, except in a small area south of Great Slave Lake. Permafrost creates serious construction problems, especially where the frozen subsoil is an unstable mixture of fine silt and water.

The northeast. East of the Mackenzie Lowlands and the tree line the climate is colder, and the terrain changes to that of the ancient and rocky Precambrian mass known as the Canadian Shield, the edge of which is straddled by the two largest lakes, Great Bear Lake, 12,275 square miles, and Great Slave Lake, 10,980 square miles. Tree growth becomes sparse and stunted and eventually disappears, to be replaced by the light but tough vegetation of the Arctic tundra. East of the tree line the Arctic climate prevails, with average winter temperatures similar to those in the sub-Arctic region to the west, but with summer temperatures not rising above an average of 50° F (10° C) even in July. In these so-called barren lands, annual precipitation is light and the soils, where they exist at all on the heavily glaciated surface, are usually sandy and thin. Mosses, lichens, and many small hardy flowering plants survive in these conditions and support a variety of animal life ranging from small burrowing mammals and their enemy, the white Arctic fox, to the large caribou and musk-ox. Seals, walrus, and polar bears are prevalent along the coasts. Bird life is plentiful in summer, with some species, notably the ptarmigans and ravens, remaining all winter. Mosquitoes and other insects abound during the summers.

Human impact and settlement. The native people of the territories once led nomadic lives, surviving by adapting to the harsh natural environment. But this balance was disturbed when Europeans established permanent settlements and introduced firearms. The numbers of barren-ground caribou were drastically reduced, and the musk-ox was almost exterminated before being protected in 1927. Fuel and game resources near settlements were rapidly depleted once people began to congregate around the trading posts and missions. Climatic and soil conditions made regeneration slow. Attempts to introduce reindeer and other domesticated animals have not been successful, in part because of the difficulty of managing animals on ranges large enough to prevent destruction of the vegetation by overgrazing.

Today, nearly all the population lives in small settlements along the Mackenzie River and along the Arctic coastlines of the mainland and northern islands. The largest town is the capital, Yellowknife, with a population of about 6,000 in 1971. Hay River, on Great Slave Lake; Fort Smith, near the Alberta border, and Inuvik, in the Mackenzie delta are other towns of over 2,500 people in the Mackenzie area. In the eastern Arctic, Frobisher Bay, on Baffin Island between Hudson Bay and the Atlantic, is the only town with more than 1,000 inhabitants.

THE PEOPLE AND THE CONDITIONS OF LIFE

Population composition. The native Indians and Eskimos comprise two distinct groups, differing in appearance, language, and culture and living apart from one another.

Indians. The approximately 6,000 Indians belong to several tribes, all part of the Athabaskan language family. Tribal organization was never strong among the northern Indians, and small bands led by individuals chosen for their skill in the hunt were the effective social unit. This arrangement was easily molded to the needs of the fur trade when it reached the Mackenzie area in the 18th century. Thereafter, the exchange of furs for imported goods became the basis of the Indian economy. Government treaties were made with the groups living south of Great Slave Lake in 1899, with those living farther north, only in 1921. No reservations were established. The decline of the fur trade in the 20th century left much of the Indian population unemployed, and they have had difficulty adapting to wage employment.

The Indian fur-trading economy

Eskimos. The origins of the approximately 11,000 Eskimos living in the territories, largely in the coastal areas, are obscure. Although several dialect groups are represented, all apparently have descended from what is known as the Thule culture, a prehistoric maritime culture living north and west. A distinct group living inland in the District of Keewatin perished when the caribou migrations upon which they depended were disrupted following the advent of the rifle. The coastal Eskimos lived in small, nomadic family groups before the coming of Europeans. Early contacts with explorers and whaling crews introduced new diseases and reduced the population during the 19th century. There was also considerable interbreeding. The fur trade was not well established in the Arctic until early in the 20th century; but the Eskimo way of life adapted quickly to it, and they, like the Indians, came to depend upon outside sources of supply for most of the necessities of life. Construction activity during World War II and in the postwar years further affected their way of life. Eskimos adapted readily to the opportunities for casual employment, and many were quick to abandon the seminomadic trapping and hunting existence for life in the settlements.

Europeans. The other approximately 15,000 people in the territories are mainly of European and mixed European and native descent. Most live in the more economically advanced Mackenzie District, where they find employment in mining, transportation, and public service. Much of this population has always been transient, living only temporarily in the north before returning to the more settled parts of North America.

Demographic trends. Until World War II the non-native population was small, numbering only 2,572 in the 1941 census. The annual rate of natural increase among the more numerous native population was only about four to five per thousand of population until the 1940s, when the birthrate rose sharply, doubling from 1940 to 1950. Subsequent reductions in the death rate due to improved living conditions raised the annual rate of increase even further. Migration into the territories increased during and after World War II and, with the natural increase, raised the total population from just over 12,000 in 1941 to 36,000 in 1972.

The economy. High production costs and transportation problems inhibit development of many of the territories' known mineral resources, including the large iron ore deposits in the eastern Arctic; but prospecting remains active even in the remote Arctic islands, especially for oil.

Mining and power. Mining is the principal industry of the territories; much of this production is accounted for by the large open-pit lead and zinc mines at Pine Point on the south shore of Great Slave Lake. Gold and some silver have been mined at Yellowknife since the late 1930s. Production of radioactive ores at Port Radium on Great Bear Lake ceased in 1960, but silver is now mined in the area. Tungsten, copper, and cadmium are also produced. Petroleum fuels for use in the territories are obtained from a small refinery located at the Norman Wells oil field. Hydroelectric power is supplied to the Pine Point area by an 18,000-kilowatt-capacity generating plant on the Taltson River, and to the Yellowknife area by plants on the Snare and Yellowknife rivers.

Exploitation of minerals

Hunting and fishing. Trapping continues to provide income for some of the native population. Muskrat, beaver, marten, mink, and lynx are the most important furs taken in the Mackenzie area, while the white fox remains the principal fur in the Arctic regions. Fishing and hunting of sea mammals also provides some employment. Whitefish and lake trout are fished commercially on Great Slave and some smaller lakes. Arctic char are fished and exported to southern markets from several communities in the eastern Arctic. Seals and small whales are hunted for food, and some sealskins are marketed commercially. Sport fishing and hunting are major attractions for the small but growing number of tourists who visit each year.

Forestry and agriculture. The western Mackenzie district has most of the 200,000 square miles of forested land, but even there, large stands of merchantable timber are not plentiful. The several sawmills process the timber only for local use. Although there are more than 1,000,000 acres of arable land in the Mackenzie district, farming is not profitable. Some field crops are grown for local use, but most foodstuffs must be imported from the outside, very much affecting their price.

Economic management. Governmental assistance in the development of major resources is provided only in the form of roads, electric-power facilities, mapping, and geological services, but the government participates in a consortium with private companies to search for oil and natural gas in the Arctic region. The government-owned Northern Canada Power Commission operates generating facilities throughout the territories. Another government enterprise, Northern Transportation Company Limited, is the principal carrier on the Mackenzie River waterway. The government also has encouraged establishment of local cooperative enterprises in such smaller businesses as commercial fishing, arts and crafts, fur-garment manufacture, retailing, and logging. By 1970, 34 such cooperatives had been established.

Transportation. Nearly all passenger and much freight traffic is carried by scheduled and chartered air services. Flights link Yellowknife and other major settlements along the Mackenzie Valley to Edmonton, Alberta. Settlements on the west coast of Hudson Bay and in Keewatin District are connected to Winnipeg, and those in the eastern Arctic to Montreal.

Airline services

Surface transportation for heavy freight is mainly by water. Fuel oil for heating, and other bulky supplies, are carried to eastern Arctic settlements by seagoing supply ships organized by the federal Department of Transport. In Mackenzie District, the Mackenzie waterway system, 1,700 miles (2,735 kilometres) long, is the main freighting facility. It is supplemented by the Mackenzie Highway, part of a highway system of about 800 miles in the southern part of the district, and by the Great Slave Lake Railway, connecting Pine Point to the trans-Canadian rail systems at Grimshaw, Alberta. Tractor trains and other overland vehicles using temporary winter roadways carry heavy freight into remote areas. Snowmobiles have largely replaced the dogsled for light winter travel.

Administration and social conditions. Ultimate constitutional responsibility for government in the territories rests with the federal government in Ottawa, but most responsibilities of a provincial nature in Canada have been delegated to a territorial administration sitting at Yellowknife. This consists of a commissioner and a council comprising seven members elected by the people and five members appointed by Ottawa. The territories are represented by one elected member of Parliament. Justice is dispensed by a territorial court, a police magistrate, and several justices of the peace. Law enforcement is by the Royal Canadian Mounted Police.

The federal government retains direct responsibility for all natural resources except game and administers them through its Department of Indian Affairs and Northern Development. The possibility of achieving provincial status for all or some part of the territories in the near future has become a lively political issue in recent years.

Education and health services. Missionaries provided nearly all the education and health care available in the territories until the 1950s, but education is now a government responsibility. A fully integrated system of elementary, secondary, and vocational education is administered by the Education Branch of the Department of Indian Affairs and Northern Development. The provision of health and welfare services has also become an increasingly heavy responsibility of government in recent years. Health services are administered by the federal Northern Health Service. A system of public health and treatment facilities is based on the five major hospitals. There is a universal hospital-insurance scheme, and free treatment is provided for several diseases, including tuberculosis, which has been a major threat to the health of the native people.

Income and housing. Salaries of professional and skilled workers are unusually high to compensate for the high cost of northern living, but because of the extremely low cash incomes earned by persons lacking such skills, the overall level of personal income is lower than in other parts of Canada. Inadequate housing has been a serious problem in most settlements. High building costs and low incomes allow few native people to afford modern housing. Government subsidized housing programs are beginning to alleviate this problem.

Cultural life. Modern forms of transportation and communication have done much to break down the isolation of life in the north, and contemporary North American popular culture is everywhere evident. Radio stations relay programs throughout the territories, and most of the larger settlements have their own newspapers and some have closed-circuit television. In the Mackenzie District, little remains of the traditional culture of the native Indians; but in the eastern regions, elements of the Eskimo culture have been preserved. Public policy in recent years has encouraged the bringing back of Eskimo traditions in arts and crafts, providing an important source of income in some Eskimo communities and making Eskimo culture familiar to collectors throughout the world. Eskimo-language broadcasting is well established, and a new written language has been introduced to encourage Eskimo writing and to facilitate communication among the widely scattered communities.

Preservation of Eskimo culture

Prospects. The immediate problems of the Northwest Territories mainly have to do with improving the social and economic conditions of the native people and other permanent residents. Their interests must be reconciled with those of the business and other outside groups responsible for organizing and financing the major resource development programs. The future of the territories appears to hinge upon constructive approaches to these problems, though any major growth in population seems unlikely in the foreseeable future.

BIBLIOGRAPHY. R.A.J. PHILLIPS, *Canada's North* (1967), the best general account of the development and history of the area; K.J. REA, *The Political Economy of the Canadian North* (1968), an economic history of the Northwest Territories and the Yukon placing emphasis on the economic and political factors in its development as a region; contains much statistical data. Useful material may also be found in

several collections and symposia: V.W. BLADEN (ed.), *Canadian Population and Northern Colonization* (1962); and F.H. UNDERHILL (ed.); *The Canadian Northwest: Its Potentialities* (1959); both contain excellent articles on various aspects of the resources and social problems of the territories. Current statistics on population are available in the 1971 census. The *Canada Year Book* provides most of what little statistical material we have on the Northwest Territories, although most of it is combined with data for the Yukon Territory. Other useful government publications are the annual reports of the Commissioner of the Northwest Territories and the federal Department of Indian Affairs and Northern Development. Some useful background on the most important political issue in the territories, the prospect of provincial status, may be found in the *Report of the Advisory Commission on the Development of Government in the Northwest Territories*, 2 vol. (1966).

(K.J.R.)

Norway

Norway, the "northern way," occupies the western half of the Scandinavian peninsula of Northern Europe. In shape, it has been characterized as resembling a fish with its tail flapping in the Arctic Ocean. Its bulging head, with a wide-open mouth about Oslofjorden, around which region nearly half the 4,000,000 inhabitants of the country live, faces to the south. Norway has always depended heavily on its economic relations with foreign countries; this has been the case during both its periods of independence and those times when it has been politically united with its fellow Scandinavian nations, Sweden and Denmark. These foreign links are illustrated by the heritage of the Vikings, who plundered the coasts from the British Isles into the Mediterranean Sea and across the Atlantic Ocean to North America. Later, in less colourful but more peaceful fashion, the Norwegians have turned to trading in fish and lumber, and the modern nation has emerged as a major maritime transporter of the world's goods. Nearly two-thirds of the physical area of Norway is classed as mountainous territory. Its more than 160,000 lakes and about the same number of islands testify to the glaciers that scourged its landscape and coastlines.

Maritime heritage

Lying on the northern outskirts of Europe and thus avoiding the characteristics of a geographical crossroad, Norway has maintained a great homogeneity among its peoples and their way of life. With Sweden, it has a larger percentage of tall, blond, and blue-eyed persons than any other nation of the world, and its projections for life expectancy, for both men and women, are among the highest in the world. Although it is split politically between Socialists and non-Socialists, the former long ago stopped insisting on nationalization of the nation's industry, and the latter have accepted extensive governmental control of the economy. Such evidences of national consensus, along with abundant waterpower and peaceful labour relations, were a major factor in the rapid growth of Norway as an industrial nation in the 20th century, creating one of the highest standards of living in the world. (For information on related topics, see the articles OSLO; EUROPE; and SCANDINAVIA, HISTORY OF.)

THE NATURAL ENVIRONMENT

Geology. Geologically, Norway belongs to northern Europe's Fennoscandian Shield, its extremely hard bedrock mostly of granite and other heat- and pressure-formed materials, ranging from 1,000,000,000 to 2,000,000,000 years in age.

Terrain. Glaciation and other forces wore down the surface and created thick sandstone, conglomerate, and limestone deposits known as sparagmite, as well as numerous extensive areas, called peneplains, whose relief has been largely eroded away. Remains of the latter include the Hardangervidda, 3,000 feet (1,000 metres) above sea level, in southern Norway, Europe's largest mountain plateau, covering about 4,600 square miles (12,000 square kilometres); and Finnmarksvidda, 1,000 feet (300 metres) above sea level, occupying most of the northernmost and largest county of Norway.

From the Cambrian through the Silurian geological periods, from about 570,000,000 to 395,000,000 years ago, most of Norway was below sea level and acquired a layer of limestone, shale, slate, and conglomerate from 100 to 160 metres thick. Folding processes in the earth then gave rise to a mountain system that is a continuation of the Caledonian System of the British Isles. Erosion was increased by the large foldings of the Tertiary Period, beginning 65,000,000 years ago, which in addition to elevating the Norwegian mountains, particularly along the west coast, also created the Himalayas in Asia, the Alps in southern Europe, and the Andes and the Rockies in the Americas.

Rivers running westward acquired tremendous erosive power. Following fracture lines marking weaknesses in the earth's crust, they dug out gorges and canyons that knifed deep into the jagged coast. To the east, the land sloped more gently, and broader valleys were formed. During the Ice Age of the Quaternary Period, less than 2,500,000 years ago, the scouring action of glaciers tonguing down the then-existing V-shaped valleys created the magnificent U-shaped drowned fjords that now grace the western coast. Enormous masses of earth, gravel, and stone were also carried by glacial action as far south as Denmark and northern Germany. Bedrock, exposed in almost a third of Norway, was scoured and polished by the movements of these materials.

Soils. In the melting periods following each ice age, however, large areas were flooded by the sea because the enormous weight of the ice had depressed the land. Thick layers of clay, silt, and sand were deposited along the present coast and in large areas in the Oslo and Trondheim regions, which rise as high as 650 feet (200 metres) above sea level today. Some very rich soils are found below these old marine coasts. In the large areas covered by forests, the main soil has been stripped of much of its mineral content, creating poor agricultural land. Norway has an average altitude of 1,600 feet (500 metres), compared to 1,000 feet (300 metres) for Europe as a whole.

Climate. Although it occupies almost the same degrees of northern latitudes as Alaska, Norway owes its warmer climate to the Gulf Stream, carrying 4,000,000 to 5,000,000 tons of tropical water per second into the surrounding seas. The Gulf Stream usually keeps the fjords from freezing, even in the Arctic Finnmark region. Even more important are the southerly air currents brought in above these warm waters, especially during the winter.

Effect of the Gulf Stream

The annual mean temperature on the west coast is 45° F (7° C), or 54° F (12° C) above average for the latitude. In Lofoten, above the Arctic Circle, the January mean is 43° F (24° C) above the average for this latitude around the world, one of the greatest thermal anomalies known. Norway lies directly in the path of the North Atlantic cyclones, which bring frequent changes in weather and gales. Western Norway has a marine climate, with comparatively cool summers, mild winters, and up to 80 inches of mean annual precipitation. Eastern Norway, sheltered by the mountains in the centre of the country, has an inland climate with warm summers, cold winters, and less than 40 inches mean annual precipitation.

Plant and animal life. Norway has about 2,000 species of plants, but only a few, mainly mountain plants, are peculiar to Norway. Thick forests of spruce and pine predominate in the broad glacial valleys up to 2,800 feet (850 metres) above sea level in eastern Norway and 2,300 feet (700 metres) in the Trondheim region. Even in the thickest spruce woods, the ground is carpeted with leafy mosses and heather, and a rich variety of deciduous trees—notably birch, ash, rowan, and aspen—grow even on the steepest hillsides. The birch zone extends from 3,000–3,900 feet (1,000–1,200 metres) above sea level, after which there is a willow belt that includes dwarf birch.

In western Norway there are virtually no conifers. North of the Arctic Circle there is no spruce, and only some pine grows in the inland valleys, amid their surprisingly rich vegetation. Wild berries grow abundantly in all regions: they include blueberries and cranberries of small size and yellow cloudberries, a fruit-bearing plant of the rose family that is little known outside Scandinavia and Great Britain.

Mammals

Reindeer, wolverine, lemming, and other Arctic animals are found throughout Norway but, in the South, only in the mountain areas. Elk are common in the large conifer forests and red deer on the west coast. Only 100 years ago large animals of prey were common in Norway, but now the bear, wolf, and lynx are found only in a few areas, mainly in the north. Foxes and many species of marten, however, are common, and in certain areas badgers and beavers thrive.

Most of the rivers and lakes have a variety of fish, notably trout and salmon; the latter is found in at least 160 rivers, often in an abundance that attracts anglers from all over the world.

Of the large variety of birds, many migrate as far as Africa for the winter. In the north, people collect eggs and down from millions of seabirds, and, as far south as Ålesund, small cliff islands often are nearly covered by several hundred thousands of nesting birds. Several kinds of grouse are common in the mountains and in the forests.

THE NATIONAL REGIONS

There are four traditional regions of Norway, three in the south and one (comprising the "flapping fishtail") in the Arctic north. The three main regions of the south are defined by wide mountain barriers. From the southernmost point northward, a swelling complex of ranges jointly described as Langfjellene ("the Long Mountains") divide eastern Norway, or Østlandet, from western Norway, or Vestlandet. An eastward sweep of the mountains separates Østlandet in the north from the Trondheim region, or Trøndelag. And, where the southern half of Norway's total length ends, northern Norway, or Nord-Norge, begins.

Eastern Norway. Østlandet has more than half of Norway's population of about 4,000,000, of whom some 700,000 live in the metropolitan area of the national capital, Oslo. Another 1,000,000 persons live close by, mainly in the many industrial cities and urban agglomerations on both sides of Oslofjorden. By the 1970s almost continuous urbanized belts were spreading from Oslo out into the countryside.

The agricultural core of the Østlandet lies in the lowlands extending eastward and southward to the Swedish border. With suitable precipitation during the growing season, the highest July temperatures in Norway, a soil consisting of relatively rich marine deposits, and large nearby markets, the land is intensively cultivated. There are even a number of large, heavily mechanized farms producing cereal grains, which generally do not grow well in such latitudes.

Most of the farms, however, are small. To supplement their income from domestic animals and vegetables, a great number of farmers pursue forestry as a secondary occupation, since most of the forests are a part of farm acreages. Norway has never had the agricultural villages that are common elsewhere in Europe. The more densely populated areas of the country have grown up around crossroads of transportation from which people have moved to the cities and suburbs. Thus, there is actually little borderline between the rural and urban populations. For many years Oslo has attracted settlers from all over the country, becoming a national melting pot surrounded by the most important agricultural as well as industrial districts of Norway.

Urban and rural populations

In the interior of Østlandet, farms are located along the sides of the broad valleys, the bottoms of which contain only washed-out deposits of soil. Frost is more frequent in spring and fall than in the coastal areas. The largest forests in Norway are found between the Swedish border and the Glåma River east of Oslo. The coastline facing Denmark across the Skagerak passage, stretching from Oslofjorden to the southern tip of Norway, is densely populated and thriving, with small towns, coastal villages, and many small farms. Centred on the city of Kristiansand, this southern area is sometimes set apart as a fifth region, Sørlandet. The idyllic coastline has developed into Norway's foremost summer-vacation area. The land is hilly, but the growing season is slightly longer than around Oslo. The interior, with narrow valleys running up into the beginnings of Langfjellene, is very sparsely populated, and the people of the scattered settlements depend more on cattle raising and forestry. Economically and in view of other practical aspects, however, Sørlandet is usually included as part of Østlandet.

In all, about 50 percent of Østlandet is forested, giving the region 55 percent of Norway's total forest resources, equivalent to its share of Norway's total area of fully cultivated land. In mining and manufacturing, Østlandet has a share of more than 60 percent of the nation's total production value and of its total trade. These lions' shares of the national wealth, combined with the concentration of economic activity around Oslofjorden, secure for Østlandet the highest average income per household of the four regions: 27,700 kroner (7.14 kroner=$1 U.S.; 17.13 kroner=£1 sterling, on December 1, 1970), compared to 25,500 for Vestlandet, 23,200 for Trøndelag, and 23,100 for Nord-Norge (1969).

Western Norway. The narrow coastal zone of Vestlandet abuts into the North Atlantic Ocean; it has many islands and steep-walled, narrow fjords cutting deep into the interior mountain region. The major exception is the wide Jæren Plain, south of Stavanger. With rich glaciated soils, exceptionally mild winters, long growing seasons, and plentiful precipitation, Jæren has the highest agricultural yields in all of Norway, and its farmers are not distracted by the necessity of forestry or fishing.

Stavanger, Norway's fourth largest city, has become an expanding industrial centre, particularly in canning and engineering. It is also the main base for the explorations in the North Sea for undersea oil deposits. To the north, the city of Haugesund thrives on fishing and other industry. The island of Karmøy comprises a notably rich agricultural area. The inland fjord districts of Vestlandet are more sheltered, with rich fruit districts specializing in apples. The flowering of the trees beneath snowcapped mountains adds an extra dimension of beauty to the world-famous "land of the fjords."

The city of Bergen is the natural centre of Vestlandet. Since the late Middle Ages, when it was an active trading centre of northern Europe, the city has had a more international character than any other in Norway. The typical "Bergenser" is often caricatured as an easy-talking city boaster with a "go-ahead spirit" matched only by the people of the Sunnmøre district, farther north on the coast. Sunnmøre, with Ålesund as its local centre, contains many engineering firms, and the bulk of Norway's growing furniture industry is gathered on its rocky coast. Fishing, which predominates farther north, around the cities of Molde and Kristiansund, is a main second occupation, as forestry is in Østlandet. Deep in the Vestlandet fjords lie many of Norway's largest smelting plants, constructed to exploit the great hydroelectric resources of the region.

Central Norway. Trøndelag is Norway's most typical agricultural region, with flat, fertile land around the wide Trondheimsfjord and the city of Trondheim. In addition, 30 percent of its area is forested, and it has a large percentage of Norway's mining industry. The main difference between Østlandet and Trøndelag is the cooler summers in the latter. The typical "Trønder" resembles the stolid farmer or lumberman of inland Østlandet. Trondheim, the second largest city of Norway, and for long periods the national capital, dominates the economic life of Trøndelag.

Northern Norway. Nord-Norge, most of which is above the Arctic Circle, is truly the "land of the midnight sun" in the summer, whereas during December and most of January no real daylight appears at all. Most of the region is filled with mountains with jagged peaks and ridges, even on the many islands. A long string of large islands jutting into the North Atlantic west of Vestfjorden form the Lofoten Wall. Numerous fjords scissor into this narrow strip of Norway's northern tail. Its few roads and severe winters make seaway transportation even more vital than in Vestlandet.

Isolation of the north

Some of the inland valleys have dense forests, but the short summers and poor soil limit agriculture to a largely

subsidiary activity, though the production of milk, meat, and potatoes is surprisingly high in many districts. Fishing continues to dominate the economy except in certain mining and industrial centres.

THE NORWEGIAN PEOPLE

In most parts of Norway, the nucleus of the population is Nordic in heritage and appearance. Between 60 and 70 percent have pure-blue eyes. An influx of Alpine and Mediterranean peoples has been strong in southwestern Norway. In the far north there has been some intermarriage with the *samer*, Lapps, or Laplanders, who live on the plateau called the Finnmarksvidda, moving their reindeer herds down to the coast for summer grazing. The Lapps, dark-haired and of short stature, came to Norway before anyone else, at least 10,000 years ago, probably from Central Asia. Nord-Norge also has about 7,000 *kvener*, immigrants from Finland.

Language and religion. The language of Norway belongs to the North Germanic branch of the Germanic language group and is a relative of English, which belongs to the West Germanic branch. The Norwegian alphabet has three extra letters, *æ, ø,* and *å,* pronounced respectively as the vowels in b*a*d, b*u*rn, and b*a*ll. Modern Norwegian has many dialects, but all of them, as well as the Swedish and Danish languages, are understood throughout all three countries. Until about 1850 there was only one written language, called Riksmål, or "official Norwegian," which was strongly influenced by Danish during the 500-year union of the two nations. Then Landsmål, or "country Norwegian," was created out of the rural dialects. After a long feud, mostly urban–rural in makeup, both forms received equal status under the terms Bokmål, or "book Norwegian," and Nynorsk, or "new Norwegian." For 80 percent of the schoolchildren, Bokmål has been chosen as the main language in local schools.

Dialect status

About 96 percent of Norwegians belong to the Evangelical Lutheran National Church, which is supported out of state funds. The more than 130,000 persons outside this establishment in 1960 included Pentecostalists, Lutheran Free Church members, Methodists, Baptists, and Roman Catholics, in descending order of membership.

Population trends. Since World War II, Norway's birthrate has remained fairly steady at 16 to 20 per 1,000 inhabitants. This compares with about 29 at the turn of the century and a slump to 15 during the 1930s. The mortality rate has declined from about 15 per 1,000 at the turn of the century to ten in the 1970s, due mainly to reduced infant mortality and campaigns against tuberculosis, once a national scourge. The average life expectancy has risen from 50 to 71 years for males and from 54 to 76 for women since 1900.

Patterns of migration

Immigration and emigration are fairly well in balance, at around 15,000 annually, and this, in conjunction with the birthrates and death rates, produces an annual population increase of only about 32,000 a year. Internal migration, however, is steadily increasing due to the shift from rural to urban occupations. In 1968 such migrants numbered more than 180,000. From 1946 to 1965, the percentage of people living in urban areas increased from 49 to 63. Demographers foresee an increase in Norway's population from the present nearly 4,000,000 to nearly 4,700,000 by 1990. Of these, it is estimated that slightly fewer than 20 percent will live in the scattered settlements outside urban centres.

THE NATIONAL ECONOMY

Economic organization. Only about one-fourth of Norway's commodity imports are consumer goods, the rest consisting of raw materials and capital goods. The rate of reinvestment has been extremely high in Norway for many years. This is reflected in the rising employment in building and construction. Even more rapid growth, however, has been registered in commercial and service occupations, as in most countries with a high standard of living. Manpower is in a continuing strong demand.

Of the nearly 18,000 industrial companies in Norway, fewer than 700 have more than 100 employees. Nonetheless, they account for half of the industrial labour force and 56 percent of the production. The smaller companies are usually family owned, whereas most of the larger ones are joint-stock companies. Foreign interests control companies accounting for about 11 percent of total production. Only a few larger concerns are state owned, and even these are usually run with almost complete independence.

Industrial ownership

Agriculture and fishing are strongly organized enterprises, and they are subsidized by the state. In remote districts, private industry may receive special incentives in the form of loans and grants or tax relief. Taxes are high, with sharply progressive income taxes and a "value added" tax of 20 percent on all economic activity imposed directly at the manufacturing level rather than as a sales tax. Total tax revenues are equivalent to almost 40 percent of the gross national product, but half of this represents transfers of income; that is, it is returned to the private sector in the form of price subsidies, social-insurance benefits, and the like. All this has added to economic problems of inflation, but increases in productivity made possible a growth of more than 43 percent in real income for industrial workers from 1959 to 1969. The strongly centralized trade unions and employer associations respect one another as well as government guidelines, thus helping to control the rapidly expanding economy.

Foreign exchange. Foreign trade, in the form of commodities exported chiefly to western Europe or as shipping services throughout the world, accounts for nearly 40 percent of Norway's national income. For more than a century, Norway has been among the four leading

Norway, Area and Population

	area*		population	
	sq mi	sq km	1960 census†	1970 estimate‡
City Counties (*byfylker*)				
Bergen	18	47	116,000	116,000
	19	50		
Oslo	166	430	476,000	487,000
	175	453		
Counties (*fylker*)				
Akershus	1,790	4,635	234,000	312,000
	1,895	4,909		
Aust-Agder	3,324	8,610	77,000	80,000
	3,557	9,212		
Buskerud	5,378	13,928	183,000	196,000
	5,766	14,933		
Finnmark	17,970	46,544	72,000	76,000
	18,783	48,649		
Hedmark	10,093	26,140	177,000	179,000
	10,558	27,344		
Hordaland	5,758	14,914	225,000	255,000
	6,017	15,584		
Møre og Romsdal	5,668	14,680	213,000	223,000
	5,821	15,076		
Nordland	14,011	36,288	237,000	243,000
	14,789	38,327		
Nord-Trøndelag	8,130	21,056	117,000	118,000
	8,673	22,463		
Oppland	9,315	24,125	166,000	172,000
	9,773	25,313		
Østfold	1,511	3,914	203,000	219,000
	1,614	4,180		
Rogaland	3,273	8,477	239,000	266,000
	3,529	9,141		
Sogn og Fjordane	6,884	17,829	100,000	101,000
	7,168	18,566		
Sør-Trøndelag	6,993	18,111	212,000	232,000
	7,305	19,919		
Telemark	5,477	14,186	150,000	157,000
	5,913	15,315		
Troms	9,699	25,121	128,000	137,000
	10,021	25,954		
Vest-Agder	2,632	6,817	109,000	123,000
	2,811	7,280		
Vestfold	825	2,137	160,000	173,000
	856	2,216		
Total Norway	118,914§	307,988‖	3,591,000‖	3,866,000‖
	125,051	323,883		

*Where two figures are given the first is the land area, the second, the total area. †De jure. ‡Norwegians on the islands of Svalbard (Spitsbergen) and Jan Mayen are considered as residents of the mainland communities in which they are registered and are thus included in the population estimate figures for Norway. §Converted area figures do not add to total given because of rounding. ‖Figures do not add to total given because of rounding.
Source: Official government figures.

MAP INDEX

NORWAY

Size of symbol indicates relative size of town • ○ ◉ ▣ ■

Elevations in metres

0 50 100 200 300 km

0 50 100 200 mi

shipping nations of the world, and Norwegian shipowners constantly have pioneered new trades and markets. The export of finished products has been boosted tremendously by the nation's membership in the European Free Trade Association (EFTA) since its founding in 1959. In 1970 Norway began negotiations for membership in the European Economic Community (EEC): in July 1972 a treaty was signed which would have made entry into the EEC effective January 1, 1973. This was rejected in a non-binding referendum in September 1972.

Agriculture and forestry. During the 1960s, the number of farms in Norway decreased by about 34,000, to fewer than 165,000, most of the abandoned acreage being absorbed into other farms. Only about 22,000 farms have more than 25 acres of farmland, while only about 40 have more than 250 acres. Labour for hire is scarce and most of the work must be done by the farmer-owner himself. Extensive mechanization and fertilization, however, have kept the total output on the increase. Although the country is more than self-sufficient in animal products, it remains dependent on imports for cereal crops.

In addition to the nation's 2,000,000 acres of farmland, the farmers own some two-thirds of Norway's 16,000,000 acres of productive forests. Forestry forms the basis for the wood-processing industry, which accounts for about 12 percent of Norway's total commodity exports, and it is of major importance for the nearly 80,000 farms that are so small that a second major source of income must be found in other occupations.

Employment in fisheries

Along the coast, fishing plays the same role that forestry does elsewhere. At the same time, it forms the basis for a large and growing fish-processing industry and offers seasonal employment for many farmers. Of the 61,000 fishermen registered in 1960, only a third had fishing as their sole occupation. Most vessels are owned by the fishermen themselves, the necessary crew members being paid by shares of gross income in a continuation of a centuries-old tradition of the sea. In spite of a decline in the total number of fishermen, mechanized methods have doubled—and in some years nearly trebled—the catch since 1962, with only two other nations surpassing the totals. A critical current problem is to avoid depleting the fish resources while maintaining the volume. About 76 percent of the catch, most of it near the coast, goes into fish meal and oil, but increasing quantities are being processed for human consumption in freezing plants. Fish offal is used as feed at mink farms.

Mining and metallurgy. Norway mines only a few ores in quantity, mainly pyrites and iron ore, but its ferrotitanium ore comes from Europe's largest deposit, in southwestern Norway. Coal is mined only at Svalbard. Nonetheless, the mountains of Norway have played a key role in the industrialization process by supplying huge quantities of hydroelectric energy. In 1969, the electrical energy production per capita was nearly twice that of Canada, the world's second largest per capita producer.

Almost half of Norway's annual production of hydroelectric power is consumed by its electrometallurgical plants, which place the country as the world's largest exporter of iron-based alloys and metals combined. It is among the major exporters of aluminum—here second to Canada—and of nickel, copper, and zinc. One company, which produces enough fertilizer to produce food for more than 20,000,000 people, is also the world's second largest producer of magnesium. Although the aluminum and nickel exports depend on raw-material imports, magnesium is made from seawater and locally available limestone or marble rock known as dolomite. Norway also has plenty of quartz for the iron-alloy industry, which is the primary feeder of these products to the steel industry of western Europe. In all, the electrometallurgical industries accounted for somewhat more than a quarter of Norway's total commodity exports in 1969, although they employed only 7 percent of the national industrial labour force.

Engineering and petroleum exploitation. Engineering is by far the largest single employer among Norway's industries. As late as in 1960, engineering remained an industry largely based on the domestic market. But with more liberalized world trade, especially that brought on by Norway's membership in EFTA, extensive specialization became a necessity. The result has been the fast-growing export of a large variety of products, from enormous marine tankers to integrated electronic circuits.

Oil and natural-gas reserves

Oil and natural-gas deposits with estimated exploitable reserves of at least 1,000,000,000 barrels of oil were found in 1969–70 in the Norwegian sector of the continental shelf stretching across the shallow North Sea to Denmark. Trial production began in the spring of 1971. Much larger finds are possible, especially on Norway's wide continental shelf in the Atlantic, extending northward all the way up the peninsula and into the Barents Sea, where there are Soviet claims to undersea oil rights. Norway's two existing oil refineries and its chemical industry account for some 7 to 8 percent of the commodity exports.

Shipping. Despite Norway's concern with producing goods for exports, commodity imports usually run about 42 percent higher than commodity exports. The deficit is sometimes completely offset, however, by the foreign-currency earnings of the merchant marine. Only Liberia, Great Britain, and Japan have larger shares of the world's merchant fleet. The age of the Norwegian fleet is well under the world average, and there is often a large backlog of orders for new tankers and giant bulk carriers. About 1,300 ships, accounting for more than 90 percent of Norway's total tonnage, sail only between foreign ports. Half of this foreign-going fleet consists of tankers. Most of the financing is secured abroad easily because of the trust in the profitability of Norwegian shipping. Of the crews, about a quarter are non-Norwegians, mainly from Spain and Asian countries.

Transportation. The elongated shape of Norway and its many mountains, large sparsely populated areas, and severe climate make special demands on transportation services. Only the Oslo region has sufficient traffic density to make public surface transportation profitable. About 1,700 vessels link the many fine ports along the sheltered coast. In most of Norway, regular overland transportation services are so expensive that the government must provide or subsidize both establishment and operation.

Bus transport plays a key role in public transportation, aided by almost 200 fjord ferries. The number of private automobiles is rapidly increasing, creating parking problems and traffic jams in the major cities. Of the 44,700 miles (72,000 kilometres) of public roads, only about 15 percent are hard-surfaced, and only shorter stretches of mostly substandard motorways have been completed. Demands for an additional 19,000 miles (30,000 kilometres) of road are growing, as are those for the comprehensive reconstruction of the many narrow, winding roads. The main Oslo-to-Bergen route still remains closed for four to six months during the winter.

Railways

The 2,600 miles (4,200 kilometres) of railway are operated by the Norwegian State Railways (NSB), which sustains large annual operating deficits. Vestlandet has never had north–south railway connections, only routes running east from Stavanger and Bergen to Oslo and from Møre og Romsdalen to one of the two routes linking Oslo and Trondheim. The connection to Bodø was completed in 1962. Farther north, the only railway is the extension of the Swedish railway system to Narvik, which is used mainly to carry iron ore for export. Of the three other links with Swedish railways, one runs from Trondheim and two from Oslo—the southernmost connecting Norway to the continent via the Swedish and Danish railways.

Aviation is an increasingly popular means of passenger transportation, especially in the northern regions and during the winter months. Extensive state aid is needed, however, to finance the building of airfields. The one private airline has about the same proportion of the domestic air traffic as the Scandinavian Airlines System (SAS), which is 50 percent state owned, with three-sevenths of total capital from Sweden and two-sevenths each from Denmark and Norway. SAS has pioneered commercial flights across the Arctic. Four commercial airports throughout the country are equipped to handle international flights.

THE NORWEGIAN GOVERNMENT

Norway is a constitutional hereditary monarchy. The king acts as administrative head of government and selects his own Cabinet. The legislative body is the 150-member Storting, elected by vote of all persons over 21 years of age.

Organization. The constitution of Norway, drafted in 1814 when Norway left the 500-year union with Denmark, was influenced by British political traditions, by the constitution of the United States, and by French revolutionary ideas. Amendments can be made by a two-thirds majority in the Storting. Unlike many parliamentary forms of legislature, the Storting cannot be dissolved during its four-year term of office. If a majority in the Storting votes against an important Cabinet issue, the minister responsible or even the whole Cabinet resigns. In legislative matters, the king has a suspending right of veto, but this has never been exercised since the 91-year union with Sweden was dissolved in 1905.

Five political parties were represented in the Storting in the 1970s. The Labour Party, during its 25 years of rule up to 1965, practiced moderate Socialism and established a few large state-owned factories. It nationalized private enterprises only under special circumstances. The Conservative Party advocates a modern progressive conservatism and accepts extensive government control and transfers of income. The Centre Party, called the Agrarian Party until 1958, has a policy directed beyond special group interests.

The Christian People's Party is concerned primarily with maintaining Christian principles in public life. The Liberal Party, supported by puritanic rural groups as well as city radicals, stresses social reform. Until 1961 the Communist Party was represented in the Storting. Since 1965 the four non-Socialist parties have ruled in coalition, mainly continuing along the lines followed by the Labour governments. Before elections the political parties nominate their candidates at membership meetings in each of Norway's 20 *fylker* (counties). Each *fylke* (county) elects from four to 13 representatives, according to its population, with party representation allotted on the basis of the percentage of the vote it received.

Proportional representation

Local administration. The cities of Oslo and Bergen constitute two of the country's 20 *fylker*. The others are divided into a total of 404 rural and 45 urban municipalities, with councils elected every fourth year, two years after the Storting elections. For the country as a whole, the municipal elections tend to mirror the party division of the Storting. The municipal councils elect a board of aldermen and a mayor. Much municipalities also employ councillors for such governmental affairs as finance, schools, social affairs, and housing. Norwegians pay direct taxes to both federal and municipal governments.

The *fylke* can levy taxes on the municipalities for roads, hospitals, secondary schools, and other joint projects. The *fylke* councils comprise delegates from the municipalities, while the *fylke* governor is appointed by the Cabinet.

Justice. The Supreme Court makes the final decision on legal practice. There are 104 primary courts and 464 conciliation councils. Norway is divided into five court-of-appeal regions. The rights of the citizen are guarded also by the ombudsman, who handles about 1,000 complaints a year and finds about 20 percent of them to be justified.

Defense. Military service of 12 to 15 months, plus refresher training, is compulsory for all fit Norwegian men between 22 and 44 years of age. Nonetheless, Norway's defense force is far too small to protect all of its territory against a major aggressor. Its strategy is designed to defend key areas, especially in the north, until forces from other members of the North Atlantic Treaty Organization (NATO) can be moved in. The Norwegian units have great mobility, and, with its important strategic location as NATO's northern flank and its myriad of fjords to serve as naval bases for fleets in the North Atlantic, Norway has the ultimate in early warning systems.

International military alliances

The NATO headquarters for Northern Europe is at Kolsås, near Oslo. Foreign troops and nuclear charges, however, are excluded from Norway by law except in cases of war or immediate threat of war. The Norwegian Air Force has a number of ground-to-air missiles and various aircraft from other Western nations. The navy consists of heavy coastal artillery and such light vessels as gunboats, torpedo boats, submarines, and corvettes up to 1,800 tons. In the 1970s the total military personnel numbered about 40,000, but an additional 70,000 men were members of local home-guard units and could be mobilized quickly in cases of emergency.

SOCIAL CONDITIONS

Education. Most of Norway's municipalities have extended compulsory schooling from seven to nine years. After the eighth grade, pupils may choose from among advanced courses in Norwegian, mathematics, and English (compulsory from the fifth grade) and such optional subjects as German. The ninth grade offers an even wider choice, including vocational training. With three years of additional high school, students may take the examinations leading to university study. About 10 percent of the more than 27,000 college and university students study abroad. The institutions in Norway are being expanded to accommodate the anticipated doubling of this student group in the final decades of the 20th century.

Nearly 85,000 students attend vocational schools and nearly 7,000 the folk high schools, boarding schools offering a one-year course designed for 17-year-old students from rural areas. Only a few of Norway's schools charge tuition, and all students can get government loans on favourable terms.

Health and welfare. Norway's several large, active health organizations have nearly 2,000,000 members. Compulsory membership in a national health-insurance system secures for all Norwegians free medical care in hospitals, compensations for doctors' fees, and free medicine for certain chronic illnesses. Cash benefits during illness or pregnancy, covered by another insurance fund, are compulsory for salaried employees and optional for the self-employed. Norway has one doctor per 700 inhabitants, most of them working in hospitals, the majority of which are owned by the state, counties, and municipalities. Extensive programs of preventive medicine have conquered Norway's ancient nemesis, tuberculosis. There is also a well-developed system of maternal and child health care, as well as compulsory school health services and free family counselling by professionals. A public dental service provides care for 90 percent of the children between seven and 15 years of age, and plans are under way to extend it downward to three years and upward to 18 years. There is one dentist per 1,200 inhabitants.

National programs in medicine and welfare

A "people's pension" was established in Norway in 1967, to ensure the entire population a standard of living reasonably close to the standard that the individual had achieved during his working life. The pension covers old age as well as cases of disability or loss of support. The premiums are paid by the individual members, employers, municipalities, and the state. The basic pension is adjusted every year, regardless of the plan's income. Supplementary pensions vary according to income and pension-earning time. The state pays a family allowance for all children up to 16 years of age.

Housing and living standards. Norway still feels the housing shortage created by World War II, and this has been aggravated by high costs in the densely populated urban areas. But housing standards have improved tremendously, and most families live in houses built since the war, a majority of them financed by state loans on favourable terms.

Norway ranks ninth among the nations of the world in gross national product per capita of population. In the postwar years individual income per capita has tripled in real terms. Tax rates progressing upward with income and the greatly increased social-security benefits, allocated mainly according to need, have contributed to a levelling of incomes. The perennial shortage of labour, especially of skilled workers, has had a parallel effect.

The consumer of the 1970s spends a smaller share of his income than formerly on food, beverages, and tobacco, since the saturation point apparently has been reached. Travel and leisure activities are increasing their share rapidly, as are such household goods as electrical appliances. In 1959–68 the number of automobiles per inhabitants increased from one in 21 to one in six. A four-week vacation every year with somewhat more than full wages was established by law in 1964. Working hours are restricted to nine hours a day and 42½ hours per week. A five-day work week had become the rule by the late 1960s.

Culture. Located on the outskirts of Europe and with much of its inland population almost completely isolated until this century, Norway has been able to preserve much of its old folk culture. On the other hand, as seafarers and traders the Norwegians have always received fresh cultural stimuli from abroad. A number of Norwegians have made important contributions in return, notably playwright Henrik Ibsen and the composer Edvard Grieg. The Norwegian recipients of the Nobel Prize for Literature are Sigrid Undset, Bjørnstjerne Bjørnson, and Knut Hamsun.

Literature and performing arts

Although Norway comprises one of the world's smallest language communities, the country is first in books published per capita. The annual number of new titles is close to 2,500, of which two-thirds are of Norwegian origin. Literature is subsidized through tax exemption and government purchasing for libraries. In all, there are 5,000 public or school libraries, which annually loan more than 14,000,000 books.

Permanent theatres have been established in four cities, and the state travelling theatre, the Riksteatret, organizes tours throughout the country, giving as many as 1,200 performances annually. Several Oslo theatres have closed down in recent years, however, and the Norwegian Opera, opened in 1959, requires state subsidies, as do five other theatres. Films in Norway are subject to censorship, primarily on grounds of violence and, to lesser extent, for erotic content. The production of Norwegian-made feature films is subsidized, but they usually number fewer than ten a year.

The visual arts

In addition to its National Art Gallery, Oslo opened a special museum in 1963 honouring Edward Munch, probably Norway's most famous painter. In 1968 the Sonja Henie–Niels Onstad Art Centre was opened near Oslo; it contains modern art from all over the world.

Norwegian painters of the 20th century have excelled in murals to an extent rivalled only by Mexican painters. Other artists are world known for their multimedia assemblages, pictorial weaving, and nonfigurative art in sculpture as well as painting. The works of Gustav Vigeland have been assembled in Oslo's Vigeland sculpture park in a spectacular display centred around a 60-foot granite monolith containing 121 struggling figures.

Architecture has drawn inspiration from medieval stave churches of upright logs as well as houses of horizontal logs notched at the corners. Private houses, almost all of wood, are made to fit snugly into the terrain. For larger buildings, steel and glass are supplemented by concrete that often is shaped and texturized with considerable imagination.

Arts and crafts and industrial design flourish side by side, often inspired by archaeological finds from the Viking age (around the year 1000), by the culture of the northern Lapps, and by advanced schools of design and the Norwegian Design Centre, one of the largest of its kind in the world. Norway has markedly increased its exports of furniture, enamelware, textiles, tableware, and jewelry, much of which incorporates design motifs reflecting these cultural heritages as well as avant-garde styles.

Recreation. The Norwegians have the special advantages of abundant space and a traditionally close contact with nature. Cross-country skiing is a national pastime in the long winter season. In the Olympic Winter Games, the tiny nation has won far more medals than any other nation. The Norwegians have more than 200,000 second homes, mainly along the sheltered coastline and in the mountains. Even from downtown Oslo, it takes only a 20-minute drive to get into deep forests, and on a nice winter Sunday the surrounding hills abound with skiers.

Scientific research. Science and research have limited means in a small country. In the natural sciences, however, reflecting the country's intimacy with an overpowering physical environment, individual efforts of Norwegians have won world fame. Among the most widely known are the explorers Fridtjof Nansen, Roald Amundsen, and Thor Heyerdahl. Norway also has pioneered many industrial breakthroughs, especially in the area of electronics.

The communications media. Some 185 newspapers are published in Norway, about half of them daily, except Sundays and holidays, on which no papers are published. Although most of them are small, average circulations increased from 7,700 to 10,000 in the period from 1950 to 1970. Most newspapers have affiliations with political parties, but this is reflected very little in readership. Even though the Labour Party is supported by nearly half of the electorate, the labour press has less than a quarter of the total circulation. Press ethics are on a high level, and the independence of the editors is universally recognized.

Broadcasting is a state monopoly, with no commercial sponsors or advertising over the one radio and one television network. Educational and informational programs are given priority over entertainment, and day-to-day management is quite independent of the state. Every home with radio and television pays an annual fee. Regular colour-television broadcasts are expected by 1973.

Prospects. With a high standard of living evenly distributed across a healthy, stable democracy, Norway now is concerned mainly with two of the newest problems of advanced industrial societies, the stresses of living in such a society and the threats of pollution posed to the environment. The latter is a growing problem, but so far none of the damage is irreparable. The development of hydroelectric resources, however, has destroyed the beauty of several spectacular waterfalls. The hope of energy-hungry Norway is that the "white coal" of the mountains will soon be supplemented with "black gold" from beneath the North Sea.

BIBLIOGRAPHY

General: *The Norway Year Book* (annual), articles by experts in various fields covering new developments; *Facts About Norway*, new ed. (1970), a popular presentation of Norway today; GUNNAR JERMAN and FINN P. NYQUIST, *New Norway* (1970), 220 pictures, most in colour, with brief texts; ERLING WELLE-STRAND, *Norway: Pictures and Facts* (1970); CENTRAL BUREAU OF STATISTICS, OSLO, *Statistical Yearbook* (annual); *Historical Statistics* (1968); PHILIP CARAMAN, *Norway* (1969), a travel book including much historical as well as current data; PHILIP BOARDMAN, *Northern Paradise* (1963).

Geography and economy: MAGNE HELVIG and VIGGO JOHANNESSEN, *Norway: Land, People, Industries*, 2nd ed. (1968), an extensive presentation in popular form; OLA HEIR and OLE RØMER SANDBERG, JR., *Norwegian Agriculture and Its Organisations* (1966); NORWEGIAN WATERCOURSE AND ELECTRICITY BOARD, OSLO, *Hydro Power Developments in Norway* (1969); NORWEGIAN SHIPOWNERS' ASSOCIATION, OSLO, *Review of Norwegian Shipping* (1970); HERBERT DORFMAN, *Labor Relations in Norway*, rev. ed. (1966), historical as well as current information.

Government: JAMES A. STORING, *Norwegian Democracy* (1963), an American's view of the political situation; TIM GREVE, *Norway and NATO* (1968); OLAV HOVE, *The System of Education in Norway* (1968); SOCIAL MINISTRY, OSLO, *The Ideas Behind the Welfare State and Its Structure* (1968).

Culture: INGEBORG LYCHE, *Promoting the Arts in Norway* (1969); ARNE ØSTVEDT, *Music and Musicians in Norway Today* (1961).

(Ja.C.)

Nō Theatre

One of the traditional theatrical forms of Japan, Nō is unlike Western theatre in which characters come onto the stage to develop a series of incidents that lead to a dramatic climax and conclusion. The term Nō denotes skill or accomplishment. The performers of Nō are neither actors nor "representers" in the usual sense; they are simply storytellers who use their visual appearances and their

movements to suggest the essence of their tale rather than to enact it. Nō is, thus, not a drama in which definite ideas can be conveyed symbolically in character and event, as in a Western play. Nō employs the visual and the auditory to evoke response beneath—or above—the level of concrete thought.

Performance of Nō

A brief sketch of one of the most famous of Nō plays, *Takasago* by Zeami Motokiyo, best suggests the quality and nature of the Nō theatrical tradition. The stage itself is not unlike a formal architectural set for an Elizabethan or Greek play. The pavilion-like, roofed stage juts into the right side of the audience and, retreating back and to the left, a covered bridgeway fronted by three pine trees leads to a brilliantly coloured curtain, the only colour on stage save for an ancient, gnarled pine painted on the back wall of the stage. A border or path of white gravel separates stage and bridgeway from the audience. A kimono-clad chorus sits at the extreme right of the stage, facing centre, while facing forward from the front of the inner stage are a flutist and three drum players. The flute begins a high plaintive air.

A heavily robed Shintō priest with two attendants crosses slowly down the bridgeway onto the stage; he turns to face his attendants. To the flute and drum taps the three sing "Today we don our travelling dress, today we don our travelling dress, long is the journey before us," words then echoed by the chorus and repeated by the trio. The Priest then says

> I am Tomonari, priest of the Aso Shrine in Higo Province in Kyūshū. Never having seen Miyako, I intend to journey thither and shall take the excellent opportunity thus offered me to visit the Bay of Takasago in Harima Province.

They then sing a "travelling song" that in a few lines describes their long voyage, after which Tomonari sits, with attendants on his right. He announces "Travelling in haste, we have now reached Takasago. Let us stop here awhile and enquire about this place."

To further music, two other actors enter the bridgeway and stop, facing one another. One carries a broom and wears an "old woman" mask and wig, the other a rake and "old man" disguise; both are kimonoed and robed in dignified manner. They sing, beginning "The spring breezes murmur in the Takasago pine. The day is closing in. . . ." The voice reveals that the "old woman" is played by a man. The action that follows includes formal song, speech, mime, and dance involving the elderly couple and Tomonari, with commentary by the chorus. The performers assume lengthy, frozen poses or move with slow and weighty precision under their heavy garments. Finally, the couple are established as poetic personifications of the twin pines of Sumiyoshi and Takasago, symbols of longevity and conjugal fidelity.

During an interlude following the couple's exit, another actor enters and, in response to Tomonari's questions, relates the legend of the twin pines. The play's short concluding section takes place in Sumiyoshi. The chorus and the deity of the shrine, who is played by the actor that played the old man in a different costume and mask, explains the essence of the legend that has been presented. The deity stamps twice and the play is ended. (From *The Noh Drama*; Charles E. Tuttle Co., Inc., 1960.)

NATURE AND OBJECTIVES

Takasago represents a mysterious, exotic species of theatre. It is a short work and others are always played with it; its dialogue is sparse, a mere frame for the movement and music; little "happens"; and the total effect is less that of a present action than of a simile or metaphor made visual. Such is the tradition of Nō theatre.

The "action" of Nō

Though Nō does put a story on the stage, it does not aim at an interest in the story as such. If story-centred theatre were to be considered prose, Nō would be an equivalent of poetry. Suggestion and not statement is the medium's principal means, and the texture of its words and movements is so dense with allusion, symbol, and fine shadings of meaning that no translation can approximate it, nor can a spectator not steeped in Japanese cultural history grasp it fully.

The author of *Takasago*, Zeami Motokiyo (1363–1443), and his father, Kan-ami Kiyotsugu (1333–84), wrote many of the most beautiful and exemplary of Nō texts. Zeami formulated the principles of the Nō theatre that guided its performers for many centuries. His *Kakyō* ("The Mirror of the Flower," 1424) described the composition, the recitation, the mime and dance of the performers, and the staging principles of Nō. As has been noted, the term Nō denotes skill and accomplishment, and Zeami described this theatre as one of "elegant imitation."

He advised on selection of properly classical characters to be portrayed, from legend or life, and on the proper integration of the visual, the melodic, and the verbal to open the eye and ear of the mind to the supreme beauty he crystallized in the word *yūgen*. Meaning literally "dark" or "obscure," *yūgen* suggested beauty only partially perceived—fully felt but barely glimpsed by the viewer.

Nō began as a festival drama at shrines and temples in the 12th or 13th century but was continually refined up to the years of the Tokugawa period (1603–1867). It became a ceremonial drama performed on auspicious occasions by professional actors for the warrior class—as, in a sense, a prayer for peace, longevity, and the prosperity of the social elite. Outside the noble houses, however, there were performances that popular audiences could attend. The collapse of the feudal order with the Meiji restoration (1868) threatened the existence of Nō, though a few famous actors maintained its traditions. After 1945, however, a large number of educated youth came to enjoy Nō not simply for its status as a "classic theatre" but as a perfected and refined contemporary stage art.

FORMS AND TRADITIONS

In writing and in all phases of staging and performing, Nō is theatre at its most stylized. The educated spectator knows the story well and can follow the many allusions to ancient literary, legendary, and historical phenomena. Like the worshipper who has innumerable times partaken in a rite, he attends not for novelty but for a renewed personal experience of the rite itself.

Nō repertoire. About 2,000 Nō texts survive in full, of which about 230 remain in modern repertoires. Their plots were refined from stories available at the time of writing from Chinese or Indian sources, but especially from tales and legends of old Japan and from anthologies of Japanese poetry written as early as the 8th century. Themes and characters were diverse: the wonders worked by Buddha and native gods, deeds of warriors, relations of families of lord and retainer, hermits and visitants from heaven, animals and plants.

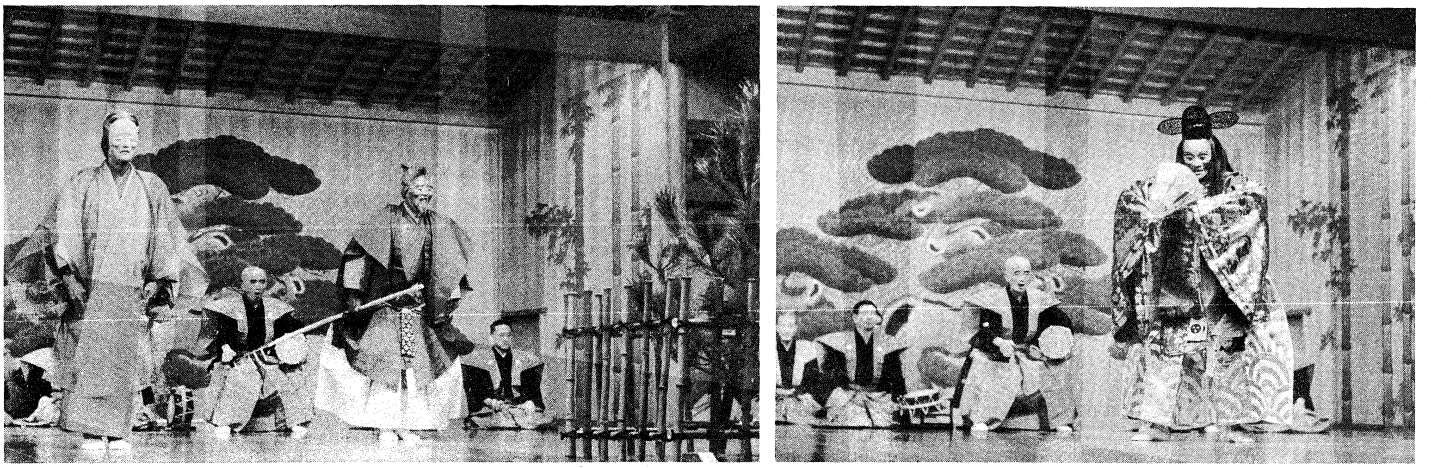
Types of Nō

The Nō repertoire includes five types of play. The first type, *waki* Nō, involves a sacred story relating the origin of, or miracle associated with, a shrine, as in *Takasago*; the second, *shura mono*, deals with defeated or, less frequently, victorious warriors; the third, *katsura mono*, or female-wig play, has a female protagonist; the group contains many of Nō's masterpieces. The fourth type is varied in content, including *genzaimono*, tales performed as if contemporary rather than from legend, and *kyōjo mono*, or "mad woman" pieces. Devils, strange beasts, and supernatural beings are featured in the *kiri* Nō, the fifth type. A program of Nō generally comprises three to five plays selected from the five types so as to achieve both an artistic unity and the desired mood; invariably a play of the *kiri* type is used as a concluding work.

Two special features of a Nō program are *okina* and *kyōgen*. *Okina* is essentially an invocation for peace and prosperity. The *kyōgen* is an interlude of humour between plays to relieve the tension engendered by them; it is also the name for the actor who relates the story between the two parts of plays, as in *Takasago*.

Structural elements of plays. The structure of Nō plays can be viewed in terms of the different types of performers, of recitation, of dance and mime, and of music, all of which have definite formulations that can be varied from play to play.

Three major Nō roles exist, *shite* ("actor"), *waki* ("by-



Scenes from Takasago.
(Left) The spirit of the pine tree assumes human form as an old man (*shite*) and woman (*tsure*) to relate its story. (Right) *Shite*, appearing in costume, mask, and headpiece indicating divinity, here represents the god of the pine tree and performs a solemn dance.
Hisao Maejima

Roles and schools of performers

stander"), and *kyōgen*. Each is a specialty having several "schools" of performers, and each has it own "acting place" on the stage. Subsidiary roles include those of attendant (*tsure*), of a "boy" (*kokata*), and of nonspeaking "walk-on" (*tomo*). In *Takasago* the *shite* played both the old man and the deity Myōjin; the *waki*, the priest Tomonari; the *kyōgen*, the "man of the place" in the interlude. The old woman was a masked *tsure*, and the priest's attendants were *waki tsure*. The Nō play is essentially for the *shite*; he plays the protagonist role be it male or female. Only he dances and mimes. In two-part plays the *shite* changes costume and mask (if used) in the interlude and appears in the second part as a different character to extend and conclude the story. The *waki* is usually a Shintō priest, an imperial envoy, a travelling monk, or similar person whose interest, despite his "bystander" role, elicits from the *shite* and *kyōgen* the details of the story. *Kyōgen* performers represent a local personage, essentially narrators of the tales basic to the plays. In a second, less representative, kind of Nō play with only one part the performers appear purely as narrators who speak entirely in the third person. In both of these the *shite* and the *waki* play major roles, and the "bystander" idea of the *waki* role is absent. The *kokata*, usually playing a boy, may also play an emperor or person of high social rank, a practice based on the belief that a young person is pure and thus sacred. This, too, is characteristic of Nō in suggesting a character's essence rather than representing his appearance. An uncostumed figure at the rear of the stage is the *kōken*, who not only looks after the actors' hand props and adjusts the costumes and stage props but also stands ready to fill in for the *shite* role if necessary.

From the early days of Nō, performers have specialized in one or another of these roles, or as musicians. Four of the five *shite* schools predate Zeami, the *waki* families go back to the late 16th century, and the *kyōgen* and musician groups began in the earliest years of Nō. *Shite-tsure* and *kokata* actors and chorus members are from the *shite* group; *waki-tsure* from the *waki*. These schools, together with the careful preservation of texts and the passing of performing arts from father to son or master to pupil, contributed greatly to the continuity of Nō traditions and styles over many centuries.

The eight-to-ten-man chorus (*jiutai*) of Nō both resembles the Greek usage and differs from it. Its most unique function is to recite words in the place of a performer; alternatively, it may provide only musical accompaniment to dance or, like the Greek chorus, describe the entrance of a deity. In the "third-person Nō," the chorus may replace a performer in reciting text.

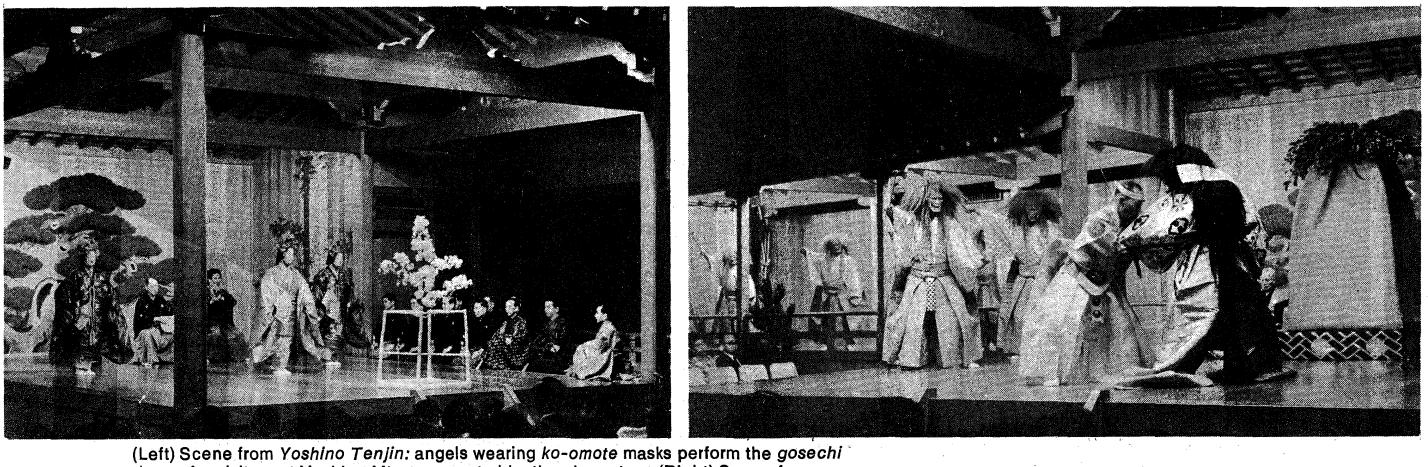
The recitation (*utai*) of Nō is the most important element in the performance. Each portion of the written text (*utaibon*) carries a prescription of the mode of recitation—as well as of accompanying movement or dance—although application of this may be varied slightly by each school or individual performer. Each type of dialogue and song has its own name: the *sashi* is like a recitative; the *uta* are the songs proper; the *rongi*, or debate, is intoned between chorus and *shite*; the *kiri* is the chorus with which the play ends. Singing is specified as high or low pitched. Plain speech is classed as name introduction (by the *waki*), dialogue, and narration. Although modern Nō is sometimes performed by all-woman casts, men generally fill all roles and, even as female characters, do not attempt to imitate a female voice. To do so would not be relevant to the spirit of Nō, for the actor, dressed and masked as a female, is on stage to suggest and never to impersonate.

Dance, mime, and music

Though at base it is more narration than drama, Nō requires at many moments a broad range of stylized movements. Several of these are carefully defined. *Furi* refers to intricate patterns of movement that accord with the words being recited, making visual their meanings; the patterns sometimes involve extravagant leaps. Pure dance, *mai*, is not ordinarily accompanied by recitation; either fast or slow, it is stately and elegant whether it is a god dance or a dance of female spirits. The number of movements of *mai* are rigidly specified, as are its gestures, postures, and steps; and the dancers carry brilliantly coloured fans. The most elaborate movement occurs in the *kiri*, or ending chorus, in which the *furi* patterns become very close to dance proper rather than to elaborate mime.

The music of Nō, *hayashi*, is played on a flute, which provides the melodic theme of songs and dances, and from a small and a large hand drum and a flat floor drum that provide the rhythmic base. Entrances of performers are marked by musical themes. The dance accompaniments evolved from an early period of Nō when persons around the dancer struck the stage with bamboo or sticks or sang to excite him into a state of rapture or possession by a god. Such primitive devices were refined into the sharp, piercing music of the dignified Nō dance.

Staging and design conventions. The traditional Nō stage was described earlier and is suggested in the accompanying photographs of productions. Until the beginning of the 17th century, however, viewers surrounded the stage on all sides. Temporary stages outside urban centres retained this aspect, in which the bridgeway is attached to the main stage by various means. Of the five main pillars on the stage, four are named for, and associated with, the performers; the fifth, the "eye-fixing pillar," serves as a landmark for the performers, whose vision may be limited by masks. Empty jars suspended under the stage act as resonators to amplify the sound of the dancer's stamping feet. Performers costume themselves in the dressing room and then, if they are to wear masks, put them on in the especially sacred "mirror room," where they "become the role" by contemplating their image.



(Left) Scene from *Yoshino Tenjin:* angels wearing *ko-omote* masks perform the *gosechi* dance for visitors at Yoshino Mt., represented by the cherry tree. (Right) Scene from *Momojigari:* female devils in *hannya* masks (expressing vengeance) attack the *waki,* a warrior figure. Structure, stage right, represents a mountain bearing maple leaves.
Hisao Maejima

The decor of the stage reflects the elemental simplicity of the Nō form. Only the painting of twisted pine and of young bamboo shoots decorate the back walls. Simple stage pieces or platforms may represent hills, a hermit's house, a rock, or a boat; they may be decorated with flowers, a tree, a sacred rope, or a candle. Smaller hand props may also be used, but by and large Nō, like the theatre of Shakespeare, leaves the connotations of time and place to the suggestion of its words and of its actors' mime and to its audience's imagination.

Costumes and masks

The sumptuous colour and material and the detailed line of the Nō costumes date from the period when support by noblemen, and even competition between nobles for actors and for brilliance of spectacle, permitted such largesse. The costumes, wigs, and head gear came to represent very elaborate aesthetic conventions, whereas the earliest Nō dress probably followed contemporary styles. A standard costume for each role evolved, but the costuming was not realistic. Poverty, for example, was not shown by rags but rather by suggestion in the kinds and arrangements of upper garments, head gear, and hand props.

The court dances known as *gigaku* and *bugaku* were imported from the Asian continent as early as the 7th century, probably the first entertainments in Japan to use masks. Names of mask makers are known from the 13th century. The mask of the deity Myōjin in *Takasago* is that of "a man of *Kantan*" created in the 14th century by a Tokuwaka to portray a noble youth. Over 100 kinds of mask have been classified as to divinities, men and women, young or aged, nonhuman figures, and the like. The principal mask is that of the divine figure *okina,* which has a movable jaw attached to the upper portion by string; others are generally in one piece. Ordinarily, the *shite* alone is masked, though a performer playing a woman or nonhuman figure is as well. The *waki,* essentially an observer, is never masked; and the kyōgen is rarely masked even when playing women, in which instances he relies on a special hair piece for suggestion of character.

CONTINUITY OF NŌ

Two factors have allowed Nō to be transmitted from generation to generation yet remain fairly close to earlier forms: first, the preservation of texts, containing detailed prescriptions of recitation, dance, mime, and music, and, second, the direct and fairly exact transmission of performing skills. On the other hand, Nō was subject to changing preferences of new audiences and new styles and patterns inevitably evolved. Further, there was constant refinement of received forms to express more clearly or intensely the objectives of Nō, but these were always and only minor deviations from traditional form. Even the differences between the five schools of *shite* performers represent only slight variations in the melodic line of the recitation or in the patterns of the *furi* or *mai* mime and dance.

Playwrights, theorists, critics

The great classic dramatists of Nō, Kan-ami, who wrote at least 18 plays, and his son Zeami, to whom 130 are attributed, performed all the essential functions of staging —text, music, dance, and movement—and themselves performed. Other playwrights include Zeami's son Motomasa (died 1432) and son-in-law Komparu Zenchiku (1405–68), Kanze Nobumitsu (1435–1516), Komparu Zempō (1454–*c.* 1520), and Kanze Nagatoshi (1488–1541). Zeami wrote his theoretical works, about 20 in number, for his descendants alone; carefully guarded and known as "the secret tradition," the writings became public only in the 20th century. Every aspect of Nō theatre, drawn from Zeami's rich personal experience, was discussed in them. Other treatises include Komparu Zenchiku's explanations, in a treatise of 1455, of Nō in metaphysical terms, but this and others were also long unknown outside the inner circles of Nō. Critics of Nō include Sakamoto Setchō (1879–1938) and Yamazaki Gakudō (1884–1944), whose commentaries on actors may have had subtle influences.

Experiments and prospects. After the late 15th century, Nō entered a long period of conservation and refinement of forms. Even the advent of the popular Kabuki theatre in the 16th and 17th centuries, with its sensuality and spectacle, failed to dim the support of Nō in high circles. In the 20th century some experimentation took place. Toki Zenmaro (1885–) and Kita Minoru (1900–) produced Nō plays that had new content but adhered to traditional conventions in production. Experiments to elaborate the humorous *kyōgen* interludes were not sustained. A so-called spotlight Nō adopted from Kabuki the long passage onto the stage through the audience and added a strong spotlight on the *shite,* but the innovation received little public acceptance.

Nō in the West

In the West the Irish poet William Butler Yeats discovered Nō through the dancing of Itō Michie and the writing of an American scholar of Japanese culture, Ernest Fenollosa. Between 1917 and 1921, Yeats wrote four dance-dramas that apparently reflected his interest. In 1954, two *shite* schools performed at the international theatre festival at Venice, and subsequently Nō companies visited several cities in Europe, performed and taught at the University of Washington in 1963, and took part in the arts exhibition at the 1968 Olympic games in Mexico City. Theatre specialists from around the world were deeply impressed by productions of Nō at the 1963 meeting of the International Theatre Institute in Tokyo.

The extreme simplicity and purity of the Nō form and its highly concentrated expression allowed much latitude for new interpretation and for an absorption of new materials. Not only did Nō seem likely to continue as a viable form in its own right, but it also offered fresh expressive means to a Western theatre that attempted to reach beyond naturalism in playwriting and production. What specific influences it would have on the performing arts of other regions remained, however, an open question.

BIBLIOGRAPHY. E.F. FENOLLOSA and EZRA POUND, *"Noh" or Accomplishment: A Study of the Classical Stage of Japan* (1916), a free translation of the texts of 15 Nō plays, including *Sotoba Komachi*, and a general outline of Nō; A. WALEY, *The Nō Plays of Japan*, new ed. (1950), full translation of 19 plays, summaries of others, one *kyōgen* play, and a bibliography; T. NOGAMI, *Japanese Noh Plays: How to See Them* (1934), an explanation written for those not familiar with Nō theatre; N. PERI, *Le nô* (1944), a bibliographic introduction to works on Nō and a general account of Nō, with a close literal translation of 10 Nō and 11 *kyōgen* plays; P.G. O'NEILL, *A Guide to Nō* (1953), a general explanation of Nō, with summaries of more than 240 plays, *Early Nō Drama: Its Background, Character and Development, 1300–1450* (1958), a study of the origin of Nō, including accounts of *shirabyōshi, kusemai, dengaku*, and *sarugaku*; Z. TOKI, *Japanese Nō Plays* (1954), a fairly detailed explanation and bibliography; NIPPON GAKUJUTSU SHINKOKAI, *Japanese Noh Drama* (1960), a general introduction and translation of 10 plays, including *Takasago*; H. BOHNER, *Nô: Die einzelnen Nô* (1956), an easy introduction to some 240 plays arranged in the conventional division into 5 parts.

(Y.H.)

Nova Scotia

One of the four British colonies federated into the Dominion of Canada in 1867, Nova Scotia comprises the peninsula of Nova Scotia, Cape Breton Island (separated from the mainland to the southwest by the narrow Strait of Canso), and a few small adjacent islands. Along the 17-mile-wide Chignecto Isthmus, which seems to thrust the peninsula into the Atlantic Ocean, runs the province's only land boundary, with New Brunswick to the west. Two arms of the Gulf of St. Lawrence, Northumberland and Cabot straits, separate it, respectively, from Prince Edward Island on the north and Newfoundland on the northeast. To the east lies the Atlantic and to the southwest the Bay of Fundy. Nova Scotia's 21,425 square miles (55,490 square kilometres) held a population of nearly 800,000 in the 1971 census, about one-third of whom lived in the two metropolitan areas of Halifax, the capital, and Sydney.

Location and general character

Nova Scotia—with New Brunswick, Prince Edward Island, and Newfoundland—is classified as one of Canada's Maritime, or Atlantic, Provinces, and both its past and its present are tied closely to the maritime life of fishing, shipbuilding, and transatlantic shipping. It was the site, in 1605, of the first permanent North American settlement north of Florida, established by the French. Among the legends that pervade the province is that found in the tale *Evangeline* by the U.S. poet Henry Wadsworth Longfellow, a fictional account of the very real deportation of the inhabitants of French descent in 1755 by a fearful British governor. The province's contemporary life retains some of the feeling of 17th-century Acadie, or Acadia, the Micmac Indian name adopted by the French for the region before Scottish colonists implanted their own name Nova Scotia, Latin for New Scotland. For information on related topics, see the articles CANADA; CANADA, HISTORY OF, and NORTH AMERICA.

The history of Nova Scotia. *Vicissitudes of settlement.* As in many of the older Canadian provinces, the 17th and 18th centuries were characterized by the instabilities of colonization, of struggles for power originating in the rivalries between London and Paris, and of migratory and military pressure from the colonies to the south that were to become the United States. The territory passed back and forth between France and England until 1713, when the French, retaining Cape Breton Island and other areas, began construction of the powerful Louisbourg fortress. Halifax was founded in 1749 as a counterbalance and populated with some 4,000 British settlers. In the 1750s the French Acadians, who refused to swear allegiance to the English crown, were expelled.

Decades as a pawn of empire

Offers of free land attracted immigrants not only from the British Isles but also from New England; it was the latter group that helped to secure the first elected assembly. By the time of the American Revolution, New Englanders comprised roughly one-half of Nova Scotia's population; though they tried to remain neutral, four delegates attended the Continental Congress in Philadelphia. During the Revolution and after, some 35,000 Loyalists added to the number of immigrants from the south. Meanwhile, Prince Edward Island had split off from Nova Scotia in 1769, and New Brunswick followed in 1784; Cape Breton Island finally was reunited with Nova Scotia in 1820. In 1848 Nova Scotia became the first British colony to exercise the prerogative of government responsible to the people through their elected representatives. There was both economic and political opposition to the proposed confederation with Ontario, Quebec, and New Brunswick, but union was carried out in 1867.

From colony to province. As a separate British colony, Nova Scotia prospered from its forestry, fisheries, and shipbuilding for the first two-thirds of the 19th century. Under the Reciprocity Treaty of 1854, the north–south flow of commerce and Nova Scotia's normal market and supply source in New England seemed secure. Nova Scotia benefitted further from the increased demand for duty-free natural products during the Civil War. The Canadian tariffs on manufactured goods, the alleged pro-Confederate sentiments of Britain and Canada, and the protectionist pressures of some states in the victorious North led to the nonrenewal of reciprocity in 1866 and the levying of further protective tariffs by both nations. Canada also lost the preferential treatment by Britain previously granted under colonial policy. Railways, meanwhile, were changing continental traffic patterns, and Canada and the U.S became rivals for the east–west inland trade. Canada's tariffs, intended both to protect its own manufactures and to assist in financing its railways from coastal ports to central Canada, were supposed to assure the province of year-round commerce, but the pioneers had not counted on the costs of transportation to the distant population centres. Nova Scotia's prospects were further set back as iron steamers replaced the wooden sailing vessels that had been the pride and chief industry of the province. The steamers usually bypassed Nova Scotian ports for those in the U.S., and even the lucrative trade with the West Indies dwindled. Further damage to the economy of the Maritime Provinces occurred with the opening of the western provinces to settlement, a step that tended to drain the older regions of some of the more vigorous elements of the population and to deprive them of investment capital and other resources, which now were directed toward the new lands.

By the early 1970s such conditions had been alleviated, and industry had made significant beginnings to add to the exploitation of natural resources. Nova Scotia, however, continued to lag well behind the more developed provinces and the Canadian averages in virtually all economic indicators.

The natural and human landscape. *Surface features.* Nova Scotia's five upland regions reach a maximum height of more than 1,700 feet (520 metres) above sea level in the Cape Breton Highlands. The most important lowlands lie along the Bay of Fundy and the Minas Basin, in the southwest and along the Northumberland Strait. Many of the tens of thousands of acres of marshland in the west created by the tremendously high tides of the Bay of Fundy have been turned to agricultural use by dikes, which were begun in the early 18th century by the Acadian French.

Lakes and rivers

Over 3,000 lakes and hundreds of short rivers and streams either have been impounded by, or have cut through, the irregularly high and low landscapes. The best known of the lakes, Bras d'Or on Cape Breton Island, is saline but tideless because of three short channels connecting it with the Atlantic. Many intruding heads of land make the lake's 360 square miles a geographical complexity; it is surrounded by steep hills and valleys.

Climate. The southwestern and southern shores of Nova Scotia have both milder and wetter climates than the rest of the province, with frost-free days numbering about 140 along the Bay of Fundy and 160 around Yarmouth, compared with 100 in the interior. Rainfall varies from 55 inches (1,400 millimetres) in the south, where

fog may occur on as many as 90 days, to 40 inches elsewhere.

Vegetation and animal life. Productive forests occupy about 75 percent of the land area. Only one-quarter of the woodland is held as crown land, some of which is used for public parks and reserves. Softwoods are by far the most numerous, led by various species of balsam, spruce, hemlock, and pine; birch and maple make up most of the hardwoods. Hunting—especially for deer and moose and for partridge, pheasant and duck among the birds—is a favourite outdoor sport for Nova Scotians and tourists. For anglers, trout and salmon are among the most popular freshwater fish; for deep-sea fishermen, Wedgeport is the scene of annual international tuna matches.

The human imprint. Subsistence living on family farms has been characteristic of Nova Scotia agriculture. In 1970 improved farmland and pastureland amounted to some 500 square miles and farm woodlands 1,700 square miles. Along the coasts many families living on marginal land have combined farming with fishing for lobsters and ground fish; in other areas farming and lumbering are often combined. Mixed farming and dairying are carried out in the fertile lands of Guysborough and other eastern counties, as well as on the Chignecto Isthmus and along the Fundy shore. Though apples and pears are also found in other sections, the Annapolis Valley orchards, dating back three centuries, are the most productive. With greater efficiency in land use and increased mechanization, both the number of farms and total acreage under cultivation have declined, but individual farms are becoming somewhat larger, production is increasing, and income improving. About 6 percent of the people live on farms, and another 35 percent are nonfarm rural dwellers.

Patterns of rural and urban employment

In the 19th century, as steamships replaced schooners, the outports dwindled. Coal and steel and textile industries drew workers to Sydney and Halifax, but some smaller towns continued enterprises like boatbuilding, woodworking, and food processing. Recently, internal population movement has been toward the Halifax–Dartmouth area, in which there is employment in the shipyards, naval dockyard, construction industries, and assembly plants. Sydney has been declining during the transition from steel and coal to other resources, but Port Hawkesbury has had a dramatic growth spurt due to the location of an oil refinery, pulp and paper plant, power generator, and other industries near the Strait of Canso.

The people of Nova Scotia. About one-eighth of Nova Scotia's people are descended from the Acadian French, who were allowed to return from exile after the British took French Canada in 1763. Most of the remaining seven-eighths are descended from settlers from the British Isles, both English and Scottish. Acadian communities are now located above Yarmouth and along St. Mary's Bay and on the west coast of Cape Breton Island. New England planters resettled the Minas Basin after 1755, and British Empire Loyalists founded Shelburne in 1783 and helped to populate Halifax. Scottish Highlanders, many victims of enclosure movement, tended to settle in the Northumberland Strait counties and Cape Breton Island. Also in the 18th century some Ulstermen established farms in the Truro and Onslow districts, and later Irish came in small family groups and settled in and near Halifax. German immigrants in the 1750s founded the seaport of Lunenburg, and others came a century later. Other minorities include small numbers of Dutch, Italian, and Hungarian who have immigrated since World War II. More significant are the Negro communities near Halifax and Shelburne, dating from the days of the West Indian slave trade and particularly the Loyalist influx after the American Revolution. Over 3,000 Micmac Indians live on reservations.

Ethnic components of the population

In 1970 over one-third of the people were Roman Catholics; among the Protestant denominations the United Church of Canada is the largest, followed by Anglicans and Baptists. Since 1960 immigration into Nova Scotia has not been numerically significant, but emigration of both the undertrained and the well educated has continued. The population has grown mainly by natural increase: the 1969 birth rate was 17.8 per 1,000 and the death rate 8.7, compared to national statistics of 17.6 for births and 7.3 for deaths. In 1971 the population of nearly 789,000 represented an increase of 2 percent in five years.

The province's economy. *Components.* With only about 5 percent of the province's land suitable for crops, agriculture has focussed mainly on livestock and poultry production and various dairying activities—though the marshlands offered the environment necessary for such crops as blueberries.

Primary resources and secondary industries

Fisheries continue as a major activity, with the major part of the catch comprising shellfish, haddock, and cod. The largest fish-processing plant in North America was opened in Lunenburg in 1964.

The generally small-scale forestry operations, about one-half of which are directed to pulp production, have been supplemented by large plants built in recent years by major companies. The pulp industry remains a significant factor within the economy, although increasing concern is evident over environmental pollution caused by plant operations. Coal seams on Cape Breton Island, mostly underwater, are becoming thin, and the future of even new mines like the one at Lingan is uncertain; but coal remains the province's single most valuable mineral resource. By 1969, however, the combined value of industrial minerals was greater than coal; among the nonmetallics, salt production was increasing, and Nova Scotia claimed the world's largest barite deposits and 85 percent of Canada's gypsum production. Such metals as silver, copper, lead, and zinc came primarily from one mine in Walton. Offshore explorations for oil, particularly those near Sable Island in 1972, gave rise to hopes for a new industry and to debates between federal and provincial officials over jurisdictional rights to such deposits. The value of export trade, particularly from the port of Halifax, with its new container piers, increased by 113 percent between 1957 and 1970.

Nova Scotia entered the 1970s with relatively unstable conditions prevailing among its industries, most of which were small-scale operations engaged in processing the province's primary products. Industrial Estates Ltd., a crown corporation, had a fair record in locating new industries by advancing up to 100 percent of factory and equipment costs, but a similar corporation met with little success in developing substitute industries on Cape Breton Island to facilitate the phasing out of the coal industry. In addition, heavy-water plants, built with considerable fanfare in the 1960s, were slow to begin output, and the provincial government showed some reluctance to join the development activities sponsored by the federal government. Bright spots included continuing growth of petrochemical facilities around the Strait of Canso and new facilities for containerized shipping in Halifax harbour. The Nova Scotia Research Foundation served to keep small industries in touch with advances in technology and science, in addition to engaging in research in pure science.

Transportation. Cape Breton Island was connected with the mainland by causeway in 1955; the other main highways, some 5,000 miles of which are paved, are located around the perimeter of both the mainland and Cape Breton Island, with a direct road as well from Truro to Halifax. Ferry services operate to ports in New Brunswick, Prince Edward Island, and the state of Maine. Both major Canadian railroads serve the province by land or ferry, and five airports handle commercial and private aviation.

Administration and social conditions. *Government and politics.* Nova Scotia's governmental structure is similar to that of all Canadian provinces. A lieutenant governor appointed by the federal government serves as representative of the crown, whereas the premier is the leader of the party in power in the provincial parliament and selects his cabinet from among his colleagues. The judiciary is federally appointed, while government employees

are recruited through a nonpolitical civil service. Local government is divided among the three cities, Halifax, Dartmouth, and Sydney; 39 towns; and 24 rural municipalities. Provincial income is derived from two virtually equal sources: the federal government and various provincial taxes and fees.

Since confederation there have been two major political parties in Nova Scotia, the dominant Liberals and the Conservatives (now the Progressive Conservatives); the New Democratic Party has had little success. In recent times provincial election campaigns have centred upon efficiency in government, economic expansion, and the quality and style of leadership. The proposed union of the Maritime Provinces is a matter of continuing study. As a proving ground for politicians, Nova Scotia had by 1970 produced three federal prime ministers and many other national figures.

Education. Nova Scotia has over 800 public schools, 13 vocational schools, 11 institutions of higher learning, and two institutes of technology. The nondenominational system of compulsory, free public education, dating from 1811, is in the final stages of consolidating the many rural school districts into fewer, larger, and more efficient districts. Considerable emphasis is placed on commercial and vocational training at the secondary as well as the college level. Dalhousie University (founded 1818) in Halifax is perhaps the best known of the institutions of higher learning, though a number of Roman Catholic and other private institutions offer diversified curricula. Agriculture, social work, and teacher training are among the specialized areas served by specific institutions. St. Francis Xavier University, in Antigonish, has attracted international interest in its adult-education programs, which have fostered social leadership through self-help projects such as cooperatives and credit unions.

Health and welfare. Since 1969 Nova Scotia has participated in the federal Medical Care Insurance Program. The premium is paid from the province's general revenues, and insured services include all medically necessary procedures by medical practitioners plus some oral surgery procedures in hospitals. The province also provides facilities for mental health, dental care, tuberculosis control, and other public-health services. In the early 1970s there were 54 hospitals. Nova Scotia's welfare services are similar to those in other provinces, covering old age assistance, allowances for the blind and disabled, social assistance, social development, child welfare and adoptions, and service to unmarried mothers and to delinquents. The Department of Public Welfare operates institutions for mentally retarded children and for delinquent youth and supervises a number of children's aid societies and other private institutions.

The standards of living approximate those of the nation generally, but costs have been higher and average incomes substantially lower. Nova Scotia is below the Canadian norm in employment, industrial wages, and housing facilities. Personal income per capita as a percentage of Canada improved from 69.7 percent in 1951 to 80.3 percent 20 years later, although the actual dollar gap widened from $349 to $610.

Cultural life and institutions. Its position as a peninsula and island thrust into the Atlantic has kept Nova Scotia somewhat removed from the mainstream of Canadian life, centred in Ontario and Quebec. The greater ease of travel and the growing impact of the nationwide communications media in recent decades have brought the province closer, making a modern way of life typical of the province and eroding in some measure the features of traditional Nova Scotian life—but many areas of the mainland peninsula and of Cape Breton Island retain an unretouched image of the past.

Preservation of traditional cultures

Scottish culture is particularly vigorous in Nova Scotia. Pictou County alone has several bagpipe bands, while St. Francis Xavier University offers courses in Celtic studies, and the Gaelic College in St. Ann's, Cape Breton, fosters piping, singing, dancing, and handicrafts. Clan gatherings take place annually at the Gaelic College to celebrate the Gaelic Mod, a festival of Highland folk arts. Both fine arts and crafts of various ethnic origins are exhibited at the arts festival held annually on the campus of Acadia University. Each summer individual and community arts are taught at Tatamagouche, and an Acadian festival is held at Clare.

Halifax is the centre of repertory drama, and the several universities have their own theatre productions. Groups of resident and visiting painters are found in such south shore communities as Mahone Bay, Lunenburg, and Peggy's Cove. Woodcarving, pottery making, hooking, and weaving are found throughout the province.

French-language programs are provided on five of the 30 radio stations and one of the 12 television stations. There is one weekly newspaper in French and 27 weeklies and 6 dailies in English.

Among the most impressive historic sites are the restored fortress of Louisbourg, Champlain's habitation at Port Royal, and the Halifax citadel. Also of historical interest are the Alexander Graham Bell National Historic Park with a museum at Baddeck and the Grand Pré memorial in the Annapolis Valley. In 1970 these sites were visited by over 1,250,000 persons, and Cape Breton Highlands National Park attracted over 700,000 tourists. A system of provincial parks and campsites is under continual development, and private accommodations and recreational facilities are rapidly increasing; but assessments of the economic value of tourism are imprecise.

Prospects. In the early 1970s, though there was little assurance of eliminating disparities that existed since confederation, Nova Scotia had realistic prospects for keeping pace with the economic growth rate of the nation. Such progress would be related to developing additional power resources, modernizing steel facilities, revitalizing the coal industry, attracting additional light and heavy industry, expanding tourism, and improving the marketing of the province's farm, fishery, and forest products. Federal and provincial efforts are coordinated under the Department of Regional Economic Expansion, the Atlantic Development Board, the Cape Breton Development Corporation, and other crown agencies. It was hoped that these objectives could be achieved without seriously eroding those cultural values and aspirations for an abundant life that have characterized the peoples of Nova Scotia for over three centuries.

BIBLIOGRAPHY. There is no comprehensive history of Nova Scotia, but numerous books provide reliable information on various phases or aspects of the province's development. In addition to the full-length works, there is a considerable body of literature in the *Collections* of the Nova Scotia Historical Society (1878–), *The Dalhousie Review*, and the *Canadian Historical Review* (both quarterly). Selected readings are J. MURRAY BECK, *The Government of Nova Scotia* (1957), a comprehensive work tracing the development of governmental institutions; W.R. BIRD, *This is Nova Scotia* (1950), a popular travelogue that captures contemporary viewpoints, conditions, and links with the past; J.B. BREBNER, *The Neutral Yankees of Nova Scotia* (1937, reprinted 1969) and *New England's Outpost: Acadia Before the Conquest of Canada* (1927, reprinted 1965), among the standard historical works on colonial Nova Scotia; ARTHUR G. DOUGHTY, *The Acadian Exiles: A Chronicle of the Land of Evangeline* (1916, reprinted 1964), a concise, more objective account than most of this controversial subject; W.S. MACNUTT, *The Atlantic Provinces: The Emergence of Colonial Society, 1712–1857* (1965), a historical synthesis of Nova Scotia in a regional context; and G.A. RAWLYK (ed.), *Historical Essays on the Atlantic Provinces* (1967), an anthology, regional in scope, with several individual essays elucidating important features of the economic, intellectual, and cultural life of Nova Scotia.

(B.O'G.)

Novel

The novel is a genre of fiction, and fiction may be defined as the art or craft of contriving, through the written word, representations of human life that instruct or divert or both. The various forms that fiction may take are best

seen less as a number of separate categories than as a continuum or, more accurately, a cline, with some such brief form as the anecdote at one end of the scale and the longest conceivable novel at the other. When any piece of fiction is long enough to constitute a whole book, as opposed to a mere part of a book, then it may be said to have achieved novelhood. But this state admits of its own quantitative categories, so that a relatively brief novel may be termed a novella (or, if the insubstantiality of the content matches its brevity, a novelette), and a very long novel may overflow the banks of a single volume and become a *roman fleuve,* or river novel. Length is very much one of the dimensions of the genre.

The term novel is a truncation of the Italian word *novella* (from the plural of Latin *novellus,* a late variant of *novus,* meaning "new"), so that what is now, in most languages, a diminutive denotes historically the parent form. The *novella* was a kind of enlarged anecdote like those to be found in the 14th-century Italian classic Boccaccio's *Decameron,* each of which exemplifies the etymology well enough. The stories are little new things, novelties, freshly minted diversions, toys; they are not reworkings of known fables or myths, and they are lacking in weight and moral earnestness. It is to be noted that, despite the high example of novelists of the most profound seriousness, such as Tolstoy, Henry James, and Virginia Woolf, the term novel still, in some quarters, carries overtones of lightness and frivolity. And it is possible to descry a tendency to triviality in the form itself. The ode or symphony seems to possess an inner mechanism that protects it from aesthetic or moral corruption, but the novel can descend to shameful commercial depths of sentimentality or pornography. It is the purpose of this article to consider the novel not solely in terms of great art but also as an all-purpose medium catering for all the strata of literacy.

The novel and the epic compared

Such early ancient Roman fiction as Petronius' *Satyricon* of the 1st century AD and Lucius Apuleius' *Golden Ass* of the 2nd century contain many of the popular elements that distinguish the novel from its nobler born relative the epic poem. In the fictional works, the medium is prose, the events described are unheroic, the settings are streets and taverns, not battlefields and palaces. There is more low fornication than princely combat; the gods do not move the action; the dialogue is homely rather than aristocratic. It was, in fact, out of the need to find—in the period of Roman decline—a literary form that was anti-epic in both substance and language that the first prose fiction of Europe seems to have been conceived. The most memorable character in Petronius is a *nouveau riche* vulgarian; the hero of Lucius Apuleius is turned into a donkey; nothing less epic can well be imagined.

The medieval chivalric romance (from a popular Latin word, probably *Romanice,* meaning written in the vernacular, not in traditional Latin) restored a kind of epic view of man—though now as heroic Christian, not heroic pagan. At the same time, it bequeathed its name to the later genre of continental literature, the novel, which is known in French as *roman,* in Italian as *romanzo,* etc. (The English term romance, however, carries a pejorative connotation.) But that later genre achieved its first great flowering in Spain at the beginning of the 17th century in an antichivalric comic masterpiece—the *Don Quixote* of Cervantes, which, on a larger scale than the *Satyricon* or *The Golden Ass,* contains many of the elements that have been expected from prose fiction ever since. Novels have heroes, but not in any classical or medieval sense. As for the novelist, he must, in the words of the contemporary British-American W.H. Auden,

> Become the whole of boredom, subject to
> Vulgar complaints like love, among the Just
> Be just, among the Filthy filthy too,
> And in his own weak person, if he can,
> Must suffer dully all the wrongs of Man.

The novel attempts to assume those burdens of life that have no place in the epic poem and to see man as unheroic, unredeemed, imperfect, even absurd. This is why there is room among its practitioners for writers of hard-boiled detective thrillers such as the contemporary American Mickey Spillane or of sentimental melodramas such as the prolific 19th-century English novelist Mrs. Henry Wood, but not for one of the unremitting elevation of outlook of a John Milton.

This article analyzes the characteristics of the novel, discusses the various uses it serves, enumerates the artistic styles it reflects and its subgenres, and treats it historically wherever it has been a significant literary form. Finally, there are discussions of its social and economic aspects and its prospects and a concluding discussion of its evaluation and study. The article is outlined as follows:

The reader may also be interested in the analogous treatment of a comparable genre in the article SHORT STORY. Critical approaches are discussed in the articles ARTS, CRITICISM OF THE; and LITERARY CRITICISM. The place of the novel within the literatures of the world is treated in articles such as LITERATURE, WESTERN; and LITERATURE, EAST ASIAN; and in articles on the arts of various peoples, such as SOUTH ASIAN PEOPLES, ARTS OF; AFRICAN PEOPLES, ARTS OF.

I. Characteristics of the novel

ELEMENTS OF THE NOVEL

Plot. The novel is propelled through its hundred or thousand pages by a device known as the story or plot. This is frequently conceived by the novelist in very simple terms, a mere nucleus, a jotting on an old envelope: for example, Charles Dickens' *Christmas Carol* (1843) might have been conceived as "a misanthrope is reformed through certain magical visitations on Christmas Eve"; or Jane Austen's *Pride and Prejudice* (1813) as "a young couple destined to be married have first to overcome the barriers of pride and prejudice"; or Fyodor Dostoyevsky's *Crime and Punishment* (1866) as "a young man commits a crime and is slowly pursued in the direction of his punishment." The detailed working out of the nuclear idea requires much ingenuity, since the plot of one novel is expected to be somewhat different from that of another, and there are very few basic human situations for the novelist to draw upon. The dramatist may take his plot ready-made from fiction or biography—a form of theft sanctioned by Shakespeare—but the novelist has to produce what look like novelties.

Original and borrowed plots

The example of Shakespeare is a reminder that the ability to create an interesting plot, or even any plot at all, is not a prerequisite of the imaginative writer's craft. At the lowest level of fiction, plot need be no more than a string of stock devices for arousing stock responses of concern and excitement in the reader. The reader's interest may be captured at the outset by the promise of conflicts or mysteries or frustrations that will eventually be resolved, and he will gladly—so strong is his desire to be moved or entertained—suspend criticism of even the most trite modes of resolution. In the least sophisticated fiction, the knots to be untied are stringently physical, and the denouement often comes in a sort of triumphant violence. Serious fiction prefers its plots to be based on psychological situations, and its climaxes come in new states of awareness—chiefly self-knowledge—on the parts of the major characters.

Melodramatic plots, plots dependent on coincidence or improbability, are sometimes found in even the most elevated fiction; E.M. Forster's *Howards End* (1910) is an example of a classic British novel with such a plot. But the novelist is always faced with the problem of whether

it is more important to represent the formlessness of real life (in which there are no beginnings and no ends and very few simple motives for action) or to construct an artifact as well balanced and economical as a table or chair; since he is an artist, the claims of art, or artifice, frequently prevail. But there are ways of constructing novels in which plot may play a desultory part or no part at all. The traditional picaresque novel—a novel with a rogue as its central character—like Alain Lesage's *Gil Blas* (1715) or Henry Fielding's *Tom Jones* (1749), depends for movement on a succession of chance incidents. In the works of Virginia Woolf, the consciousness of the characters, bounded by some poetic or symbolic device, sometimes provides all the fictional material. Marcel Proust's great *roman fleuve, À la recherche du temps perdu* (1913–27; *Remembrance of Things Past*), has a metaphysical framework derived from the time theories of the philosopher Henri Bergson, and it moves toward a moment of truth that is intended to be literally a revelation of the nature of reality. Strictly, any scheme will do to hold a novel together—raw action, the hidden syllogism of the mystery story, prolonged solipsist contemplation—so long as the actualities or potentialities of human life are credibly expressed, with a consequent sense of illumination, or some lesser mode of artistic satisfaction, on the part of the reader.

Character. The inferior novelist tends to be preoccupied with plot; to the superior novelist the convolutions of the human personality, under the stress of artfully selected experience, are the chief fascination. Without character it was once accepted that there could be no fiction. In the period since World War II, the creators of what has come to be called the French *nouveau roman* (*i.e.*, new novel) have deliberately demoted the human element, claiming the right of objects and processes to the writer's and reader's prior attention. Thus, in books termed *chosiste* (literally "thing-ist"), they make the furniture of a room more important than its human incumbents. This may be seen as a transitory protest against the long predominance of character in the novel, but, even on the popular level, there have been indications that readers can be held by things as much as by characters. Henry James could be vague in *The Ambassadors* (1903) about the provenance of his chief character's wealth; if he wrote today he would have to give his readers a tour around the factory or estate. The popularity of much undistinguished but popular fiction has nothing to do with its wooden characters; it is machines, procedures, organizations that draw the reader. The success of Ian Fleming's British spy stories in the 1960s had much to do with their hero, James Bond's car, gun, and preferred way of mixing a martini.

Objects as "characters"

But the true novelists remain creators of characters—prehuman, such as those in William Golding's *Inheritors* (1955); animal, as in Henry Williamson's *Tarka the Otter* (1927) or Jack London's *Call of the Wild* (1903); caricatures, as in much of Dickens; or complex and unpredictable entities, as in Tolstoy, Dostoyevsky, or Henry James. The reader may be prepared to tolerate the most wanton-seeming stylistic tricks and formal difficulties because of the intense interest of the central characters in novels as diverse as James Joyce's *Ulysses* (1922) and *Finnegans Wake* (1939) and Laurence Sterne's *Tristram Shandy* (1760–67).

Characters as symbols

It is the task of literary critics to create a value hierarchy of fictional character, placing the complexity of the Shakespearean view of man—as found in the novels of Tolstoy and Joseph Conrad—above creations that may be no more than simple personifications of some single characteristic, like some of those by Dickens. It frequently happens, however, that the common reader prefers surface simplicity—easily memorable cartoon figures like Dickens' never-despairing Mr. Micawber and devious Uriah Heep—to that wider view of personality, in which character seems to engulf the reader, subscribed to by the great novelists of France and Russia. The whole nature of human identity remains in doubt, and writers who voice that doubt—like the French exponents of the *nouveau roman* Alain Robbe-Grillet and Nathalie Sarraute, as well as many others—are in effect rejecting a purely romantic view of character. This view imposed the author's image of himself—the only human image he properly possessed—on the rest of the human world. For the unsophisticated reader of fiction, any created personage with a firm position in time–space and the most superficial parcel of behavioral (or even sartorial) attributes will be taken for a character. Though the critics may regard it as heretical, this tendency to accept a character is in conformity with the usages of real life. The average person has at least a suspicion of his own complexity and inconsistency of makeup, but he sees the rest of the world as composed of much simpler entities. The result is that novels whose characters are created out of the author's own introspection are frequently rejected as not "true to life." But both the higher and the lower orders of novel readers might agree in condemning a lack of memorability in the personages of a work of fiction, a failure on the part of the author to seem to add to the reader's stock of remembered friends and acquaintances. Characters that seem, on recollection, to have a life outside the bounds of the books that contain them are usually the ones that earn their creators the most regard. Depth of psychological penetration, the ability to make a character real as oneself, seems to be no primary criterion of fictional talent.

Scene, or setting. The makeup and behaviour of fictional characters depend on their environment quite as much as on the personal dynamic with which their author endows them: indeed, in Émile Zola, environment is of overriding importance, since he believed it determined character. The entire action of a novel is frequently determined by the locale in which it is set. Thus, Gustave Flaubert's *Madame Bovary* (1857) could hardly have been placed in Paris, because the tragic life and death of the heroine have a great deal to do with the circumscriptions of her provincial milieu. But it sometimes happens that the main locale of a novel assumes an importance in the reader's imagination comparable to that of the characters and yet somehow separable from them. Wessex is a giant brooding presence in Thomas Hardy's novels, whose human characters would probably not behave much differently if they were set in some other rural locality of England. The popularity of Sir Walter Scott's "Waverley" novels is due in part to their evocation of a romantic Scotland. Setting may be the prime consideration of some readers, who can be drawn to Conrad because he depicts life at sea or in the East Indies; they may be less interested in the complexity of human relationships that he presents.

Regional novels

The regional novel is a recognized species. The sequence of four novels that Hugh Walpole began with *Rogue Herries* (1930) was the result of his desire to do homage to the part of Cumberland, in England, where he had elected to live. The great Yoknapatawpha cycle of William Faulkner, a classic of 20th-century American literature set in an imaginary county in Mississippi, belongs to the category as much as the once-popular confections about Sussex that were written about the same time by the English novelist Sheila Kaye-Smith. Many novelists, however, gain a creative impetus from avoiding the same setting in book after book and deliberately seeking new locales. The English novelist Graham Greene apparently needed to visit a fresh scene in order to write a fresh novel. His ability to encapsulate the essence of an exotic setting in a single book is exemplified in *The Heart of the Matter* (1948); his contemporary Evelyn Waugh stated that the West Africa of that book replaced the true remembered West Africa of his own experience. Such power is not uncommon: the Yorkshire moors have been romanticized because Emily Brontë wrote of them in *Wuthering Heights* (1847), and literary tourists have visited Stoke-on-Trent, in northern England, because it comprises the "Five Towns" of Arnold Bennett's novels of the early 20th century. Others go to the Monterey, California, of John Steinbeck's novels in the expectation of experiencing a *frisson* added to the locality by an act of creative imagination. James Joyce, who remained inexhaustibly stimulated by Dublin, has exalted that city in a manner that even the guidebooks recognize.

The setting of a novel is not always drawn from a real-life locale. The literary artist sometimes prides himself on his ability to create the totality of his fiction—the setting as well as the characters and their actions. In the Russian expatriate Vladimir Nabokov's *Ada* (1969) there is an entirely new space–time continuum, and the English scholar J.R.R. Tolkien in his *Lord of the Rings* (1954–55) created an "alternative world" that appeals greatly to many who are dissatisfied with the existing one. The world of interplanetary travel was imaginatively created long before the first moon landing. The properties of the future envisaged by H.G. Wells's novels or by Aldous Huxley in *Brave New World* (1932) are still recognized in an age that those authors did not live to see. The composition of place can be a magical fictional gift.

Whatever the locale of his work, every true novelist is concerned with making a credible environment for his characters, and this really means a close attention to sense data—the immediacies of food and drink and colour—far more than abstractions like "nature" and "city." The London of Charles Dickens is as much incarnated in the smell of wood in lawyers' chambers as in the skyline and vistas of streets.

Narrative method and point of view. Where there is a story, there is a storyteller. Traditionally, the narrator of the epic and mock-epic alike acted as an intermediary between the characters and the reader; the method of Fielding is not very different from the method of Homer. Sometimes the narrator boldly imposed his own attitudes; always he assumed an omniscience that tended to reduce the characters to puppets and the action to a predetermined course with an end implicit in the beginning. Many novelists have been unhappy about a narrative method that seems to limit the free will of the characters, and innovations in fictional technique have mostly sought the objectivity of the drama, in which the characters appear to work out their own destinies without prompting from the author.

The epistolary method, most notably used by Samuel Richardson in *Pamela* (1740) and by Jean-Jacques Rousseau in *La nouvelle Héloïse* (1761), has the advantage of allowing the characters to tell the story in their own words, but it is hard to resist the uneasy feeling that a kind of divine editor is sorting and ordering the letters into his own pattern. The device of making the narrator also a character in the story has the disadvantage of limiting the material available for the narration, since the narrator-character can know only those events in which he participates. There can, of course, be a number of secondary narratives enclosed in the main narrative, and this device—though it sometimes looks artificial—has been used triumphantly by Conrad and, on a lesser scale, by W. Somerset Maugham. A, the main narrator, tells what he knows directly of the story and introduces what B and C and D have told him about the parts that he does not know.

Seeking the most objective narrative method of all, Ford Madox Ford used, in *The Good Soldier* (1915), the device of the storyteller who does not understand the story he is telling. This is the technique of the "unreliable observer." The reader, understanding better than the narrator, has the illusion of receiving the story directly. Joyce, in both his major novels, uses different narrators for the various chapters. Most of them are unreliable, and some of them approach the impersonality of a sort of disembodied parody. In *Ulysses*, for example, an episode set in a maternity hospital is told through the medium of a parodic history of English prose style. But, more often than not, the sheer ingenuity of Joyce's techniques draws attention to the manipulator in the shadows. The reader is aware of the author's cleverness where he should be aware only of the characters and their actions. The author is least noticeable when he is employing the stream of consciousness device, by which the inchoate thoughts and feelings of a character are presented in interior monologue—apparently unedited and sometimes deliberately near-unintelligible. It is because this technique seems to draw fiction into the psychoanalyst's consulting room (presenting the raw material of either art or science, but certainly not art itself), however, that Joyce felt impelled to impose the shaping devices referred to above. Joyce, more than any novelist, sought total objectivity of narration technique but ended as the most subjective and idiosyncratic of stylists.

The presence of the author

The problem of a satisfactory narrative point of view is, in fact, nearly insoluble. The careful exclusion of comment, the limitation of vocabulary to a sort of reader's lowest common denominator, the paring of style to the absolute minimum—these puritanical devices work well for an Ernest Hemingway (who, like Joyce, remains, nevertheless, a highly idiosyncratic stylist) but not for a novelist who believes that, like poetry, his art should be able to draw on the richness of word play, allusion, and symbol. For even the most experienced novelist, each new work represents a struggle with the unconquerable task of reconciling all-inclusion with self-exclusion. It is noteworthy that Cervantes, in *Don Quixote*, and Nabokov, in *Lolita* (1955), join hands across four centuries in finding most satisfactory the device of the fictitious editor who presents a manuscript story for which he disclaims responsibility. But this highly useful method presupposes in the true author a scholarly, or pedantic, faculty not usually associated with novelists.

Scope, or dimension. No novel can theoretically be too long, but if it is too short it ceases to be a novel. It may or may not be accidental that the novels most highly regarded by the world are of considerable length—Cervantes' *Don Quixote*, Dostoyevsky's *Brothers Karamazov*, Tolstoy's *War and Peace*, Dickens' *David Copperfield*, Proust's *À la recherche du temps perdu*, and so on. On the other hand, since World War II, brevity has been regarded as a virtue in works like the later novels of the Irish absurdist author Samuel Beckett and the *ficciones* of the Argentine Jorge Luis Borges, and it is only an aesthetic based on bulk that would diminish the achievement of Ronald Firbank's short novels of the post-World War I era or the Evelyn Waugh who wrote *The Loved One* (1948). It would seem that there are two ways of presenting human character—one, the brief way, through a significant episode in the life of a personage or group of personages; the other, which admits of limitless length, through the presentation of a large section of a life or lives, sometimes beginning with birth and ending in old age. The plays of Shakespeare show that a full delineation of character can be effected in a very brief compass, so that, for this aspect of the novel, length confers no special advantage. Length, however, is essential when the novelist attempts to present something bigger than character—when, in fact, he aims at the representation of a whole society or period of history.

No other cognate art form—neither the epic poem nor the drama nor the film—can match the resources of the novel when the artistic task is to bring to immediate, sensuous, passionate life the somewhat impersonal materials of the historian. *War and Peace* is the great triumphant example of the panoramic study of a whole society—that of early-19th-century Russia—which enlightens as the historians enlightens and yet also conveys directly the sensations and emotions of living through a period of cataclysmic change. In the 20th century, another Russian, Boris Pasternak, in his *Doctor Zhivago* (1957), expressed—though on a less than Tolstoyan scale—the personal immediacies of life during the Russian Revolution. Though of much less literary distinction than either of these two books, Margaret Mitchell's *Gone with the Wind* (1936) showed how the American Civil War could assume the distanced pathos, horror, and grandeur of any of the classic struggles of the Old World.

Needless to say, length and weighty subject matter are no guarantee in themselves of fictional greatness. Among American writers, for example, James Jones's celebration of the U.S. Army on the eve of World War II in *From Here to Eternity* (1951), though a very ambitious project, repels through indifferent writing and sentimental characterization; Norman Mailer's *Naked and the Dead* (1948), an equally ambitious military novel, succeeds much more because of a tautness, a concern with compression, and an astringent objectivity that Jones was

unable to match. Frequently the size of a novel is too great for its subject matter—as with Marguerite Young's *Miss MacIntosh, My Darling* (1965), reputedly the longest single-volume novel of the 20th century, John Barth's *Giles Goatboy* (1966), and John Fowles's *Magus* (1965). Diffuseness is the great danger in the long novel, and diffuseness can mean slack writing, emotional self-indulgence, sentimentality.

Even the long picaresque novel—which, in the hands of a Fielding or his contemporary Tobias Smollett, can rarely be accused of sentimentality—easily betrays itself into such acts of self-indulgence as the multiplication of incident for its own sake, the coy digression, the easygoing jogtrot pace that subdues the sense of urgency that should lie in all fiction. If Tostoy's *War and Peace* is a greater novel than Fielding's *Tom Jones* or Dickens' *David Copperfield*, it is not because its theme is nobler, or more pathetic, or more significant historically; it is because Tolstoy brings to his panoramic drama the compression and urgency usually regarded as the monopolies of briefer fiction.

Works in several books

Sometimes the scope of a fictional concept demands a technical approach analogous to that of the symphony in music—the creation of a work in separate books, like symphonic movements, each of which is intelligible alone but whose greater intelligibility depends on the theme and characters that unify them. The French author Romain Rolland's *Jean-Christophe* (1904–12) sequence is, very appropriately since the hero is a musical composer, a work in four movements. Among works of English literature, Lawrence Durrell's *Alexandria Quartet* (1957–60) insists in its very title that it is a tetralogy rather than a single large entity divided into four volumes; the concept is "relativist" and attempts to look at the same events and characters from four different viewpoints. Anthony Powell's *Dance to the Music of Time*, a multivolume series of novels that began in 1951 (collected 1962), may be seen as a study of a segment of British society in which the chronological approach is eschewed, and events are brought together in one volume or another because of a kind of parachronic homogeneity. C.P. Snow's *Strangers and Brothers*, a comparable series that began in 1940 and continued to appear throughout the '50s and into the '60s, shows how a fictional concept can be realized only in the act of writing, since the publication of the earlier volumes antedates the historical events portrayed in later ones. In other words, the author could not know what the subject matter of the sequence would be until he was in sight of its end. Behind all these works lies the giant example of Proust's *roman fleuve*, whose length and scope were properly coterminous with the author's own life and emergent understanding of its pattern.

Myth, symbolism, significance. The novelist's conscious day-to-day preoccupation is the setting down of incident, the delineation of personality, the regulation of exposition, climax, and denouement. The aesthetic value of his work is frequently determined by subliminal forces that seem to operate independently of the writer, investing the properties of the surface story with a deeper significance. A novel will then come close to myth, its characters turning into symbols of permanent human states or impulses, particular incarnations of general truths perhaps only realized for the first time in the act of reading. The ability to perform a quixotic act anteceded *Don Quixote*, just as *bovarysme* existed before Flaubert found a name for it.

But the desire to give a work of fiction a significance beyond that of the mere story is frequently conscious and deliberate, indeed sometimes the primary aim. When a novel—like Joyce's *Ulysses* or John Updike's *Centaur* (1963) or Anthony Burgess' *Vision of Battlements* (1965)—is based on an existing classical myth, there is an intention of either ennobling a lowly subject matter, satirizing a debased set of values by referring them to a heroic age, or merely providing a basic structure to hold down a complex and, as it were, centrifugal picture of real life. Of *Ulysses*, Joyce said that his Homeric parallel (which is worked out in great and subtle detail) was a bridge across which to march his 18 episodes; after the march the bridge could be "blown skyhigh." But there is no doubt that, through the classical parallel, the account of an ordinary summer day in Dublin is given a richness, irony, and universality unattainable by any other means.

The mythic or symbolic intention of a novel may manifest itself less in structure than in details which, though they appear naturalistic, are really something more. The shattering of the eponymous golden bowl in Henry James's 1904 novel makes palpable, and hence truly symbolic, the collapse of a relationship. Even the choice of a character's name may be symbolic. Sammy Mountjoy, in William Golding's *Free Fall* (1959), has fallen from the grace of heaven, the mount of joy, by an act of volition that the title makes clear. The eponym of *Doctor Zhivago* is so called because his name, meaning "the living," carries powerful religious overtones. In the Russian version of the Gospel According to St. Luke, the angels ask the women who come to Christ's tomb: "Chto vy ischyote zhivago mezhdu myortvykh?"—"Why do you seek the living among the dead?" And his first name, Yuri, the Russian equivalent of George, has dragon-slaying connotations.

The symbolism of names

The symbol, the special significance at a subnarrative level, works best when it can fit without obtrusion into a context of naturalism. The optician's trade sign of a huge pair of spectacles in F. Scott Fitzgerald's *Great Gatsby* (1925) is acceptable as a piece of scenic detail, but an extra dimension is added to the tragedy of Gatsby, which is the tragedy of a whole epoch in American life, when it is taken also as a symbol of divine myopia. Similarly, a cinema poster in Malcolm Lowry's *Under the Volcano* (1947), advertising a horror film, can be read as naturalistic background, but it is evident that the author expects the illustrated fiend—a concert pianist whose grafted hands are those of a murderer—to be seen also as a symbol of Nazi infamy; the novel is set at the beginning of World War II, and the last desperate day of the hero, Geoffrey Firmin, stands also for the collapse of Western civilization.

There are symbolic novels whose infranarrative meaning cannot easily be stated, since it appears to subsist on an unconscious level. Herman Melville's *Moby Dick* (1851) is such a work, as is D.H. Lawrence's novella *St. Mawr* (1925), in which the significance of the horse is powerful and mysterious.

USES OF THE NOVEL

As an expression of an interpretation of life. Novels are not expected to be didactic, like tracts or morality plays; nevertheless, in varying degrees of implicitness, even the "purest" works of fictional art convey a philosophy of life. The novels of Jane Austen, designed primarily as superior entertainment, imply a desirable ordered existence, in which the comfortable decorum of an English rural family is disturbed only by a not too serious shortage of money, by love affairs that go temporarily wrong, and by the intrusion of self-centred stupidity. The good, if unrewarded for their goodness, suffer from no permanent injustice. Life is seen, not only in Jane Austen's novels but in the whole current of bourgeois Anglo-American fiction, as fundamentally reasonable and decent. When wrong is committed, it is usually punished, thus fulfilling Miss Prism's summation in Oscar Wilde's play *The Importance of Being Earnest* (1895), to the effect that in a novel the good characters end up happily and the bad characters unhappily: "that is why it is called fiction".

That kind of fiction called realistic, which has its origins in 19th-century France, chose the other side of the coin, showing that there was no justice in life and that the evil and the stupid must prevail. In the novels of Thomas Hardy there is a pessimism that may be taken as a corrective of bourgeois Panglossianism—the philosophy that everything happens for the best, satirized in Voltaire's *Candide* (1759)—since the universe is presented as almost impossibly malevolent. This tradition is regarded as morbid, and it has been deliberately ignored by most popular novelists. The "Catholic" novelists such as François Mauriac in France, Graham Greene in England, and

others—see life as mysterious, full of wrong and evil and injustice inexplicable by human canons but necessarily acceptable in terms of the plans of an inscrutable God. Between the period of realistic pessimism, which had much to do with the agnosticism and determinism of 19th-century science, and the introduction of theological evil into the novel, writers such as H.G. Wells attempted to create a fiction based on optimistic liberalism. As a reaction, there was the depiction of "natural man" in the novels of D.H. Lawrence and Ernest Hemingway.

For the most part, the view of life common to American and European fiction since World War II posits the existence of evil—whether theological or of that brand discovered by the French Existentialists, particularly Jean-Paul Sartre—and assumes that man is imperfect and life possibly absurd. The fiction of the Communist states is based on a very different assumption, one that seems naïve and old-fashioned in its collective optimism to readers in the disillusioned democracies. It is to be noted that in the Soviet Union aesthetic evaluation of fiction has been replaced by ideological judgment. The works of the popular British writer A.J. Cronin, since they seem to depict personal tragedy as an emanation of capitalistic infamy, are rated higher than those of Conrad, James, and their peers. The novel as an art form stands or falls by its capacity to express a view of life acceptable to the Soviet authorities.

The Communist interpretation

As entertainment or escape. In a period that takes for granted that the written word should be "committed"—to the exposure of social wrong or the propagation of progressive ideologies—novelists who seek merely to take the reader out of his dull or oppressive daily life are not highly regarded, except by that reading public that has never expected a book to be anything more than a diversion. Nevertheless, the provision of laughter and dreams has been for many centuries a legitimate literary occupation. It can be condemned by serious devotees of literature only if it falsifies life through oversimplification and tends to corrupt its readers into belief that reality is as the author presents it. The novelettes once beloved of mill girls and domestic servants, in which the beggar maid was elevated to queendom by a king of high finance, were a mere narcotic, a sort of enervating opium of the oppressed; the encouragement of such subliterature might well be one of the devices of social oppression. Adventure stories and spy novels may have a healthy enough astringency, and the very preposterousness of some adventures can be a safeguard against any impressionable young reader's neglecting the claims of real life to dream of becoming a secret agent. The subject matter of some humorous novels—such as the effete British aristocracy created by P.G. Wodehouse, which is no longer in existence if it ever was—can never be identified with a real human society; the dream is accepted as a dream. The same may be said of Evelyn Waugh's early novels—such as *Decline and Fall* (1928) and *Vile Bodies* (1930)—but these are raised above mere entertainment by touching, almost incidentally, on real human issues (the relation of the innocent to a circumambient malevolence is a persistent theme in all Waugh's writing).

Any reader of fiction has a right to an occasional escape from the dullness or misery of his existence, but he has the critical duty of finding the best modes of escape—in the most efficiently engineered detective or adventure stories, in humor that is more than sentimental buffoonery, in dreams of love that are not mere pornography. The fiction of entertainment and escape frequently sets itself higher literary standards than novels with a profound social or philosophical purpose. Books like James Buchan's *Thirty-nine Steps* (1915), Graham Greene's *Travels with My Aunt* (1969), Dashiell Hammett's *Maltese Falcon* (1930), and Raymond Chandler's *Big Sleep* (1939) are distinguished pieces of writing that, while diverting and enthralling, keep a hold on the realities of human character. Ultimately, all good fiction is entertainment, and if it instructs or enlightens it does so best through enchanting the reader.

As propaganda. The desire to make the reader initiate certain acts—social, religious, or political—is the essence of all propaganda, and, though it does not always accord well with art, the propagandist purpose has often found its way into novels whose prime value is an aesthetic one. The *Nicholas Nickleby* (1839) of Charles Dickens attacked the abuses of schools to some purpose, as his *Oliver Twist* (1838) drew attention to the horrors of poorhouses and his *Bleak House* (1853) to the abuses of the law of chancery. The weakness of propaganda in fiction is that it loses its value when the wrongs it exposes are righted, so that the more successful a propagandist novel is, the briefer the life it can be expected to enjoy. The genius of Dickens lay in his ability to transcend merely topical issues through the vitality with which he presented them, so that his contemporary disclosures take on a timeless human validity—chiefly through the power of their drama, character, and rhetoric.

The ephemeral quality of propaganda

The purely propagandist novel—which Dickens was incapable of writing—quickly becomes dated. The "social" novels of H.G. Wells, which propounded a rational mode of life and even blueprinted utopias, were very quickly exploded by the conviction of man's irredeemable irrationality that World War I initiated and World War II corroborated, a conviction the author himself came to share toward the end of his life. But the early scientific romances of Wells remain vital and are seen to have been prophetic. Most of the fiction of the Soviet Union, which either glorifies the regime or refrains from criticizing it, is dull and unreal, and the same may be said of Communist fiction elsewhere. Propaganda too frequently ignores man as a totality, concentrating on him aspectively—in terms of politics or sectarian religion. When a didactic attack on a system, as in Harriet Beecher Stowe's attack on slavery in the United States in *Uncle Tom's Cabin* (1852), seems to go beyond mere propaganda, it is because the writer makes the reader aware of wrongs and injustices that are woven into the permanent human fabric. The reader's response may be a modification of his own sensibility, not an immediate desire for action, and this is one of the legitimate effects of serious fiction. The propagandist Dickens calls for the immediate righting of wrongs, but the novelist Dickens says, mainly through implication, that all men—not just schoolmasters and state hirelings—should become more humane. If it is possible to speak of art as possessing a teaching purpose, this is perhaps its only lesson.

As reportage. The division in the novelist's mind is between his view of his art as a contrivance, like a Fabergé watch, and his view of it as a record of real life. The versatile English writer Daniel Defoe, on the evidence of such novels as his *Journal of the Plague Year* (1722), a re-creation of the London plague of 1665, believed that art or contrivance had the lesser claim and proceeded to present his account of events of which he had had no direct experience in the form of plain journalistic reportage. This book, like his *Robinson Crusoe* (1719) and *Moll Flanders* (1722), is more contrived and cunning than appears, and the hurried, unshaped narrative is the product of careful preparation and selective ordering. His example, which could have been a very fruitful one, was not much followed until the 20th century, when the events of the real world became more terrifying and marvellous than anything the novelist could invent, and seemed to ask for that full imaginative treatment that only the novelist's craft can give.

In contemporary American literature, John Hersey's *Hiroshima* (1946), though it recorded the actual results of the nuclear attack on the Japanese city in 1945, did so in terms of human immediacies, not scientific or demographic abstractions, and this approach is essentially novelistic. Truman Capote's *In Cold Blood* (1965) took the facts of a multiple murder in the Midwest of the United States and presented them with the force, reality, tone, and (occasionally) overintense writing that distinguish his genuine fiction. Norman Mailer, in *The Armies of the Night* (1968), recorded, in great personal detail but in a third-person narration, his part in a citizens' protest march on Washington, D.C. It would seem that Mailer's talent lies in his ability to merge the art of fiction and the craft of reportage, and his *Of a Fire on the Moon* (1970),

which deals with the American lunar project, reads like an episode in an emergent *roman fleuve* of which Mailer is the central character in a story made up of true public events.

The presentation of factual material as art is the purpose of such thinly disguised biographies as Somerset Maugham's *Moon and Sixpence* (1919), undisguised biographies fleshed out with supposition and imagination like Helen Waddell's *Peter Abelard* (1933), and a great number of autobiographies served up—out of fear of libel or of dullness—as novels. Conversely, invented material may take on the lineaments of journalistic actuality through the employment of a Defoe technique of flat understatement. This is the way of such science fiction as Michael Crichton's *Andromeda Strain* (1969), which uses sketch maps, computer projections, and simulated typewritten reports.

As an agent of change in the language and thought of a culture. Novelists, being neither poets nor philosophers, rarely originate modes of thinking and expression. Poets such as Chaucer and Shakespeare have had much to do with the making of the English language, and Byron was responsible for the articulation of the new romantic sensibility in it in the early 19th century. Books like the Bible, Karl Marx's *Kapital*, and Adolf Hitler's *Mein Kampf* may underlie permanent or transient cultures, but it is hard to find, except in the early Romantic period, a novelist capable of arousing new attitudes to life (as opposed to aspects of the social order) and forging the vocabulary of such attitudes.

Sentiment

With the 18th-century precursors of Romanticism—notably Richardson, Sterne, and Rousseau—the notion of sentiment entered the European consciousness. Rousseau's *Nouvelle Héloïse* (1761) fired a new attitude toward love—more highly emotional than ever before—as his *Émile* (1762) changed educated views on how to bring up children. The romantic wave in Germany, with Goethe's *Sorrows of Werther* (1774) and the works of Jean-Paul Richter a generation later, similarly aroused modes of feeling that rejected the rational constraints of the 18th century. Nor can the influence of Sir Walter Scott's novels be neglected, both on Europe and on the American South (where Mark Twain thought it had had a deplorable effect). With Scott came new forms of regional sentiment, based on a romantic reading of history.

It is rarely, however, that a novelist makes a profound mark on a national language, as opposed to a regional dialect (to which, by using it for a literary end, he may impart a fresh dignity). It is conceivable that Alessandro Manzoni's *I promessi sposi* (1825–27; *The Betrothed*), often called the greatest modern Italian novel, gave 19th-century Italian intellectuals some notion of a viable modern prose style in an Italian that might be termed "national," but even this is a large claim. Günter Grass, in post-Hitler Germany, sought to revivify a language that had been corrupted by the Nazis; he threw whole dictionaries at his readers, in the hope that new freedom, fantasy, and exactness in the use of words might influence the publicists, politicians, and teachers in the direction of a new liberalism of thought and expression.

It is difficult to say whether the French Existentialists, such as Sartre and Albert Camus, have influenced their age primarily through their fiction or their philosophical writings. Certainly, Sartre's early novel *Nausea* (1938) established unforgettable images of the key terms of his philosophy, which has haunted a whole generation, as Camus's novel *The Stranger* (1942) created for all time the lineaments of "Existential man." In the same way, the English writer George Orwell's *Nineteen Eighty-four* (1949) incarnated brilliantly the nature of the political choices that are open to 20th-century humanity, and, with terms like "Big Brother" (*i.e.*, the leader of an authoritarian state) and "doublethink" (belief in contradictory ideas simultaneously), modified the political vocabulary. But no novelist's influence can compare to that of the poet's, who can give a language a soul and define, as Shakespeare and Dante did, the scope of a culture.

As an expression of the spirit of its age. The novelist, like the poet, can make the inchoate thoughts and feelings of a society come to articulation through the exact and imaginative use of language and symbol. In this sense, his work seems to precede the diffusion of new ideas and attitudes and to be the agent of change. But it is hard to draw a line between this function and that of expressing an existing climate of sensibility. Usually the nature of a historical period—that spirit known in German as the *Zeitgeist*—can be understood only in long retrospect, and it is then that the novelist can provide its best summation. The sickness of the Germany that produced Hitler had to wait some time for fictional diagnosis in such works as Thomas Mann's *Doctor Faustus* (1947) and, later, Günter Grass's *Tin Drum* (1959). Evelyn Waugh waited several years before beginning, in the trilogy *Sword of Honour*, to depict that moral decline of English society that started to manifest itself in World War II, the conduct of which was both a cause and a symptom of the decay of traditional notions of honour and justice.

The lag between events and their use in novels

The novel can certainly be used as a tool for the better understanding of a departed age. The period following World War I had been caught forever in Hemingway's *Sun Also Rises* (1926; called *Fiesta* in England), F. Scott Fitzgerald's novels and short stories about the so-called Jazz Age, the *Antic Hay* (1923) and *Point Counter Point* (1928) of Aldous Huxley, and D.H. Lawrence's *Aaron's Rod* (1922) and *Kangaroo* (1923). The spirit of the English 18th century, during which social, political, and religious ideas associated with rising middle classes conflicted with the old Anglican Tory rigidities, is better understood through reading Smollett and Fielding than by taking the cerebral elegance of Pope and his followers as the typical expression of the period.

Similarly, the unrest and bewilderment of the young in the period after World War II still speak in novels like J.D. Salinger's *Catcher in the Rye* (1951) and Kingsley Amis' *Lucky Jim* (1954). It is notable that with novels like these—and the beat-generation books of Jack Kerouac; the American-Jewish novels of Saul Bellow, Bernard Malamud, and Philip Roth; and the Negro novels of Ralph Ellison and James Baldwin—it is a segmented spirit that is expressed, the spirit of an age group, social group, or racial group, and not the spirit of an entire society in a particular phase of history. But probably a *Zeitgeist* has always been the emanation of a minority, the majority being generally silent. The 20th century seems, from this point of view, to be richer in vocal minorities than any other period in history.

As a creator of life styles and an arbiter of taste. Novels have been known to influence, though perhaps not very greatly, modes of social behaviour and even, among the very impressionable, conceptions of personal identity. But more young men have seen themselves as Hamlet or Childe Harold than as Julien Sorel, the protagonist of Stendhal's novel *The Red and the Black* (1830), or the sorrowing Werther. Richardson's novel may popularize Pamela, or Galsworthy's *Forsyte Saga* (1906–22) Jon, as a baptismal name, but it rarely makes a deeper impression on the mode of life of literate families. On the other hand, the capacity of Oscar Wilde's *Picture of Dorian Gray* (1891) to influence young men in the direction of sybaritic amorality, or of D.H. Lawrence's *Lady Chatterley's Lover* (1928) to engender a freer attitude to sex, has never been assessed adequately. With the lower middle class reading public, the effect of devouring *The Forsyte Saga* was to engender genteelisms —cucumber sandwiches for tea, supper renamed dinner —rather than to learn that book's sombre lesson about the decline of the old class structure. Similarly, the ladies who read Scott in the early 19th century were led to barbarous ornaments and tastefully arranged folk songs.

Fiction has to be translated into one of the dramatic media—stage, film, or television—before it can begin to exert a large influence. *Tom Jones* as a film in 1963 modified table manners and coiffures and gave American visitors to Great Britain a new (and probably false) set of expectations. The stoic heroes of Hemingway, given to drink, fights, boats, and monosyllables, became influential only when they were transferred to the screen. They engendered other, lesser heroes—incorruptible pri-

Fiction in other media

vate detectives, partisans brave under interrogation—who in their turn have influenced the impressionable young when seeking an identity. Ian Fleming's James Bond led to a small revolution in martini ordering. But all these influences are a matter of minor poses, and such poses are most readily available in fiction easily adapted to the mass media—which means lesser fiction. Proust, though he recorded French patrician society with painful fidelity, had little influence on it, and it is hard to think of Henry James disturbing the universe even fractionally. Films and television programs dictate taste and behaviour more than the novel ever could.

STYLE

Romanticism. The Romantic movement in European literature is usually associated with those social and philosophical trends that prepared the way for the French Revolution, which began in 1789. The somewhat subjective, anti-rational, emotional currents of romanticism transformed intellectual life in the revolutionary and Napoleonic periods and remained potent for a great part of the 19th century. In the novel, the romantic approach to life was prepared in the "sentimental" works of Richardson and Sterne and attained its first major fulfillment in the novels of Rousseau. Sir Walter Scott, in his historical novels, turned the past into a great stage for the enactment of events motivated by idealism, chivalry, and strong emotional impulse, using an artificially archaic language full of remote and magical charm. The exceptional soul—poet, patriot, idealist, madman—took the place of dully reasonable fictional heroes, such as Tom Jones, and sumptuous and mysterious settings ousted the plain town and countryside of 18th-century novels.

The romantic novel must be seen primarily as a historical phenomenon, but the romantic style and spirit, once they had been brought into being, remained powerful and attractive enough to sustain a whole subspecies of fiction. The cheapest love story can be traced back to the example of Charlotte Brontë's *Jane Eyre* (1847), or even Rousseau's earlier *Nouvelle Héloïse*. Similarly, best-selling historical novels, even those devoid of literary merit, can find their progenitor in Scott, and science fiction in Mary Shelleys' *Frankenstein* (1818), a romantic novel subtitled *The Modern Prometheus,* as well as in Jules Verne and H.G. Wells. The aim of romantic fiction is less to present a true picture of life than to arouse the emotions through a depiction of strong passions, or to fire the imagination with exotic, terrifying, or wonderful scenes and events. When it is condemned by critics, it is because it seems to falsify both life and language; the pseudo-poetical enters the dialogue and *récit* alike, and humanity is seen in only one of its aspects—that of feeling untempered with reason.

If such early romantic works as those of Scott and of the Goethe of *The Sorrows of Werther* have long lost their original impact, the romantic spirit still registers power and truth in the works of the Brontës—particularly in Emily Brontë's *Wuthering Heights,* in which the poetry is genuine and the strange instinctual world totally convincing. Twentieth-century romantic fiction records few masterpieces. Writers like Daphne du Maurier, the author of *Jamaica Inn* (1936), *Rebecca* (1938), and many others, are dismissed as mere purveyors of easy dreams. It is no more possible in the 20th century to revive the original romantic élan in literature than it is to compose music in the style of Beethoven. Despite the attempts of Lawrence Durrell to achieve a kind of decadent romantic spirit in his *Alexandria Quartet,* the strong erotic feeling, the exotic setting, the atmosphere of poetic hallucination, the pain, perversion, and elemental force seem to be contrivances, however well they fulfill the original romantic prescription.

Decline of the romantic novel

Realism. Certain major novelists of the 19th century, particularly in France, reacted against romanticism by eliminating from their work those "softer" qualities—tenderness, idealism, chivalric passion, and the like—which seemed to them to hide the stark realities of life in a dreamlike haze. In Gustave Flaubert's works there are such romantic properties—his novel *Salammbô* (1862), for instance, is a sumptuous representation of a remote pagan past—but they are there only to be punctured with realistic irony. On one level, his *Madame Bovary* may be taken as a kind of parable of the punishment that fate metes out to the romantic dreamer, and it is the more telling because Flaubert recognized a strong romantic vein in himself: "Madame Bovary, c'est moi" ("Madame Bovary is myself"). Stendhal and Balzac, on the other hand, admit no dreams and present life in a grim nakedness without poetic drapery.

Balzac's mammoth fictional work—the 20-year succession of novels and stories he published under the collective title *The Human Comedy*—and Stendhal's novels of the same period, *The Red and the Black* (1830) and *The Charterhouse of Parma* (1839), spare the reader nothing of those baser instincts in man and society that militate against, and eventually conquer, many human aspirations. Rejecting romanticism so energetically, however, they swing to an extreme that makes "realism" a synonym for unrelenting pessimism. Little comes right for the just or the weak, and base human nature is unqualified by even a modicum of good. But there is a kind of affirmative richness and energy about both writers that seems to belie their pessimistic thesis.

The pessimistic view

In England, George Eliot in her novel *Middlemarch* (1871–72) viewed human life grimly, with close attention to the squalor and penury of rural life. If "nature" in works by romantic poets like Wordsworth connoted a kind of divine benevolence, only the "red in tooth and claw" aspect was permitted to be seen in the novels of the realists. George Eliot does not accept any notion of Divine Providence, whether Christian or pantheistic, but her work is instinct with a powerful moral concern: her characters never sink into a deterministic morass of hopelessness, since they have free will, or the illusion of it. With Thomas Hardy, who may be termed the last of the great 19th-century novelists, the determinism is all-pervasive, and his final novel, *Jude the Obscure* (1896), represents the limit of pessimism. Behind him one is aware of the new science, initiated by the biologists Charles Darwin and T.H. Huxley, which displaces man as a free being, capable of choice, by a view of him as the product of blind mechanistic forces over which he has little control.

Realism in this sense has been a continuing impulse in the 20th-century novel, but few writers would go so far as Hardy in positing man's near-total impotence in a hostile universe, with the gods killing human creatures for their sport. Realism in the Existentialist fiction of 20th-century France, for instance, makes man not merely wretched but absurd, yet it does not diminish his power of self-realization through choice and action. Realism has frequently been put in the service of a reforming design, which implies a qualified optimism. War novels, novels about the sufferings of the oppressed (in prison, ghetto, totalitarian state), studies of human degradation that are bitter cries against man-made systems—in all of these the realistic approach is unavoidable, and realistic detail goes much further than anything in the first realists. But there is a difference in the quality of the anger the reader feels when reading the end of Hardy's *Tess of the D'Urbervilles* (1891) and that generated by Upton Sinclair's *Jungle* (1906) or Erich Maria Remarque's *All Quiet on the Western Front* (1929). In Hardy's novel, pessimistic determinism, reducing human character to pain, frustration, and impotent anger, was—paradoxically—appropriate to an age that knew no major cataclysms or oppressions. The novels of Sinclair and Remarque reflect the 20th century, which saw the origin of all wrong in the human will, and set on a program of diagnosis and reform.

Realism as reform

Naturalism. The naturalistic novel is a development out of realism, and it is, again, in France that its first practitioners are to be found, with Émile Zola leading. It is difficult to separate the two categories, but naturalism seems characterized not only by a pessimistic determinism but also by a more thoroughgoing attention to the physical and biological aspects of human existence. Man is less a soul aspiring upward to its divine source than a

product of natural forces, as well as genetic and social influences, and the novelist's task is to present the physical essence of man and his environment. The taste of Balzac's and Stendhal's audiences was not easily able to accommodate itself to utter frankness about the basic processes of life, and the naturalists had to struggle against prejudice, and often censorship, before their literary candour was able to prevail. The 20th century takes the naturalistic approach for granted, but it is more concerned with a technique of presentation than with the somewhat mechanistic philosophy of Zola and his followers.

Naturalism received an impetus after World War I, when novelists felt they had a duty to depict the filth, suffering, and degradation of the soldier's life, without euphemism or circumlocution. Joyce's *Ulysses*, when it appeared in 1922, was the first novel to seek to justify total physical candour in terms of its artistic, as opposed to moral, aim—which was to depict with almost scientific objectivity every aspect of an ordinary urban day. Though Joyce had read Zola, he seems to invoke the spirit of a very much earlier naturalistic writer—the ribald French author of the 16th century, François Rabelais—and this is in keeping with the Catholic tradition that Joyce represents. Zola, of course, was an atheist.

It would have been a sin against his aesthetic canons for Joyce to have shown Leopold Bloom—the protagonist of *Ulysses*—eating breakfast or taking a bath and yet not defecating or masturbating. The technique of the interior monologue, which presented the unedited flow of a character's unspoken thought and emotion, also called for the utmost frankness in dealing with natural functions and urges. Joyce, it is now recognized, had no prurient or scatological intention; his concern was with showing life as it is (without any of the didactic purpose of Zola), and this entailed the presentation of lust, perversion, and blasphemy as much as any of the traditionally acceptable human functions.

The naturalistic novelists have had their social and legal problems—obscenity indictments, confiscation, emasculation by timid publishers—but the cause was ultimately won, at least in Great Britain and the United States, where there are few limits placed on the contemporary novelist's proclaimed right to be true to nature. In comparison with much contemporary fiction the pioneer work of Zola seems positively reticent.

Impressionism. The desire to present life with frank objectivity led certain early-20th-century novelists to question the validity of long-accepted narrative conventions. If truth was the novelist's aim, then the tradition of the omniscient narrator would have to go, to be replaced by one in which a fallible, partially ignorant character—one involved in the story and hence himself subject to the objective or naturalistic approach—recounted what he saw and heard. But the Impressionist painters of late-19th-century France had proclaimed a revision of the whole seeing process: they distinguished between what the observer assumed he was observing and what he actually observed. That cerebral editing which turned visual data into objects of geometric solidity had no place in Impressionist painting; the visible world became less definite, more fluid, resolving into light and colour.

The influence of the painters

The German novelists Thomas Mann and Hermann Hesse, moving from the realist tradition, which concentrated on closely notated detail in the exterior world, sought the lightness and clarity of a more elliptical style, and were proclaimed Impressionists. But in England Ford Madox Ford went much further in breaking down the imagined rigidities of the space–time continuum, liquidating step-by-step temporal progression and making the visual world shimmer, dissolve, reconstitute itself. In Ford's tetralogy *Parade's End* (1924–28), the reader moves freely within the time continuum, as if it were spatial, and the total picture is perceived through an accumulation of fragmentary impressions. Ford's masterpiece, *The Good Soldier* (1915), pushes the technique to its limit: the narrator tells his story with no special dispensation to see or understand more than a fallible being can, and, in his reminiscences, he fragments whole sequences of events as he ranges freely through time (such freedom had traditionally been regarded as a weakness, a symptom of the disease of inattention).

In the approach to dialogue manifested in a book that Ford wrote jointly with Conrad—*The Inheritors* (1901)—a particular aspect of literary impressionism may be seen whose suggestiveness has been ignored by other modern novelists. As the brain imposes its own logical patterns on the phenomena of the visual world, so it is given to editing into clarity and conciseness the halting utterances of real-life speech; the characters of most novels are impossibly articulate. Ford and Conrad attempted to present speech as it is actually spoken, with many of the meaningful solidities implied rather than stated. The result is sometimes exasperating, but only as real-life conversation frequently is.

The interior monologue, which similarly resists editing, may be regarded as a development of this technique. To show pre-articulatory thought, feeling, and sensuous perception unordered into a rational or "literary" sequence is an impressionistic device that, beginning in Edouard Dujardin's minor novel *Les Lauriers sont coupés* (1887; *We'll to the Woods No More*), served fiction of high importance, from Dorothy Richardson, Joyce, and Virginia Woolf to William Faulkner and Samuel Beckett.

Novelists like Ronald Firbank and Evelyn Waugh (who studied painting and was a competent draftsman) learned, in a more general sense, how to follow the examples of the Impressionist and Postimpressionist painters in their fiction. A spare brilliance of observation, like those paintings in which a whole scene is suggested through carefully selected points of colour, replaced that careful delineation of a whole face, or inventorying of a whole room, that had been the way of Balzac and other realists. In four or five brief lines of dialogue Waugh can convey as much as the 19th-century novelists did in as many pages.

Expressionism. Expressionism was a German movement that found its most congenial media in painting and drama. The artist's aim was to express, or convey the essence of, a particular theme, to the exclusion of such secondary considerations as fidelity to real life. The typical Expressionist play, by Bertolt Brecht, for example, concerns itself with a social or political idea that is hurled at the audience through every possible stage device—symbols, music, cinematic insertions, choral speech, dance. Human character is less important than the idea of humanity, and probability of action in the old realist sense is the least of the dramatist's concerns. The emotional atmosphere is high-pitched, even ecstatic, and the tone is more appropriate to propaganda than to art. Expressionistic technique, as the plays of Brecht prove, was an admirable means of conveying a Communist program, and it was in the service of such a program that John Dos Passos, in the trilogy of novels *U.S.A.* (1937), used literary devices analogous to the dramatic ones of Brecht—headlines, tabloid biographies, popular songs, lyric soliloquies, and the like.

Expressionism as alienation

But the Austrian Franz Kafka, the greatest of the Expressionist novelists, sought to convey what may crudely be termed man's alienation from his world in terms that admit of no political interpretation. Joseph K., the hero of Kafka's novel *The Trial* (1925), is accused of a nameless crime, he seeks to arm himself with the apparatus of a defense, and he is finally executed—stabbed with the utmost courtesy by two men in a lonely place. The hallucinatory atmosphere of that novel, as also of his novel *The Castle* (1926), is appropriate to nightmare, and indeed Kafka's work has been taken by many as an imaginative forecast of the nightmare through which Europe was compelled to live during the Hitler regime. But its significance is more subtle and universal; one of the elements is original sin and another filial guilt. In the story *The Metamorphosis* (1915) a young man changes into an enormous insect, and the nightmare of alienation can go no further.

Kafka's influence has been considerable. Perhaps his most distinguished follower is the English writer Rex Warner, whose *Wild Goose Chase* (1937) and *Aero-*

drome (1941) use fantasy, symbol, and improbable action for an end that is both Marxist and Freudian; the filial guilt, however, seems to be taken directly from Kafka, with an innocent hero caught in a monstrously oppressive web that is both the totalitarian state and paternal tyranny. More recently, the American expatriate writer William Burroughs has developed his own expressionistic techniques in *The Naked Lunch* (1959), which is concerned with the alienation from society of the drug addict. His later novels *Nova Express* (1964) and *The Ticket That Exploded* (1962) use obscene fantasy to present a kind of metaphysical struggle between free spirit and enslaved flesh, evidently an extrapolation of the earlier drug theme. Burroughs is a didactic novelist, and didacticism functions best in a fictional ambience that rejects the complexities of character and real-life action.

Avant-gardism. Many innovations in fiction can be classified under headings already considered. Even so revolutionary a work as Joyce's *Finnegans Wake* represents an attempt to show the true nature of a dream; this can be regarded as a kind of Impressionism pushed so far that it looks like Surrealism. The brief novels of Samuel Beckett (which, as they aim to demonstrate the inadequacy of language to express the human condition, become progressively more brief) seem to have a kind of Expressionist derivation, since everything in them is subordinated to a central image of man as a totally deprived creature, resentful of a God he does not believe in. The French anti-novel, dethroning man as a primary concern of fiction, perhaps represents the only true break with traditional technique that the 20th-century novel has seen.

Technical innovation

Dissatisfaction not only with the content of the traditional novel but with the manner in which readers have been schooled to approach it has led the contemporary French novelist Michel Butor, in *Mobile*, to present his material in the form of a small encyclopaedia, so that the reader finds his directions obliquely, through an alphabetic taxonomy and not through the logic of sequential events. Nabokov, in *Pale Fire* (1962), gives the reader a poem of 999 lines and critical apparatus assembled by a madman; again the old sense of direction (beginning at the beginning and going on to the end) has been liquidated, yet *Pale Fire* is a true and highly intelligible novel. In England, B.S. Johnson published similar "false-directional" novels, though the influence of Sterne makes them seem accessible, even cozily traditional. One of Johnson's books is marketed as a bundle of disjunct chapters—which may thus be dealt aleatorially and read in any order—packed in a box, not a binding.

Available avant-garde techniques are innumerable, though not all of them are salable. There is the device of counterpointing a main narrative with a story in footnotes, which eventually rises like water and floods the other. A novel has been written, though not published, in which the words are set (rather like the mouse's tail or tale in *Alice in Wonderland*) to represent graphically the physical objects in the narrative. Burroughs has experimented with a tricolumnar technique, in which three parallel narratives demand the reader's attention. But the writers like Borges and Nabokov go beyond mere technical innovation: they ask for a reconsideration of the very essence of fiction. In one of his *ficciones*, Borges strips from the reader even the final illusion that he is reading a story, for the story is made to dissolve, the artist evidently losing faith in his own artifact. Novels, as both Borges and Nabokov show, can turn into poems or philosophical essays, but they cannot, while remaining literature, turn into compositions disclaiming all interest in the world of feeling, thought, and sense. The novelist can do anything he pleases with his art so long as he interprets, or even just presents, a world that the reader recognizes as existing, or capable of existing, or capable of being dreamed of as existing.

TYPES OF NOVEL

Historical. For the hack novelist, to whom speedy output is more important than art, thought, and originality, history provides ready-made plots and characters. A novel on Alexander the Great or Joan of Arc can be a flimsy and superficial as any schoolgirl romance. But historical themes, to which may be added prehistoric or mythical ones, have inspired the greatest novelists, as Tolstoy's *War and Peace* and Stendhal's *Charterhouse of Parma* reveal. In the 20th century, distinguished historical novels such as Arthur Koestler's *The Gladiators* (1939), Robert Graves's *I, Claudius* (1934), Zoé Oldenbourg's *Destiny of Fire* (1960), and Mary Renault's *The King Must Die* (1958) exemplify an important function of the fictional imagination—to interpret remote events in human and particular terms, to transform documentary fact, with the assistance of imaginative conjecture, into immediate sensuous and emotional experience.

There is a kind of historical novel, little more than a charade, which frequently has a popular appeal because of a common belief that the past is richer, bloodier, and more erotic than the present. Such novels, which include such immensely popular works as those of Georgette Heyer, or Baroness Orczy's Scarlet Pimpernel stories in England in the early 20th century, and *Forever Amber* (1944) by Kathleen Winsor in the United States, may use the trappings of history but, because there is no real assimilation of the past into the imagination, the result must be a mere costume ball. On the other hand, the American novelist John Barth showed in *The Sot-Weed Factor* (1960) that mock historical scholarship—preposterous events served up with parodic pomposity—could constitute a viable, and not necessarily farcical, approach to the past. Barth's history is cheerfully suspect, but his sense of historical perspective is genuine.

It is in the technical conservatism of most European historical novels that the serious student of fiction finds cause to relegate the category to a secondary place. Few practitioners of the form seem prepared to learn from any writer later than Scott, though Virginia Woolf—in *Orlando* (1928) and *Between the Acts* (1941)—made bold attempts to squeeze vast tracts of historical time into a small space and thus make them as fictionally manageable as the events of a single day. And John Dos Passos' *U.S.A.*, which can be taken as a historical study of a phase in America's development, is a reminder that experiment is not incompatible with the sweep and amplitude that great historical themes can bring to the novel.

Picaresque. In Spain, the novel about the rogue or *pícaro* was a recognized form, and such English novels as Defoe's *Jonathan Wild* (1725) and *The Fortunate Mistress* (1724) can be regarded as picaresque in the etymological sense. But the term has come to connote as much the episodic nature of the original species as the dynamic of roguery. Fielding's *Tom Jones*, whose hero is a bastard, amoral, and very nearly gallows-meat, has been called picaresque, and the *Pickwick Papers* of Dickens—whose eponym is a respectable and even childishly ingenuous scholar—can be accommodated in the category. The requirements for a picaresque novel are apparently length, loosely linked episodes almost complete in themselves, intrigue, fights, amorous adventure, and such optional items as stories within the main narrative, songs, poems, or moral homilies. Perhaps inevitably, with such a structure or lack of it, the driving force must come from a wild or roguish rejection of the settled bourgeois life, a desire for the open road, with adventures in inn bedrooms and meetings with questionable wanderers. In the modern period, Saul Bellow's *Adventures of Augie March* (1953) and Jack Kerouac's *Dharma Bums* (1959) have something of the right episodic, wandering, free, questing character. But in an age that lacks the unquestioning acceptance of traditional morality against which the old picaresque heroes played out their villainous lives, it is not easy to revive the *novela picaresca* as the anonymous author of *Lazarillo de Tormes* (1554) conceived it, or as such lesser Spanish writers of the beginning of the 17th century as Mateo Alemán, Vicente Espinel, and Luis Vélez de Guevara developed it. The modern criminal wars with the police rather than with society, and his career is one of closed and narrow techniques, not compatible with the gay abandon of the true *pícaro*.

Sentimental. The term sentimental, in its mid-18th-century usage, signified refined or elevated feeling, and it

is in this sense that it must be understood in Laurence Sterne's *Sentimental Journey* (1768). Richardson's *Pamela* (1740) and Rousseau's *Nouvelle Héloïse* (1761) are sentimental in that they exhibit a passionate attachment between the sexes that rises above the merely physical. The vogue of the sentimental love novel was one of the features of the Romantic movement, and the form maintained a certain moving dignity despite a tendency to excessive emotional posturing. The germs of mawkishness are clearly present in Sterne's *Tristram Shandy* (1760–67), though offset by a diluted Rabelaisianism and a certain cerebral quality. The debasement by which the term sentimental came to denote a self-indulgence in superficial emotions occurred in the Victorian era, under the influence of sanctimony, religiosity, and a large commercial demand for bourgeois fiction. Sentimental novels of the 19th and 20th centuries are characterized by an invertebrate emotionalism and a deliberately lachrymal appeal. Neither Dickens nor Thackeray was immune to the temptations of sentimentality—as is instanced by their treatment of deathbed scenes. The reported death of Tiny Tim in *A Christmas Carol* (1843) is an example of Dickens' ability to provoke two tearful responses from the one situation—one of sorrow at a young death, the other of relief at the discovery that the death never occurred. Despite such patches of emotional excess, Dickens cannot really be termed a sentimental novelist. Such a designation must be reserved for writers like Mrs. Henry Wood, the author of *East Lynne* (1861). That the sentimental novel is capable of appeal even in the Atomic Age is shown by the success of *Love Story* (1970), by Erich Segal. That this is the work of a Yale professor of classics seems to indicate either that not even intellectuals disdain sentimental appeal or that tear-jerking is a process to be indulged in coldly and even cynically. Stock emotions are always easily aroused through stock devices, but both the aim and the technique are generally eschewed by serious writers.

Gothic. The first Gothic fiction appeared with works like Horace Walpole's *Castle of Otranto* (1765) and Matthew Gregory Lewis' *Monk* (1796), which countered 18th-century "rationalism" with scenes of mystery, horror, and wonder. Gothic (the spelling "Gothick" better conveys the contemporary flavour) was a designation derived from architecture, and it carried—in opposition to the Italianate style of neoclassical building more appropriate to the Augustan Age—connotations of rough and primitive grandeur. The atmosphere of a Gothic novel was expected to be dark, tempestuous, ghostly, full of madness, outrage, superstition, and the spirit of revenge. Mary Shelley's *Frankenstein*, which maintains its original popularity and even notoriety, has in overplus the traditional Gothic ingredients, with its weird God-defying experiments, its eldritch shrieks, and, above all, its monster. Edgar Allan Poe developed the Gothic style brilliantly in the United States, and he has been a considerable influence. A good deal of early science fiction, like H.G. Wells's *Island of Doctor Moreau* (1896), seems to spring out of the Gothic movement, and the Gothic atmosphere has been seriously cultivated in England in the later novels of Iris Murdoch and in the Gormenghast sequence beginning in 1946 of Mervyn Peake. It is noteworthy that Gothic fiction has always been approached in a spirit of deliberate suspension of the normal canons of taste. Like a circus trick, a piece of Gothic fiction asks to be considered as ingenious entertainment; the pity and terror are not aspects of a cathartic process but transient emotions to be, somewhat perversely, enjoyed for their own sake.

Ingredients of the typical Gothic novel

Psychological. The psychological novel first appeared in 17th-century France, with Madame de La Fayette's *Princesse de Clèves* (1678), and the category was consolidated by works like the Abbé Prévost's *Manon Lescaut* (1731) in the century following. More primitive fiction had been characterized by a proliferation of action and incidental characters; the psychological novel limited itself to a few characters whose motives for action could be examined and analyzed. In England, the psychological novel did not appear until the Victorian era, when George Eliot became its first great exponent. It has been assumed since then that the serious novelist's prime concern is the workings of the human mind, and hence much of the greatest fiction must be termed psychological. Dostoyevsky's *Crime and Punishment* deals less with the ethical significance of a murder than with the soul of the murderer; Flaubert's interest in Emma Bovary has less to do with the consequences of her mode of life in terms of nemesic logic than with the patterns of her mind; in *Anna Karenina*, Tolstoy presents a large-scale obsessive study of feminine psychology that is almost excruciating in its relentless probing. The novels of Henry James are psychological in that the crucial events occur in the souls of the protagonists, and it was perhaps James more than any serious novelist before or since who convinced frivolous novel-readers that the "psychological approach" guarantees a lack of action and excitement.

The theories of Sigmund Freud are credited as the source of the psychoanalytical novel. Freud was anticipated, however, by Shakespeare (in, for example, his treatment of Lady Macbeth's somnambulistic guilt). Two 20th-century novelists of great psychological insight—Joyce and Nabokov—professed a disdain for Freud. To write a novel with close attention to the Freudian or Jungian techniques of analysis does not necessarily produce new prodigies of psychological revelation; Oedipus and Electra complexes have become commonplaces of superficial novels and films. The great disclosures about human motivation have been achieved more by the intuition and introspection of novelists and dramatists than by the more systematic work of the clinicians.

The movement away from Freud

The novel of manners. To make fiction out of the observation of social behaviour is sometimes regarded as less worthy than to produce novels that excavate the human mind. And yet the social gestures known as manners, however superficial they appear to be, are indices of a collective soul and merit the close attention of the novelist and reader alike. The works of Jane Austen concern themselves almost exclusively with the social surface of a fairly narrow world, and yet she has never been accused of a lack of profundity. A society in which behaviour is codified, language restricted to impersonal formulas, and the expression of feeling muted, is the province of the novel of manners, and such fiction may be produced as readily in the 20th century as in the era of Fanny Burney or Jane Austen. Such novels as Evelyn Waugh's *Handful of Dust* (1934) depend on the exact notation of the manners of a closed society, and personal tragedies are a mere temporary disturbance of collective order. Even Waugh's trilogy *Sword of Honour* is as much concerned with the minutiae of surface behaviour in an army, a very closed society, as with the causes for which that army fights. H.H. Munro ("Saki"), in *The Unbearable Bassington* (1912), an exquisite novel of manners, says more of the nature of Edwardian society than many a more earnest work. It is conceivable that one of the novelist's duties to posterity is to inform it of the surface quality of the society that produced him; the great psychological profundities are eternal, manners are ephemeral and have to be caught. Finally, the novel of manners may be taken as an artistic symbol of a social order that feels itself to be secure.

Epistolary. The novels of Samuel Richardson arose out of his pedagogic vocation, which arose out of his trade of printer—the compilation of manuals of letter-writing technique for young ladies. His age regarded letter writing as an art on which could be expended the literary care appropriate to the essay or to fiction, and, for Richardson, the creation of epistolary novels entailed a mere step from the actual world into that of the imagination. His *Pamela* (1740) and *Clarissa* (1748) won phenomenal success and were imitated all over Europe, and the epistolary novel—with its free outpouring of the heart—was an aspect of early romanticism. In the 19th century, when the letter-writing art had not yet fallen into desuetude, it was possible for Wilkie Collins to tell the mystery story of *The Moonstone* (1868) in the form of an exchange of letters, but it would be hard to conceive of a detective novel using such a device in the 20th century, when the well-wrought letter is considered

artificial. Attempts to revive the form have not been successful, and Christopher Isherwood's *Meeting by the River* (1967), which has a profoundly serious theme of religious conversion, seems to fail because of the excessive informality and chattiness of the letters in which the story is told. The 20th century's substitute for the long letter is the transcribed tape recording—more, as Beckett's play *Krapp's Last Tape* indicates, a device for expressing alienation than a tool of dialectic. But it shares with the Richardsonian epistle the power of seeming to grant direct communication with a fictional character, with no apparent intervention on the part of the true author.

Pastoral. Fiction that presents rural life as an idyllic condition, with exquisitely clean shepherdesses and sheep immune to foot-rot, is of very ancient descent. Longus' *Daphnis and Chloe,* written in Greek in the 2nd or 3rd century AD, was the remote progenitor of such Elizabethan pastoral romances as Sir Philip Sidney's *Arcadia* (1590) and Thomas Lodge's *Rosalynde* (1590), the source book for Shakespeare's *As You Like It.* The *Paul et Virginie* of Bernardin de St. Pierre (1787), which was immensely popular in its day, seems to spring less from the pastoral utopian convention than from the dawning romanticism that saw in a state of nature only goodness and innocence. Still, the image of a rural Eden is a persistent one in Western culture, whatever the philosophy behind it, and there are elements of this vision even in D.H. Lawrence's *Rainbow* (1915) and, however improbable this may seem, in his *Lady Chatterley's Lover* (1928). The more realistic and ironic pictures of the pastoral life, with poverty and pig dung, beginning with George Crabbe's late-18th-century narrative poems, continuing in George Eliot, reaching sour fruition in Thomas Hardy, are usually the work of people who know the country well, while the rural idyll is properly a townsman's dream. The increasing stresses of urban life make the country vision a theme still available to serious fiction, as even a work as sophisticated as Saul Bellow's *Herzog* (1964) seems to show. But, since Stella Gibbons' satire *Cold Comfort Farm* (1932), it has been difficult for any British novelist to take seriously pastoral lyricism.

Realism in the pastoral novel

Apprenticeship. The *Bildungsroman,* or novel about upbringing and education, seems to have its beginnings in Goethe's work, *Wilhelm Meisters Lehrjahre* (1796), which is about the processes by which a sensitive soul discovers its identity and its role in the big world. A story of the emergence of a personality and a talent, with its implicit motifs of struggle, conflict, suffering, and success, has an inevitable appeal for the novelist; many first novels are autobiographical and attempt to generalize the author's own adolescent experiences into a kind of universal symbol of the growing and learning processes. Charles Dickens embodies a whole *Bildungsroman* in works like *David Copperfield* (1850) and *Great Expectations* (1861), but allows the emerged ego of the hero to be absorbed into the adult world, so that he is the character that is least remembered. H.G. Wells, influenced by Dickens but vitally concerned with education because of his commitment to socialist or utopian programs, looks at the agonies of the growing process from the viewpoint of an achieved utopia in *The Dream* (1924) and, in *Joan and Peter* (1918), concentrates on the search for the right modes of apprenticeship to the complexities of modern life.

The school story established itself in England as a form capable of popularization in children's magazines, chiefly because of the glamour of elite systems of education as first shown in Thomas Hughes's *Tom Brown's School Days* (1857), which is set at Rugby. In France, *Le Grand Meaulnes* (1913) of Alain-Fournier is the great exemplar of the school novel. The studies of struggling youth presented by Hermann Hesse became, after his death in 1962, part of an American campus cult indicating the desire of the serious young to find literary symbols for their own growing problems.

Samuel Butler's *Way of All Flesh,* which was written by 1885 but not published until 1903, remains one of the greatest examples of the modern *Bildungsroman;* philosophical and polemic as well as moving and comic, it presents the struggle of a growing soul to further, all unconsciously, the aims of evolution, and is a devastating indictment of Victorian paternal tyranny. But probably James Joyce's *Portrait of the Artist as a Young Man* (1916), which portrays the struggle of the nascent artistic temperament to overcome the represssions of family, state, and church, is the unsurpassable model of the form in the 20th century. That the learning novel may go beyond what is narrowly regarded as education is shown in two remarkable works of the 1950s—William Golding's *Lord of the Flies* (1955), which deals with the discovery of evil by a group of shipwrecked middle class boys brought up in the liberal tradition, and J.D. Salinger's *Catcher in the Rye* (1951), which concerns the attempts of an adolescent American to come to terms with the adult world in a series of brief encounters, ending with his failure and his ensuing mental illness.

Roman à clef. Real, as opposed to imaginary, human life provides so much ready-made material for the novelist that it is not surprising to find in many novels a mere thinly disguised and minimally reorganized representation of actuality. When, for the fullest appreciation of a work of fiction, it is necessary for the reader to consult the real-life personages and events that inspired it, then the work is a *roman à clef,* or novel that needs a key. In a general sense, every work of literary art requires a key or clue to the artist's preoccupations (the jail in Dickens; the mysterious tyrants in Kafka, both leading back to the author's own father), but the true *roman à clef* is more particular in its disguised references. Chaucer's "Nun's Priest's Tale" has puzzling naturalistic details that can be cleared up only by referring the poem to an assassination plot in which the Earl of Bolingbroke was involved. Swift's *Tale of a Tub* (1704), Dryden's *Absalom and Achitophel* (1681), and Orwell's *Animal Farm* (1945) make total sense only when their hidden historical content is disclosed. These, of course, are not true novels, but they serve to indicate a literary purpose that is not primarily aesthetic. Lawrence's *Aaron's Rod* requires a knowledge of the author's personal enmities, and to understand Aldous Huxley's *Point Counter Point* fully one must know, for instance, that the character of Mark Rampion is D.H. Lawrence himself and that of Denis Burlap is the critic John Middleton Murry. Proust's *À la recherche du temps perdu* becomes a richer literary experience when the author's social milieu is explored, and Joyce's *Finnegans Wake* has so many personal references that it may be called the most massive *roman à clef* ever written. The more important the *clef* becomes to full understanding, the closer the work has come to a special kind of didacticism. When it is dangerous to expose the truth directly, then the novel or narrative poem may present it obliquely. But the ultimate vitality of the work will depend on those elements in it that require no key.

Anti-novel. The movement away from the traditional novel form, in France in the form of the *nouveau roman,* tends to an ideal that may be called the anti-novel—a work of the fictional imagination that ignores such properties as plot, dialogue, human interest. It is impossible, however, for a human creator to create a work of art that is completely inhuman. Contemporary French writers like Alain Robbe-Grillet in *Jealousy* (1957), Nathalie Sarraute in *Tropisms* (1939) and *The Planetarium* (1959), and Michel Butor in *Passing Time* (1957) and *Degrees* (1960) wish mainly to remove the pathetic fallacy from fiction, in which the universe, which is indifferent to man, is made to throw back radar reflections of man's own emotions. Individual character is not important, and consciousness dissolves into sheer "perception." Even time is reversible, since perceptions have nothing to do with chronology, and, as Butor's *Passing Time* shows, memories can be lived backward in this sort of novel. Ultimately, the very appearance of the novel—traditionally a model of the temporal treadmill—must change; it will not be obligatory to start at page 1 and work through to the end; a novel can be entered at any point, like an encyclopaedia.

The two terms most heard in connection with the

French anti-novel are *chosisme* and *tropisme*. The first, with which Robbe-Grillet is chiefly associated, relates to the novelist's concern with things in themselves, not things as human symbols or metaphors. The second, which provided a title for Nathalie Sarraute's early novel, denotes the response of the human mind to external stimuli—a response that is general and unmodified by the apparatus of "character." It is things, the furniture of the universe, that are particular and variable; the multiplicity of human observers melts into an undifferentiable mode of response. Needless to say, there is nothing new in this epistemology as applied to the novel. It is present in Laurence Sterne (in whom French novelists have always been interested), as also in Virginia Woolf.

Chosisme and tropisme

Such British practitioners of the anti-novel as Christine Brooke-Rose and Rayner Heppenstall (both French scholars, incidentally) are more empirical than their French counterparts. They object mainly to the falsification of the external world that was imposed on the traditional novel by the exigencies of plot and character, and they insist on notating the minutiae of the surface of life, concentrating in an unhurried fashion on every detail of its texture. A work like Heppenstall's *Connecting Door* (1962), in which the narrator-hero does not even possess a name, is totally unconcerned with action but very interested in buildings, streets, and the sound of music. This is properly a fresh approach to the materials of the traditional novel rather than a total liberation from it. Such innovations as are found in the *nouveau roman* can best show their value in their influence on traditional novelists, who may be persuaded to observe more closely and be wary of the seductions of swift action, contrived relationships, and neat resolutions.

Cult, or coterie, novels. The novel, unlike the poem, is a commercial commodity, and it lends itself less than the materials of literary magazines to that specialized appeal called coterie, intellectual or elitist. It sometimes happens that books directed at highly cultivated audiences—like *Ulysses, Finnegans Wake,* and Djuna Barnes's *Nightwood* (1936)—achieve a wider response, sometimes because of their daring in the exploitation of sex or obscenity, more often because of a vitality shared with more demotic fiction. The duplicated typescript or the subsidized periodical, rather than the commercially produced book, is the communication medium for the truly hermetic novel.

The novel that achieves commercial publication but whose limited appeal precludes large financial success can frequently become the object of cult adulation. In the period since World War II, especially in the United States, such cults can have large memberships. The cultists are usually students (who, in an era of mass education, form a sizable percentage of the total population of the United States), or fringes of youth sharing the student ethos, and the novels chosen for cult devotion relate to the social or philosophical needs of the readers. The fairy stories of Tolkien, *The Lord of the Flies* of Golding, the science fiction of Kurt Vonnegut, Jr., have, for a greater or lesser time, satisfied a hunger for myth, symbols, and heterodox ideas, to be replaced with surprising speed by other books. The George Orwell cult among the young was followed by a bitter reaction against Orwell's own alleged reactionary tendencies, and such a violent cycle of adoration and detestation is typical of literary cults. Adult cultists tend, like young ones, to be centred in universities, from which they circulate newsletters on *Finnegans Wake,* Anthony Powell's *Music of Time* sequence, and the works of Evelyn Waugh. Occasionally new public attention becomes focussed on a neglected author through his being chosen as a cult object. This happened when the novellas of Ronald Firbank, the anonymous comic novel *Augustus Carp, Esq.*, and G.V. Desani's *All About Mr. Hatterr* got back into print because of the urging of minority devotees. Despite attempts to woo a larger public to read it, Malcolm Lowry's *Under the Volcano* obstinately remained a cult book, while the cultists performed their office of keeping the work alive until such time as popular taste should become sufficiently enlightened to appreciate it.

Student cultists

Detective, mystery, thriller. The terms detective story, mystery, and thriller tend to be employed interchangeably. The detective story thrills the reader with mysterious crimes, usually of a violent nature, and puzzles his reason until their motivation and their perpetrator are, through some triumph of logic, uncovered. The detective story and mystery are in fact synonymous, but the thriller frequently purveys adventurous *frissons* without mysteries, like the spy stories of Ian Fleming, for example, but not like the spy stories of Len Deighton, which have a bracing element of mystery and detection. The detective novel began as a respectable branch of literature with works like Poe's "Murders in the Rue Morgue" (1841), Dickens' unfinished *Edwin Drood* (1870), and Wilkie Collins' *Moonstone* (1868) and *Woman in White* (1860). With the coming of the Sherlock Holmes stories of Sir Arthur Conan Doyle, at the beginning of the 20th century, the form became a kind of infra-literary subspecies, despite the intellectual brilliance of Holmes's detective work and the high literacy of Doyle's writing. Literary men like G.K. Chesterton practiced the form on the margin, and dons read thrillers furtively or composed them pseudonymously (*e.g.*, J.I.M. Stewart, reader in English literature at Oxford, wrote as "Michael Innes"). Even the British poet laureate, C. Day Lewis, subsidized his verse through writing detective novels as "Nicholas Blake." Dorothy L. Sayers, another Oxford scholar, appeared to atone for a highly successful career as a mystery writer by turning to religious drama and the translating of Dante, as well as by making her last mystery novel —*Gaudy Night* (1935)—a highly literary, even pedantic, confection.

Such practitioners as Agatha Christie, Ellery Queen, Erle Stanley Gardner, Raymond Chandler, to say nothing of the highly commercial Edgar Wallace and Mickey Spillane, have given much pleasure and offended only the most exalted literary canons. The fearless and intelligent amateur detective, or private investigator, or police officer has become a typical hero of the modern age. And those qualities that good mystery or thriller writing calls for are not to be despised, since they include economy, skillful sustention of suspense, and very artful plotting.

The mystery novel was superseded in popularity by the novel of espionage, which achieved a large vogue with the James Bond series of Ian Fleming. Something of its spirit, if not its sadism and eroticism, had already appeared in books like John Buchan's *Thirty-nine Steps* and the "entertainments" of Graham Greene, as well as in the admirable novels of intrigue written by Eric Ambler. Fleming had imitators, as well as a more than worthy successor in Len Deighton, but the espionage novel had a comparatively short-lived popularity. It appeared that the novel of crime and detection, or of intrigue and adventure unrelated to the Cold War in which James Bond was engaged, had a sempiternal vitality, while the spy book was of very ephemeral appeal.

Western. Man's concern with taming wild land, or advancing frontiers, or finding therapy in reversion from the civilized life to the atavistic is well reflected in adventure novels, beginning with James Fenimore Cooper's novels of the American frontier *The Pioneers* (1823) and *The Last of the Mohicans* (1826). As the 19th century advanced, and new tracts of America were opened up, a large body of fiction came out of the men who were involved in pioneering adventure. Mark Twain's *Roughing It* (1872) may be called a frontier classic. Bret Harte wrote shorter fiction, like "The Luck of Roaring Camp" (1868), but helped to spread an interest in frontier writing to Europe, where the cult of what may be termed the western novel is as powerful as in America. Owen Wisters' *Virginian* (1902), Andy Adams' near-documentary *Log of a Cowboy* (1903), Emerson Hough's *Covered Wagon* (1922), from which the first important western film was made in 1923, Hamlin Garland's *Son of the Middle Border* (1917), and O.E. Rölvaag's *Giants in the Earth* (1927) all helped to make the form popular, but it is to Zane Grey—who wrote more than 50 western novels —that lovers of frontier myth have accorded the greatest devotion. The western is now thought of predominantly

as a cinematic form, but it arose out of literature. Other frontier fiction has come from another New World, the antipodes—South Africa as well as the Australian outback—but the American West has provided the best mythology, and it is still capable of literary treatment. Sophisticated literary devices may be grafted onto the western—surrealistic fantasy or parallels to Shakespeare or to the ancient classics—but the peculiar and perennial appeal of the western lies in its ethical simplicity, the frequent violence, the desperate attempt to maintain minimal civilized order, as well as the stark, near-epic figures from true western history, such as Billy the Kid, Calamity Jane, Wyatt Earp, Annie Oakley, and Jesse James.

The enduring and the ephemeral

The best seller. A distinction should be made between novels whose high sales are an accolade bestowed on literary merit and novels that aim less at aesthetic worth than at profits. The works of Charles Dickens were best sellers in their day, but good sales continue, testifying to a vitality that was not purely ephemeral. On the other hand, many best-selling novels have a vogue that is destined not to outlast the time when they were produced. It is a characteristic of this kind of best seller that the writing is less interesting than the content, and that the content itself has a kind of journalistic oversimplification that appeals to unsophisticated minds. The United States is the primary home of the commercial novel whose high sales accrue from careful, and sometimes cold-blooded, planning. A novel in which a topical subject—such as the Mafia, or corruption in government, or the election of a new pope, or a spate of aircraft accidents, or the censorship of an erotic book—is treated with factual thoroughness, garnished with sex, enlivened by quarrels, fights, and marital infidelities, presented in nonliterary prose, and given lavish promotion by its publisher may well become a best seller. It is also likely to be almost entirely forgotten a year or so after its publication. The factual element in the novel seems to be necessary to make the reader feel that he is being educated as well as diverted. Indeed, the conditions for the highest sales seem to include the reconciliation of the pornographic and the didactic.

A novel with genuine aesthetic vitality often sells more than the most vaunted best seller, but the sales are more likely to be spread over decades and even centuries rather than mere weeks and months. The author of such a book may, in time, enrich others, but he is unlikely himself to attain the opulence of writers of best sellers such as Harold Robbins or Irving Wallace.

Fantasy and prophecy. The term science fiction is a loose one, and it is often made to include fantastic and prophetic books that make no reference to the potentialities of science and technology for changing human life. Nevertheless, a novel like Keith Roberts' *Pavane* (1969), which has as a premise the conquest of England by Spain in 1588, and the consequent suppression rather than development of free Protestant intellectual inquiry, is called science fiction, though such terms as "fiction of hypothesis" and "time fantasy" would be more fitting. The imaginative novelist is entitled to remake the existing world or present possible future worlds, and a large corpus of fiction devoted to such speculative visions has been produced in the last hundred years, more of it based on metaphysical hypotheses than on scientific marvels. Jules Verne and H.G. Wells pioneered what may be properly termed science fiction, mainly to an end of diversion. Since the days of Wells's *Time Machine* (1895) and *Invisible Man* (1897), the fiction of hypothesis has frequently had a strong didactic aim, often concerned with opposing the very utopianism that Wells—mainly in his nonfictional works—built on the potentialities of socialism and technology. Aldous Huxley's *Brave New World* (1932) showed how dangerous utopianism could be, since the desire for social stability might condone conditioning techniques that would destroy the fundamental human right to make free choices. Toward the end of his life Huxley produced a cautious utopian vision in *Island* (1962), but the dystopian horrors of his earlier novel and of his *Ape and Essence* (1948) remain more convincing. Orwell's *Nineteen Eighty-four* (1949) showed a world in which a tyrannic unity is imposed by a collective solipsism, and contradictions are liquidated through the constant revision of history that the controlling party decrees. Anthony Burgess' *Clockwork Orange* (1962) and *Wanting Seed* (1962) portray ghastly futures that extrapolate, respectively, philosophies of crime control and population control out of present-day tendencies that are only potentially dangerous.

A large number of writers practice prophetic fantasy with considerable literary skill and careful factual preparation—Kurt Vonnegut, Jr., Ray Bradbury, Italo Calvino, Isaac Asimov, J.G. Ballard, to name a few—and novelists whose distinction lies mainly in more traditional fields have attempted the occasional piece of future-fiction, like L.P. Hartley with his *Facial Justice* (1961) and Evelyn Waugh in *Love Among the Ruins* (1953). The fantasist who fantasizes without prophetic or warning intent is rarer, but works like Nabokov's *Ada*, Tolkien's *Lord of the Rings* cycle, and Christine Brooke-Rose's *Out* (1964) represent legitimate and heartening stretching of the imagination, assurances that the novelist has the right to create worlds, as well as characters, of his own. But the dystopian novel can have a salutary influence on society, actively correcting regressive or illiberal tendencies, and *Brave New World* and *Nineteen Eighty-four* can be cherished as great didactic landmarks, not just as works of literary art.

Proletarian. The novel that, like Dickens' *Hard Times* (1854), presents the lives of workingmen or other members of the lower orders is not necessarily an example of proletarian fiction. The category properly springs out of direct experience of proletarian life and is not available to writers whose background is bourgeois or aristocratic. Thus, William Godwin's *Caleb Williams* (1794), or Robert Bage's *Hermsprong* (1796), though, like *Hard Times*, sympathetic to the lot of the oppressed worker, is more concerned with the imposition of reform from above than with revolution from within, and the proletarian novel is essentially an intended device of revolution. The Russian Maksim Gorky, with works like *Foma Gordeyev* (1900) and *Mother* (1907), as well as many short stories portraying the bitterness of poverty and unemployment (the pseudonym *Gorky* means "bitter"), may be taken as an exemplary proletarian writer. The United States has produced a rich crop of working-class fiction. Such Socialist writers as Jack London, Upton Sinclair, John Dos Passos, and Edward Dahlberg, however, did not witness the triumph of the workers' revolution in their own country, as Gorky did in his, and it is the fate of the American proletarian novelist, through literary success, either to join the class he once dreamed of overthrowing or to become anarchic and frustrated. In the Soviet Union the proletarian novel was doomed to disappear in the form that Gorky knew; a present-day Gorky would presumably write, in exile, novels about Soviet oppression or else join Yuly Daniel (who writes under the pseudonym Nikolay Arzhak) and Aleksandr Solzhenitsyn in enforced silence and disgrace. For it is the essence of the revolutionary novel only to possess vitality and validity when written under capitalist "tyranny." England has produced its share of working-class novelists exuding bitterness, such as Alan Sillitoe, with his *Saturday Night and Sunday Morning* (1958), but conditions apt for revolution have not existed in Britain for over a century. British novelists who emerged after World War II, such as John Braine (*Room at the Top*), Keith Waterhouse (*There is a Happy Land*), Kingsley Amis (*Lucky Jim*), and Stan Barstow (*A Kind of Loving*), showed a solution to working-class frustration in a fluid system of class promotion: revolution is an inadmissible dream. Generally speaking, in the novel, which is preoccupied with individuals rather than with groups, it is difficult to make the generalized political statements that are meat and drink to the revolutionary propagandist.

Reform versus revolution

Other types. The categories briefly discussed above are among the most common fictional forms. Theoretically there is no limit to the number available, since changing social patterns provide fresh subjects and fresh taxonomies, and new metaphysical and psychological doctrines

may beget new fictional approaches to both content and technique. Other categories of fictional art include the erotic novel (which may or may not be pornographic), the satirical novel, the farcical novel, the novel for or about children, the theological novel, the allegorical novel, and so on. Types of fiction no longer practiced, since their real-life referents no longer exist, include the colonial novel—such as E.M. Forster's *Passage to India* (1924), Henri Fauconnier's *Malaisie* (1930), and the African sequence of Joyce Cary—and space fantasy like H.G. Wells's *First Men in the Moon* (1901). One may read examples of a departed category with pleasure and profit, but the category can no longer yield more than parody or pastiche. New kinds of fiction fill in the gaps, like the novel of negritude, the structuralist novel (following the linguistic sociologists and anthropologists), the homosexual novel, the novel of drug hallucination, and so on. So long as human society continues to exist, the novel will exist as its mirror, an infinitude of artistic images reflecting an infinitude of life patterns.

Outmoded types

II. Development of the novel: the novels of specific nations or languages

THE NOVEL IN ENGLISH

The United Kingdom. England's chief literary achievements lie in the fields of drama and poetry, and the attitude of English novelists to their form was, for a long time, cheerfully empirical and even amateurish. Elizabethan novels, *novelle* rather, imitated the Spanish picaresque story, and Thomas Nashe's *Unfortunate Traveller* (1594) is a good, bustling, vital example of a rapidly composed commercial work more concerned with sensational incident and language than with shape or character. Daniel Defoe (1660?-1731) is often considered to be the true progenitor of the long English novel, but his *Robinson Crusoe* and *Moll Flanders* are loosely constructed, highly episodic, and presented as mock-biography rather than real fiction. It is with *Pamela*, by Samuel Richardson (1689–1761), that the tradition of serious, moral fiction in English may be said to begin, but later 18th-century novelists reverted to the picaresque and comic. Henry Fielding (1707–54) wrote his first novel, *Joseph Andrews*, as a parody on *Pamela*, but his masterpiece, *Tom Jones*, is original if shapeless, an example of English literary genius sinking thankfully back into the casual and improvisatory. Laurence Sterne (1713–68) produced a great mad work, *Tristram Shandy*, that, in its refusal to truckle to any rules of structure, remains still a quarry for avant-garde novelists; and Tobias Smollett (1721–71) wrote picaresque satire in *Roderick Random* and *Peregrine Pickle*, full-blooded portraits of the age that impress more with their vigour than with their art.

The Romantic Age brought, paradoxically, the cool and classically shaped novels of Jane Austen (1775–1817), a major practitioner and still a model for apprentices in the craft. Sir Walter Scott (1771–1832), a Scotsman who wrote about romantic historical Scotland, must be regarded as an international figure whose influence was greater even than Richardson's, since he more than anyone established the historical novel as the primary fictional form in Europe. In his work, nevertheless, the traditional faults of the British novel may be descried—episodic formlessness, an ebullience of texture rather than a clean narrative line. The Victorian Age moved out of the romantic past, or, as with William Makepeace Thackeray (1811–63), stayed with it only to deromanticize it. Charles Dickens (1812–70) was indebted to the picaresque tradition but turned reformist eyes on his own age. Thackeray and Dickens are complementary in that the first attacks the upper classes while the second showers sympathy, sometimes of a mawkish kind, on the lower classes. With George Eliot (1819-80), the first true English psychological novels appear, strong in their moral content, and George Meredith (1828–1909) may be said to have anticipated—in *The Ordeal of Richard Feverel* and *The Egoist*—the approach in depth that characterized the psychological novel of the 20th century.

The Romantic and Victorian ages

Both Charlotte Brontë (1816–55) and Emily Brontë (1818–48) exemplify the capacity of the English novel to achieve solitary "sports" unrelated to any current or tradition. Both *Wuthering Heights*—a superb evocation of the soul of a locality, with a love story that is fierce and primitive but recounted with poetic sophistication—and *Jane Eyre*, an exceedingly frank and still shocking study of a love that rides over Victorian conventions, are unlike any other books of their time, or of any other time, though their qualities have been diluted into hundreds of popular 20th-century romances. Later Victorians, particularly Samuel Butler (1835–1902) and Thomas Hardy (1840–1928), reflected those changes in the educated English sensibility that had been brought about by the new science. Hardy's world is one in which the Christian God has been replaced by a malevolent Providence—the poet-novelist's theologization of scientific determinism. Butler's *Way of All Flesh*, a work that contrives to be both bitterly realistic and high comic, demonstrates the working of Darwinian evolution in social institutions such as the family and even the church. In many ways, Butler leads English fiction into the modern age.

John Galsworthy (1867–1933) showed himself interested in the processes by which old institutions, such as a great Whig family, decay as history advances, but in both style and characterization he is firmly set in the Victorian Age. Of Arnold Bennett (1867–1931) it may be said that he brought to a kind of Thackerayan social realism something of the spirit of the French novel, particularly the anglicized tones of Balzac and Zola. W. Somerset Maugham (1874–1965), though his most considerable achievements lie in the short story form, wrote novels, like *Of Human Bondage*, that are infused with the delicious acerbities of French naturalism, and all his work, long or short, is exalted by the influence of Guy de Maupassant. Early 20th-century British fiction needed the impact of an alien tradition to jolt it out of bourgeois empiricism, and perhaps the two major influences were both foreigners who had elected to write fiction in England—Joseph Conrad (1857–1924), Polish-born, to whom English was a second language, and Henry James (1843–1916), an American who had drunk deeply at French fountains and brought to the exercise of his craft a scrupulousness and a concern with aesthetic values that was almost obsessive and, it may be said, very un-English. James's influence on such major English-born novelists as Virginia Woolf (1882–1941), E.M. Forster (1879–1970), and Ford Madox Ford (1873–1939) was in the direction of that concern with style which the English novel, unlike the French, has always tried to resist, and this influence remains a potent one on succeeding generations of novelists, not only in England.

Other important Englishmen have remained more interested in content than in style, though in Graham Greene (1904–), a Catholic convert, who is primarily known for his "theological" content, there is a very interesting attempt to bring to this a Conradian concern for the solitary-man theme and a Jamesian preoccupation with style. D.H. Lawrence (1885–1930) remains the great modern English novelist who reconciles a highly traditional style (in *The Rainbow*, for example, he often resembles George Eliot) with a subject matter that is revolutionary in the profundity of its human relationships. And Aldous Huxley (1894–1963), though he indulged in formal experiment in *Point Counter Point* and *Eyeless in Gaza*, was always essentially a didactic writer who used the novel form somewhat casually. The same may be said of George Orwell (1903–50), who eschewed stylistic complexity in the interests of a clear message. On the other hand, Evelyn Waugh (1902–66) and Anthony Powell (1905–), in a cautious and conservative way, consulted the claims of allusive and evocative prose, as did Wyndham Lewis (1884–1957), who also brought to his novels the aesthetic of an alien art—that of painting.

Novelists like Kingsley Amis (1922–) and Angus Wilson (1913–) are seen to belong to traditions already old in the time of Samuel Butler. Amis derives from Fielding (Amis' *Take a Girl Like You* has a moral quality reminiscent of Fielding's *Amelia*), and Wilson never pretends to be other than a disciple of novelists like George Eliot and Dickens. The Victorian kind of novel,

as practiced by writers like J.B. Priestley (1894–), satisfies large numbers of British readers, and Jamesian scrupulosity and Joycean experiment alike are regarded with amiable suspicion. Such Englishmen as have experimented in fictional technique have usually found a larger audience abroad than at home. Thus Lawrence Durrell (1912–) found his *Alexandria Quartet* hailed as a masterpiece in France, while English readers merely liked or disliked it. *A Clockwork Orange* by Anthony Burgess (1917–) achieved a large readership in the United States, but it recorded little positive response in the country of its origin.

Irish and Scottish novels

The fiction of the British Isles is not clearly differentiated into national or regional groups, and to speak of Irish novelists is often to do no more than refer to an irrelevant place of birth. Oscar Wilde (1854–1900) was more French than Irish; his one novel, *The Picture of Dorian Gray*, has no progenitors in English fiction, but it owes much to the late-19th-century French novelist Joris-Karl Huysmans. George Moore (1852–1933) was born in County Mayo, Ireland, but studied art in Paris, and works like *A Modern Lover* and *A Mummer's Wife* were intended to teach the British novel how to absorb the spirit of Flaubert and Zola. Moore's *Lake*, written under the influence of the Irish literary revival, is dutifully set in Ireland, but this essentially international writer was quick to move back to France for *Héloïse and Abélard* and on to Palestine for *The Brook Kerith*. James Joyce (1882–1941), conceivably the greatest novelist in English in the 20th century, never moves from his native Dublin in any of his fiction, but no less parochial writer can well be imagined. *Ulysses*, in the depth of its historical coverage, may be regarded as one of the last great artistic monuments of the Austro-Hungarian Empire, and *Finnegans Wake* was stimulated by the avant-garde climate of Paris, the resting place of so many Irish "wild geese." It is only perhaps negatively that Ireland shaped Joyce's literary personality; the oppressiveness of the new nationalism made him react in the direction of internationalism, and it eventually forced him to a lifelong exile. Flann O'Brien, possibly Joyce's true successor, stayed in Dublin and draws, in works like *At Swim-Two-Birds* (1939) and *The Hard Life* (1961), on the native demotic experience, but his techniques come from Europe, not from the Anglo-Irish bourgeois stockpot.

The Scottish novel hardly exists as a national entity; there is nothing in fiction that matches the exploitation of Lallans in the poetry of Burns or of Hugh MacDiarmid (C.M. Grieve). Scott, as has been indicated, belongs to European literature, and later novelists like Robert Louis Stevenson (1850–94) and Sir James Barrie (1860–1937), though Scottish themes and settings are featured triumphantly in their fiction, are essentially men whose literary metropolis was London rather than Edinburgh.

Wales sedulously cultivates Welsh as a living artistic medium, but few novels by Welshmen reach the English or international market, unless—like Emyr Humphreys—they write their works in English as well as Welsh. The English-language novel from Wales hardly exists, except in the form of best sellers like Richard Llewellyn's *How Green Was My Valley* (1940)—a deliberate capitalization on Welsh picturesqueness—and the very Joycean prose of the poet Dylan Thomas (1914–53), whose one novel, *Adventures in the Skin Trade*, has very little of Wales in it.

The United States. If the American novel appears to begin with the *Wieland* and *Edgar Huntly* of Charles Brockden Brown (1771–1810), then the fiction of the mother country may be said to have no start on that of the newly independent daughter. Admittedly, Brown owes something to the English Gothic novel, but already a typically American note is struck in the choice of a violent and bizarre form that, in various mutations, was to prove fruitful in the development of American fiction. James Fenimore Cooper (1789–1851) struck out, in the pentateuch called the "Leatherstocking" tales, in another direction suitable to the American genius and experience: *The Last of the Mohicans* and *The Prairie*, though their prose style is perhaps overelaborate, are full of the spirit of a young nation confronting the wilderness and advancing its frontiers. Harriet Beecher Stowe (1811–96) produced, in *Uncle Tom's Cabin*, an anti-slavery novel whose sensational devices owe something to the Gothic movement but whose style is rooted in that European romantic tradition fathered by Scott. Even Edgar Allan Poe (1809–49) and Nathaniel Hawthorne (1804–64) are incompletely emancipated from Europe, though Hawthorne's *Scarlet Letter* shows a preoccupation with sin that finds no fictional counterpart in the main stream of the European novel. America, aware of the darkness and mystery of a land still mainly undiscovered, was correspondingly aware, in both Gothic and eschatological fiction, of the dark places of the mind.

The true emancipation from Europe comes in the works of Hawthorne's friend Herman Melville (1819–91), whose *Moby Dick* creates a totally American fusion by combining plain adventure with a kind of Manichaean symbolism—implying a belief that the universe is under the dominion of two opposing principles, one good and the other evil. Mark Twain (1835–1910) brought the Mississippi frontier region into the literary geography of the world with *Tom Sawyer* and *Huckleberry Finn*, combining very American humour with harsh social criticism, as also in the unique *Puddn'head Wilson*. From Twain on, the new regions of the United States become the materials of a wealth of fiction. Thus, while Bret Harte (1836–1902) wrote about California and George W. Cable (1844–1925) on Louisiana, Indiana was celebrated in *The Hoosier Schoolmaster* of Edward Eggleston (1837–1902) and the *Penrod* stories of Booth Tarkington (1869–1947).

American innocence and European sophistication

But America could not wholly turn its back on Europe, and the early 20th century was characterized by the discovery of the phenomenon called American "innocence" in confrontation with the decadent wisdom and sophistication of the Old World. If, in *Innocents Abroad*, Mark Twain and his fellow travellers remained unimpressed by Europe, Henry James (1843–1916) devoted millions of words to the response of impressionable Americans to the subtle deciduous culture their ancestors had abandoned in the search for a new life. Zolaesque realism entered the American novel with William Dean Howells (1837–1920). Theodore Dreiser (1871–1945) shocked his audiences with the candour and pessimism of *Sister Carrie* and *An American Tragedy*, while Stephen Crane (1871–1900) and Frank Norris (1870–1902) made realism enter, respectively, the traditionally romantic field of war (*The Red Badge of Courage*) and the pastoral life of the promised land of California (*The Octopus*). American innocence ceased to be a fictional property. The brutality in Jack London (1876–1916) joins hands with the depictions of "natural man" in the works of Ernest Hemingway (1899–1961), while the urban fiction of James T. Farrell (1904–) and John O'Hara (1905–70) has a wholly American realism that leaves Zola far behind.

The 20th-century American novel is "compartmentalized" in a manner that seems to give the lie to any allegation of cultural unity. The cult of the regional novel has continued. Ohio came under the microscopic scrutiny of Sherwood Anderson (1876–1941) in *Winesburg, Ohio*, just as the small-minded hypocrisy and materialism of the Midwest fired the brilliant sequence of Sinclair Lewis (1885–1951) from *Main Street* on. But the best regional fiction has come from the South, with William Faulkner (1897–1962), whose technical master is Joyce, Ellen Glasgow (1874–1945), Erskine Caldwell (1903–), and Robert Penn Warren (1905–). The immigrant communities of Nebraska were the subject of the *My Ántonia* of Willa Cather (1873–1947), while Ole Rölvaag (1876–1931) dealt with the South Dakota Norwegians in *Giants in the Earth*. The urban Jew has become the very spokesman of the contemporary American experience with writers like Saul Bellow (1915–), Herbert Gold (1924–) and Bernard Malamud (1914–); while the black American has been made highly articulate in works like *Go Tell It on the Mountain* by James Baldwin (1924–), and *Invisible Man* by Ralph Ellison (1914–).

Meanwhile, the fiction of social criticism that derives from Howells and was made popular by men like Upton Sinclair (1878–1968) goes on, either in "muckraking" best sellers or in such specialized attacks on the American ethos as are exemplified by war books like Joseph Heller's *Catch-22* and Kurt Vonnegut's *Slaughterhouse-Five*. American realism has gained a reputation for candour unequalled in any other literature of the world, so that the *Tropic of Cancer* and *Tropic of Capricorn* of Henry Miller (1891–) remain models for sexual frankness that encouraged writers like Hubert Selby, Jr. (*Last Exit to Brooklyn*), and Philip Roth (*Portnoy's Complaint*) to uncover areas of eroticism still closed to the European novel. But to balance the concern with naked subject matter, American fiction has also fulfilled the promise of its earlier expatriates beginning with Gertrude Stein (1874–1946) to match the French in a preoccupation with style. The distinction of F. Scott Fitzgerald (1896–1940) lay in the lyrical intensity of his prose as much as in his "jazz age" subject matter, and writers like Truman Capote (1924–) and John Updike (1932–) have dedicated themselves similarly to perfecting a prose instrument whose effects (like those of Joyce) touch the borders of poetry. Perhaps the glories and potentialities of American fiction are best summed up in the novels of Vladimir Nabokov (1899–). His earlier works belong to Europe, but when he took to writing in English he created a sophisticated, cosmopolitan, and highly poetic style. His work, as in *Lolita* and *Pale Fire*, concentrates with unparalleled intensity on the immediacies of American life in the 20th century.

The British Commonwealth. Canada has two literatures—one in French as well as one in English—and, in the taxonomy of this article, it will be as well to forget the course of history and consider French Canadian novelists along with those of the separated, and officially abandoned, mother country. The Canadian novel in English begins, appropriately enough, with a Richardson—John Richardson (1796–1852), author of *Wacousta*, a story of the Indian uprising led by Pontiac. But, with the exception of James DeMille (1836–80), author of a remarkable novel, *A Strange Manuscript Found in a Copper Cylinder*, and William Kirby (1817–1906), whose *Golden Dog* is an interesting long romance of 18th-century Quebec, no novelist of stature emerged in what was a great period of literary expansion in the United States. The historical novels of Sir Gilbert Parker (1862–1932), the western romances of Ralph Connor (1860–1937), and the world-famous *Anne of Green Gables* by Lucy Maud Montgomery (1874–1942) exhibit a stylistic tameness, along with that jejune oversimplification of psychology that is a mark of the deliberately "popular" novel. More distinctive work came in the early 20th century, as in the prairie novels of Robert Stead (1880–1959) and Frederick Philip Grove (1871–1948). Morley Callaghan (1903–), a chronicler of urban life, may be seen as a writer approaching international stature, and that tough "European" realism that braces his work is also to be found, along with a rather exotic elegance, in the novels of Robertson Davies (1913–). Younger novelists, like Brian Moore (1921–) and Mordecai Richler (1931–), have sought artistic stimulation outside Canada, and they show general uneasiness about the lack of a cultivated audience in their great but underpopulated land.

Australian and New Zealand novels

A similar complaint may be heard in Australia and New Zealand, where the fictional tradition has long accommodated itself to the simple literary needs of unsophisticated settlers. The first true Australian novel was probably *For the Term of His Natural Life*, by Marcus Clarke (1846–81), which dealt, appropriately, considering the penal origins of the Australian settlements, with life in prison; and Rolf Boldrewood, the pseudonym of Thomas A. Browne (1826–1915), had a cognate approach of excessive simplicity and melodrama to the days of the gold rush in his *Robbery Under Arms*. Australian farm life was the theme of the novels that Arthur Hoey Davis (1868–1935) wrote under the pen name of Steele Rudd, beginning with *On Our Selection*. But something that can only be termed native Australianism—which has less to do with trades and scenic backgrounds than with a whole new language and collective philosophy of life—appears for the first time in *My Brilliant Career* (1901), by Miles Franklin (1879–1954). The expression of a national personality is not, however, enough. Progressive 20th-century Australian novelists look hungrily toward Europe, aware that native conservatism will not support the kind of experimentalism in literature that it is prepared to take for granted in painting and architecture. The most considerable modern Australian novelist is Patrick White (1912–), but even such massive achievements as *Voss* and *Riders in the Chariot* are set firmly in the Dostoyevskian tradition. Morris West (1916–) is Australian only by birth; he seeks international themes and the lure of the American best-seller market. It seems that an Australian writer can succeed only if he renounces his Australianism and goes into exile. It is still a source of humiliation to Australian litterateurs that the finest novel about Australia was written by a mere visitor—*Kangaroo* (1923), by D.H. Lawrence.

New Zealand, with its much smaller population, presents even greater problems for the native novelist seeking an audience, and this—along with New Zealand's closer tie with the mother country—explains why such fiction writers as Katherine Mansfield (1888–1923) and Dame Ngaio Marsh (1899–) made London their centre. Young novelists like Janet Frame, Ian Cross, and Maurice Shadbolt show, however, a heartening willingness to render the New Zealand scene in idioms and techniques more progressive than anything to be found in Australia, and they are prepared to resist the temptations of the expatriate life.

Inevitably, territories like India and Africa entered the federation of English literature very much later than those dominions founded on the English language. With few exceptions, African and Indian novelists employ English as a second language, and one of the charms of their novels lies in a creative tension between the adopted language and the native vernacular (needless to say, this is usually self-consciously exploited—often for poetic, but more frequently for comic, effect).

English as a second language of the novelist

Of Indian novelists, R.K. Narayan (1906–), in works like *The English Teacher* (U.S. title *Grateful to Life and Death*) and *The Man-Eater of Malgudi*, exhibits an individual combination of tenderness and humour, as well as a sharp eye for Indian foibles. Raja Rao (1909–), whose best known novel is *The Serpent and the Rope*, achieves remarkable prosodic effects through allowing Sanskrit rhythm and idiom to fertilize English. Of younger Indian novelists, Balachandra Rajan (1920–) is notable, in *The Dark Dancer* and *Too Long in the West*, for an ability to satirize the Anglo-American way of life with the same suave elegance that informs his tragicomic view of the East. Khushwant Singh (1915–) presents, in *I Shall Not Hear the Nightingale*, a powerful chronicle of Sikh life during that period of imperial dissolution that began with World War II. English seems established as the medium for the Indian novel, and it is interesting to note an ability on the part of nonnative Indian residents who have practiced the form to be absorbed into a new "Indo-Anglian" tradition. Rudyard Kipling's *Kim* is respected by Indian writers, and E.M. Forster's *Passage to India* is a progenitor of one kind of Indian novel. The novels of Paul Scott (1920–)—such as *Johnnie Sahib* and *The Birds of Paradise*—spring out of a love of the country and an understanding of its complexities not uncommon among former British soldiers and administrators. The Indian novel is perhaps a product of territory rather than of blood.

African novelists

The most important of the new African writers come from the West Coast, traditional fount of artists, and they are mostly characterized by immense stylistic vigour, a powerful realism, and, often, a satirical candour unseduced by the claims of the new nationalism. Chinua Achebe (1930–), in *Things Fall Apart* and *No Longer at Ease*, renders remarkably the tones of Umuaro speech and thought, and exhibits, as also in *Arrow of God*, a concern for that rich native culture whose extirpation is

threatened by imported Western patterns of life and government. His *Man of the People* deals sharply with corruption and personality cult in a newly independent African state modelled on his own Nigeria. A fellow Nigerian, Amos Tutuola (1920–), has gained an international reputation with *The Palm-Wine Drinkard*—a richly humorous novel permeated with the spirit of folklore. Cyprian Ekwensi (1921–) is best known for his *Jagua Nana*, a wry study of the impact of the new materialism (symbolized by the "Jagua" car of the title) on the tribal mind. Onuora Nzekwu (1928–), in *Blade Among the Boys*, has a graver theme: the conflict between Ibo religion and imported Christianity in the upbringing of a sensitive and confused young man.

South African novels have, traditionally, dealt with those pioneering themes still exemplified in the work of Laurens van der Post and Stuart Cloete, but the territory has made its entry into world literature comparatively recently chiefly because the official racist policies have inspired a powerful fiction of protest. William Plomer (1903–) may be regarded as the father of anti-apartheid literature, although his *Turbott Wolfe* appeared in 1925, long before the state doctrine was articulated. The theme of this novel was the necessity for white and black blood to mix and ensure a liberal South African future not given over purely to white domination. Alan Paton (1903–), in *Cry, the Beloved Country*, has produced the most popular novel of protest, but Dan Jacobson (1929–), Nadine Gordimer (1923–), and especially Doris Lessing (1919–) have amplified mere protest into what may be termed a kind of philosophical fiction, often distinguished enough to rank with the best work of Europe and America. Mrs. Lessing's sequence *Children of Violence* finds in South Africa a kind of starting point for denunciation of wrongs that turn out to be social and sexual as well as racial.

Caribbean works

The varied fictional achievements of the Caribbean are large. Trinidad has produced the two best known West Indian writers—Samuel Selvon and V.S. Naipaul (1932–), both of East Indian descent. Selvon's work (*A Brighter Sun, An Island is a World, The Lonely Londoners*) is grim, bitter, capable of vivid evocation of the Trinidadian scene, but Naipaul, after *A House for Mr. Biswas*, has shown signs of habituation to his English exile, so that his *Mr. Stone and the Knights Companion* seems to be a novel totally nourished by the London world in which it is set. Most Caribbean novelists, finding their publishers and their audiences in England, transfer themselves thither and cut themselves off from all but remembered roots. This is true of Edgar Mittelhölzer (1909–65), an ebullient and prolific writer whose later books all had English settings. Wilson Harris has, in English exile, created an astonishing Guianan tetralogy in which poetry and myth and symbolic difficulty have a place. George Lamming (1927–) and John Hearne (1926–) are both notable for firm and economical prose and masterly scene painting. The aesthetic prospects for the West Indian novel seem excellent, but the absorption of its practitioners into the larger English-speaking world represents a symptom that plagues the practice of literature in so many Commonwealth countries—the lack of adequate publishing facilities and, more than that, the failure of a cultivated readership to emerge. It is undoubtedly unhealthy for an author to have to seek primary communication with foreigners.

EUROPE

Russian. The Russian novel properly begins with Nikolay Karamzin (1766–1826), who introduced into Russian literature not only the exotic sentimental romanticism best seen in his novels *Poor Lisa* and *Natalya, the Boyar's Daughter* but that large Gallic vocabulary which was to remain a feature of the literary language. But those qualities that best distinguish Russian fiction—critical realism and spirituality—first appeared in the work of Mikhail Lermontov (1814–41), whose *Hero of Our Time* is the pioneering Russian psychological novel. Nikolay Gogol (1809–52), satirizing provincial mores in *Dead Souls*, ushered in, perhaps unintentionally, a whole fictional movement—the "literature of accusation." Vissarion Belinsky (1811–48), the father of Russian literary criticism, formulated the theory of literature in the service of society, and Ivan Turgenev (1818–83), produced a classic "accusatory" novel in *A Sportsman's Sketches*. But he, and the other major novelists who emerged in the mid-19th century, can hardly be considered in terms of literary movements. Fyodor Dostoyevsky (1821–81), with *The Possessed, The Idiot, Crime and Punishment*, and *The Brothers Karamazov*, affirmed with idiosyncratic power the great spiritual realities, and Leo Tolstoy (1828–1910) produced two of the greatest novels of all time—*War and Peace* and *Anna Karenina*, revelatory of the Russian soul but also of the very nature of universal man and human society. The later days of the 19th century saw a shift in fictional radicalism. Marx influenced the accusatory writers in the direction of the plight of the proletariat, not, as had been the old way, that of the peasantry. Maksim Gorky (1868–1936), Leonid Andreyev (1871–1919), Aleksandr Kuprin (1870–1938) and the Nobel Prize winner (1933) Ivan Bunin (1870–1953) wrote of the Russian urban experience and helped to create the literary climate of the Soviet regime. Generally, since the beginning of the first five-year plan in 1928, there has been a division between what the regime regards as valuable in the practice of the novel and what the rest of the world thinks. Mikhail Sholokhov (1905–) depicted, with no evident propagandist slanting, the Revolution and civil war in *Quiet Flows the Don*, and Fyodor Gladkov (1883–1958) was one of the few novelists readable outside the ranks of the Soviet devout in the new category of economic or industrial fiction. The Russian novel of the age following World War II, however, has undergone a dramatic schism, in which writers like Aleksandr Solzhenitsyn (1918–) and Boris Pasternak (1890–1960) receive the Nobel Prize but are officially condemned in the U.S.S.R., while the fictional darlings of the regime are recognized, in the non-Communist world, as possessing little or no aesthetic merit.

The influence of the Soviet government

German. Goethe (1749–1832), who practiced so many arts with such notable brilliance, may be regarded as the first major novelist of Germany. His life covers the whole period of the Enlightenment, with its insistence on a national literary spirit, the *Sturm und Drang* movement, and that phase of *Weltschmerz* which *The Sorrows of Werther* fostered. It was *Werther* more than any other work that carried German literature into the world arena; it remained influential in Europe when German Romanticism had already burned itself out. The German reaction against Romanticism was expressed in the regionalism of Theodor Storm (1817–88) and Fritz Reuter (1810–74), who sought to render with objective fidelity the life of their native provinces (northwestern and northeastern Germany, respectively). Swiss writers like Gottfried Keller (1819–90) belonged to the movement of mainstream German regional realism, featuring the picturesque solidities of Switzerland. But the post-Goethean major achievements in the novel had to wait for the Impressionist movement, which produced the works of Thomas Mann (1875–1955) and Hermann Hesse (1877–1962)—fiction concerned less with a roughly hacked slice of life than with form and aesthetic delicacy. World War I brought Expressionism—the nightmares of the German-Czech Franz Kafka (1883–1924) and the psychoanalytical novels of Jakob Wassermann (1873–1934), with their pleas for humanity and justice. When the true historical nightmare of the Nazi regime followed—predicted, in a sense, by Kafka—liberal German fiction was suppressed, and liberal German novelists like Mann and Erich Maria Remarque (1898–1970), author of *All Quiet on the Western Front*, went into exile. The brutal, philistine, and nationalistic novels of the true Nazi novelists exist only as curious and frightening relics of an era of infamy. The task of postwar novelists like Günter Grass (1927–)—author of *The Tin Drum* and *Dog Years*—and Uwe Johnson (1934–) has been to diagnose the long sickness and force German fiction into new directions—often with the help of surrealism, irony, and verbal experiment.

Suppression of the novel by the Nazis

French. Although the mid-16th-century satirical piece *Gargantua and Pantagruel* of François Rabelais has had, and is still having, a profound influence on the world novel, it would be wrong to place it in the line of true French fiction, which has always manifested a preoccupation with form, order, and economy—qualities totally un-Rabelaisian. The first notable French novel—*The Princess of Clèves* by Madame de La Fayette (1634–93)—shows none of the vices of the English novels of a century later: it is firmly constructed and takes character seriously, as does the *Manon Lescaut* of l'Abbé Prévost (1697–1763). With such works the psychological novel was established in Europe. A keen hold on reality and a concern with the problems of man as a social being animates even the fiction of the Romantic school—works like *Indiana* and *Lélia* by George Sand (1804–76) and the heavyweight romances of Victor Hugo (1802–85). But the true glories of French fiction come with the reaction to romanticism and sentimentality, as exemplified in the novels of Jean-Jacques Rousseau (1712–78). The great fathers of realism are Stendhal (1783–1842), Balzac (1799–1850), and Flaubert (1821–80), and their influence is still active. Émile Zola (1840–1902) moved away from the artistic detachment that Flaubert preached and practiced, but, in his *Chronicles of the Rougon Macquart Family*, he tried to emulate the encyclopaedic approach to the novel of Balzac, whose *Human Comedy* is meant to be a history of society in a hundred episodes. The leader of the Naturalist school, Zola saw human character as a product of heredity and environment. It was left to the 20th-century French novel to cast doubt on a mechanistic or deterministic view of man, to affirm the irrational element in his makeup, and to emphasize the primacy of the will. The temporal treadmill of Balzac and Zola has no place in the masterpiece of Marcel Proust (1871–1922); *Remembrance of Things Past*, if it has a philosophy, says more about the creative élan vital of Bergson, the human essences that underlie the shifting phenomena of time and space, than the social jungle the realists had taken for reality. André Gide (1869–1951) seems to make a plea for human aloofness from environment so that the essentially human capacity for change and growth may operate. André Malraux (1901–), the forerunner of the Existentialist novelists, demonstrated, in *Man's Fate*, the necessity for human involvement in action as the only answer to the absurdity of his position in a huge and indifferent or malevolent universe. Jean-Paul Sartre (1905–) and Albert Camus (1913–60) similarly emphasized man's freedom to choose, to say no to evil, to define himself through action.

The realist masters

Other French novelists have been more concerned with recording the minutiae of human life as a predominantly sensuous and emotional experience, like Colette (1873–1954), or with taking the religious sensibility as a fictional theme, like François Mauriac (1855–1970). The practitioners of the *anti-roman*—Butor, Robbe-Grillet, Nathalie Sarraute—pursue their attempts at liquidating human character in the traditional novelistic sense. Samuel Beckett (1906–), an Irishman who has turned himself into a major French stylist, goes his own way, presenting—in works like *Molloy, Malone Dies, Watt* and *The Unnamable*—mankind reduced to degradation and absurdity but somehow admirable because he survives.

French-Canadian fiction inevitably suffers from comparison with the glories of the mother country. Before 1900 there is little of value to record—except perhaps the historical romance of Philippe de Gaspé (1786–1871), *Les Anciens Canadiens*—but Louis Hémon (1880–1913) produced a genuine classic in *Maria Chapdelaine*, a story of Canadian pioneer life, original, moving, and sensitive. The somewhat provincial character of French-Canadian life, dominated by the Church and by outmoded notions of morality, has not been conducive either to fictional candour or to formal experiment, and metropolitan France is unimpressed, for the most part, by the literature of the separated brethren. But there is time for an avant-garde to develop and great masters to appear.

French-Canadian novels

Spanish. The great age known as "El Siglo de Oro," or Golden Age, produced what is conceivably (it must contest this claim with *War and Peace*) the most magnificent of all novels—the *Don Quixote* of Miguel de Cervantes Saavedra (1547–1616). A satire on chivalry which ends as a humane affirmation of the chivalric principle, it encloses—in tender or comic distortion—other fictional forms that flourished in the Spanish Golden Age. The *novela picaresca* fathered a whole European movement, and its best monuments are perhaps the anonymous *Lazarillo de Tormes* (1554), the *Guzmán de Alfarache* of Mateo Alemán (1547–?1614) and *The Limping Devil* (*El diablo cojuelo*) of Luis Vélez de Guevara (1579–1644). The pastoral novel, another popular but highly stylized form, was less true fiction than a sort of prose poem, in which lovers in shepherd's disguise bewailed an unattainable or treacherous mistress. But here, since the fictional lovers were often real-life personages in the cloak of a rustic name, the germs of the *roman à clef* are seen stirring. The *novela Morisca* was an inimitable Spanish form, a kind of fictional documentary about the wars between Christians and Muslims, as in *Guerras civiles de Granada* (*Civil Wars of Granada*) by Ginés Pérez de Hita (1544?–?1619).

The Golden Age in Spain

The decline of Spain as a European power is associated, in literature, with feeble nostalgia for the Golden Age or feebler imitations of French classicism. The Spanish novel began to re-emerge only in the early 19th century, when a kind of journalism celebrating regional customs encouraged the development of the realistic regional novel, with Fernán Caballero (1796–1877), Armando Palacio Valdés (1853–1938), and the important Vicente Blasco Ibáñez (1867–1928), whose *Sangre y arena* (*Blood and Sand*), *Mare nostrum*, and *Los cuatro jinetes del Apocalipsis* (*The Four Horsemen of the Apocalypse*) achieved universal fame and were adapted for the screen in the 1920s. The fiction of the so-called Generation of 1898—which took its name from the year of the Spanish-American War, a cataclysmic event for Spain that bereaved it of the last parts of a once great empire—wasted no time on national nostalgia or self-pity, but concentrated on winning a new empire of style. Ramón Mariá del Valle-Iclán (1866–1936) and Pérez de Ayala (1880–1962) brought a highly original lyricism to the novel, while Pío Baroja (1872–1956) concentrated on representing a world of discrete events, unbound by a unifying philosophy. The fiction that came out of the Civil War of 1936–39 returned to a kind of didactic realism, as with José Mariá Gironella (1917–), who depicted a ravaged and suffering Spain. Camilo José Cela (1916–), perhaps the most important modern Spanish novelist, combines realism with a highly original style. His *Familia de Pascual Duarte*—harrowing, compassionate, brilliantly economical—is a novel of towering merit.

Italian. Though Italy originated the *novella*, it was slow in coming to the full-length novel. There is little to record before *I promessi sposi* (*The Betrothed*) by Alessandro Manzoni (1785–1873), a romantic and patriotic novel that describes life in Italy under Spanish domination in the 17th century. That combination of regionalism and realism already noted in the fiction of Germany and Spain did not appear in Italy until after the unification in 1870, when Giovanni Verga (1840–1922) celebrated his native Sicily in *I malavoglia* (*The House by the Medlar Tree*) and Antonio Fogazzaro (1842–1911) showed life in northern Italy during the struggle for unification in *Piccolo mondo antico* (*Little World of the Past*). Gabriele D'Annunzio (1863–1938) and Luigi Pirandello (1867–1936) were too original for easy classification, and both worked in all the literary media. Pirandello, especially, helped to bring Italian literature into the modern world through such philosophical novels as *Il fu Mattia Pascal* (*The Late Matthew Pascal*), which raises very profound questions about the nature of human identity and yet contrives to be witty, sunny, and essentially Italianate. The importance of Italo Svevo (1861–1928) was obscured for some time because of the difficulty of Italian literati in accepting the Triestine dialect as a literary medium, but works like *Zeno* and *Senilitá* (a title which James Joyce, Svevo's friend and English teacher, translated *As a Man Grows Older*) are recog-

nized as being major contributions to the international novel.

The significant fiction of the Mussolini regime was produced by anti-Fascist exiles like Giuseppe Borgese (1882–1952) and Ignazio Silone (1900–), whose *Pane e vino* (*Bread and Wine*) is accepted as a 20th-century classic. Alberto Moravia (1907–), with his *Romana* (*Woman of Rome*) and *Noia* (*The Empty Canvas*), is perhaps the most popular Italian novelist outside Italy, but he is probably less important than Giuseppe Berto (1914–), Cesare Pavese (1908–1950), and Elio Vittorini (1908–1966). Giuseppe di Lampedusa (1896–1957) created a solitary masterpiece in *Il gattopardo* (*The Leopard*). The experimental writing of Italo Calvino and Carlo Emilio Gadda is becoming better known outside Italy, but, generally speaking, the Italian novel remains linguistically conservative and needs the impact of some new powerfully iconoclastic literary figure like James Joyce.

Scandinavian languages. Norway is better known for Ibsen's contribution to the drama and Grieg's to music than for fiction of the first quality. Nevertheless, the novels of Knut Hamsun (1859–1952) earned him the Nobel Prize in 1920, and Sigrid Undset (1882–1949) received the same honour in 1928, though the work of both has failed to engage the lasting attention of world readers of fiction. In Denmark, Johannes Vilhelm Jensen (1873–1950), another Nobel prizeman, Isak Dinesen, the pen name of Baroness Karen Blixen (1885–1962), and Martin Nexö (1869–1954) have contributed to their country's fictional literature but made little mark beyond. The achievement of Sweden's novelists is very inconsiderable, but Halldór Kiljan Laxness (1902–) has restored to Iceland the literary fame it once earned for its sagas.

Slavic and east European. Czechoslovakia's greatest novelist, Franz Kafka, wrote in German, but its writers in the vernacular include world-famous names, such as Jaroslav Hasek (1883–1923), whose *Good Soldier Schweik* is acknowledged to be a comic masterpiece, and Karel Čapek (1890–1938), best-known for the plays *R.U.R.* (which gave "robot" to the world's vocabulary) and *The Life of the Insects*, but also notable for the novels *The Absolute at Large*, *Krakatit*, and *War with the Newts*. The Czech fictional genius tends to the satirical and the fantastic. The fiction of Yugoslavia has made little impact on the progress of the novel, but Radomir Konstantinovič (1928–) should be noted: his *Exitus* is a remarkable study of the death of Christ. Poland can claim two Nobel prizewinning novelists in Henryk Sienkiewicz (1846–1916), the author of *Quo Vadis?*, and Władysław S. Reymont (1868–1925), whose novel *The Peasants* is an aromatic piece of bucolic realism. Witold Gombrowicz (1904–1969) wrote a novel, *Ferdydurke*, which has been subjected to two separate modes of suppression—first Fascist, later Communist. A remarkable surrealist essay on "anal tyranny" and depersonalisation, it still awaits the acclaim that is due. Romania, which produced outstanding novelists in E. Lovínescu and Titu Maiorescu, has suffered, like other Balkan countries, from the totalitarian suppression of the free creative spirit; but Dumitru Radu Popescu, author of *The Blue Lion*, was bold enough in the 1960s to question Communist orthodoxy through the medium of fiction. The Hungarian Gusztáv Rab (1901–1966) awaits recognition. His brilliant *Sabaria* develops very courageously the theme of the conflict between Communism and Christianity and reaches conclusions favourable to neither. It is a disturbing and beautifully composed book. Modern Greece, more famous for its poets George Seferis and Constantine Cavafy, has produced at least one major novelist in Nikos Kazantzakis (1885–1957), whose *Zorba the Greek* became famous through its film adaptation. *The Last Temptation*, which presents the life of Christ as a struggle to overcome "the dark immemorial forces of the Evil One, human and pre-human," glows with the writer's personality and bristles with the dialectic that was one of the first Greek gifts to Western civilization.

The Jewish novel. The literature of the Diaspora—the dispersion of the Jews after their exile from ancient Babylonia—records many large achievements in the languages of exile and that dialect of Low German—Yiddish—which the Ashkenazi Jews have taken around the world. Perhaps the most interesting of modern Jewish novelists in Yiddish is Isaac Bashevis Singer (1904–), an American who refuses to be absorbed linguistically into America, unlike Bellow and Malamud, who have brought to the Anglo-American language typical tones and rhythms of the ghetto. Israel, which is producing its own rich crop of national writers, began with an existing corpus of literature in modern Hebrew, a language promoted and nurtured by such scholars as Eliezer ben Yehuda (1858–1923) and the members of the neologizing Hebrew Academy of Israel. Among the early novelists is Abraham Mapu (1808–1867); among the later ones is the brilliant Moshe Shamir (1921–). The first Nobel Prize for Literature ever awarded to a Hebrew novelist went, justly, to Shmuel Yosef Agnon (1888–1970). The contemporary Hebrew novel is notable for a fusion of international sophistication and earthy homegrown realism.

ASIA, AFRICA, LATIN AMERICA

China. The tradition of storytelling is an ancient one in China, and the full-length novel can be found as far back as the late 16th century. The fiction of the 18th century shows a variety of themes and techniques not dissimilar to those of Europe, with social satire, chivalric romance, and adventure stories. Ts'ao Hsüeh-ch'in (died 1763) wrote a novel called *Hung lou meng* (translated as *The Dream of the Red Chamber* in 1892), which has features not unlike those of Galsworthy's *The Forsyte Saga* and Mann's *Buddenbrooks*—a story of a great aristocratic family in decline, garnished with love interest and shot through with pathos. The early days of the 20th century saw the foundation of the Chinese republic and the development of popular fiction in the national language (*pai-hua*). From 1917 until the Sino-Japanese War (1937–1945), a period of social and intellectual ferment, there was a great influx of Western novels, and, under their influence, movements like those devoted to realism and naturalism in Europe produced a great number of didactic novels. Novels like Mao Tun's *Rickshaw Boy* and Pa Chiu's *Chinese Earth* appeared in the postwar period, but the Communist regime has quelled all but propagandist fiction. Such novels as Liu Ching's *Wall of Bronze*, Chao Shu-li's *Changes in Li Village*, and Ting Ling's *Sun Shines over the Sangkan River* are typical glorifications of the Mao philosophy and the socialist achievements of the people.

Japan. A literature that admires economy, like the Japanese, is bound to favour, in fictional art, the short story above the full-length novel. Nevertheless, some of the ancient pillow-books, with their diary jottings and anecdotes, have the ring of autobiographical novels; while Murasaki Shikibu's *Tale of Genji*, produced nearly a thousand years ago, is a great and sophisticated work of fictional history. The 20th-century Japanese novel has been developed chiefly under Western influence, like the work of Akutagawa Ryūnosuke (1892–1927), whose *Yabunoaka* became the highly applauded film *Rashomon*. Tanizaki Jun-ichirō (1886–) is well known in America and Europe for *The Key*, *The Makioka Sisters*, and *Diary of a Mad Old Man*. His novels, all set in a modern, Westernized Japan, are assured of universal popularity because of their frank and lavish sexual content. Mishima Yukio (1925–1970), who committed ceremonial suicide at the height of his reputation, was perhaps the most successful, and certainly the most prolific, of all modern Japanese novelists. Ten of his fictions had been filmed; he had won all the major Japanese literary awards and appeared to be destined for the Nobel Prize. His works are characterized by ruthless violence and a perversity that, however much it seems to derive from the pornographic excesses of Western fiction, is certainly in the Japanese tradition. His work is hard to judge, but he was the most considerable literary figure of the East.

Western influences

The Indies. The Indian novel is a branch of British Commonwealth literature. That is to say, it is practiced by writers who have received a British literary education and, for the most part, publish their books in London.

There is no evidence of any great development of fiction in any of the native Indian tongues: a taste for reading novels, as opposed to seeing films or reading short stories in the Hindi or Punjabi press, is acquired in India along with an education in English, which is still the unifying tongue of the subcontinent. In Malaysia, where the same tradition holds among the literate, short stories are being written in Malay, Chinese, and Tamil, but the full-length novel is almost exclusively written in English. It is as much a matter of literary markets as of literary education. No novel in any of the tongues of the peninsula is likely to be a remunerative publishing proposition, as Han Suyin (1917–)—a best selling Chinese woman doctor from Johore, who found large fame with *A Many-Splendoured Thing*—would be the first to admit, for all her dislike of the British colonial tradition. Indonesia's abandonment of all vestiges of its Dutch colonial past is associated with the encouragement of a literature in Bahasa (a dialect of Malay), and there are a number of Indonesian novelists still looking for a large educated audience—inevitably in translation, in the West. Among these Mochtar Lubis (1922–) is notable; his *Twilight in Djakarta* is a bitter indictment of the Sukarno regime, which promptly sent him to prison. This was not the kind of fiction that the new Indonesia had in mind.

Africa. What applies to India and the East Indies applies also to Africa: there is still an insufficiently large audience for fiction written in any of the major African languages, and African novelists brought up on English are only too happy to continue working in it. Other European languages—chiefly Afrikaans (a South African variety of Dutch) and French—are employed as a fictional medium. Arthur Fula, who wrote *Janie Giet die Beeld* (*Janie Casts the Image*), is an Afrikaans novelist whose reputation has not, as yet, stretched to Europe or America, but the novelists of the former French possessions are gaining a name among serious students of the African novel. Mongo Beti (1932–), from the Cameroons, is known for his *Pauvre Christ de Bomba* and *Roi miraculé*; the Ivory Coast has Aké Loba, and Hamidou Kane (1928–) represents Senegal. This new African French deserves to be regarded as a distinct literary language, unrelated to that of Paris or Quebec, but the critical and linguistic tools for appraising the fiction of French-speaking Africans are not yet available.

Latin America. One of the greatest contemporary writers of fiction, Jorge Luis Borges (1899–), is an Argentinian, but his *ficciones* are very short short stories and must, with regrets, be excluded from any survey of the novel. It is significant, however, that the circumstances for the creation of a great fictional literature are in existence in Latin America, signifying (unlike the case in many former British dependencies) a shedding of the old colonial provincialism, which relied on the opinion of Madrid or, in Brazil, of Lisbon. The first Latin American novel was probably *The Mangy Parrot*, by the Mexican José Joaquín Fernández de Lizardi (1776–1827), a picaresque work satirizing colonial conditions. In Argentina, José Mármol (1817–1871) published the first major novel of the continent—*Amalia*, a powerful study of the fear and degradation that were rife in Buenos Aires during the dictatorship of the corrupt and tyrannical Rosas. The Romantic period in Europe had its counterpart in the sentimental wave that overtook such novelists as the Colombian Jorge Isaacs (1837–1895), while the new humanitarianism found a voice in *Birds Without a Nest*, a protest-novel on the conditions forced on the Indians of Peru, written by Clorinda Matto de Turner (1854–1901). Juan León Mera followed the same trend in *Cumandá*, a novel about the oppression of the Ecuadorian Indians. As Latin America moved toward the modern age, the inevitable novels of urban social protest made their appearance. Alberto Blest Gana (1830–1920) of Chile, Carlos Reyles (1868–1938) of Uruguay, and Gustavo Martínez Zuviría of the Argentine are names of some historical significance; and Reyles's naturalistic novel *La Raza de Caín* (*Cain's Race*) is original in that it finds a parallel between the breeding of stock and the building of a human society.

Rise of nativism

A 20th-century reaction against the bourgeois novel led to the movement known as nativism, with its concentration on the land itself, the lot of the indigenous peoples, the need for an anti-racist revolution with the aim of true egalitarianism. *The Underdogs*, by the Mexican Mariano Azuela (1873–1952), and *El Señor Presidente*, by Miguel Angel Asturias (1899–) from Guatemala, typify the new revolutionary novel. Much of the didactic energy that went to the making of such work resulted in a lack of balance between subject matter and style: the literature of revolt tends to be shrill and crude. The inevitable reaction to a more sophisticated kind of literature, in which the individual became more important than society and the unconscious more interesting than the operation of reason, led to the highly refined experimentation that is the mark of Borges, the Brazilian Érico Veríssimo, and Eduardo Barrios of Chile.

The fiction of Brazil is in many ways more interesting than the fiction of the mother country, which has produced only one major novelist in the realist José Maria de Eça de Queirós (1845–1900). Gregório de Matos Guerra, as early as the 17th century, wrote bitterly of colonial administration and painted a realistic picture of Brazilian life. Irony and keen observation have been characteristic of the Brazilian novel ever since—as in the *Brás Cubas* and *Dom Casmurro* of Joaquim Maria Machado de Assis (1839–1908); the *Canaan* of José Pereira da Graça Aranha (1868–1931), a remarkable study of the disillusion of German settlers in a new land; and the *Os Sertões* (*Rebellion in the Backlands*) of Euclides da Cunha (1866–1909), an account of a revolt against the newly formed Brazilian republic. The social consciousness of Brazilian novelists is remarkable, especially when it is associated with a strong concern for stylistic economy and grace. Monteiro Lobato (1882–1948), with his rustic hero Jeca Tatú; José Lins do Rêgo (1901–1957), chronicler of the decay of the old plantation life; Jorge Amado (1912–), a socialist novelist much concerned with slum life in Bahia; Érico Veríssimo (1905–), an experimental writer grounded in the traditional virtues of credible plot and strong characterization—these attest a vigour hardly to be found in the fiction published in Lisbon.

III. Social and economic aspects of the novel

Though publishers of fiction recognize certain obligations to art, even when these are unprofitable (as they usually are), they are impelled for the most part to regard the novel as a commercial property and to be better pleased with large sales of indifferent work than with the mere unremunerative acclaim of the intelligentsia for books of rare merit. For this reason, any novelist who seeks to practice his craft professionally must consult the claims of the market and effect a compromise between what he wishes to write and what the public will buy. Many worthy experimental novels, or novels more earnest than entertaining, gather dust in manuscript or are circulated privately in photocopies. Indeed, the difficulty that some unestablished novelists find in gaining a readership (which means the attention of a commercial publisher) has led them to take the copying machine as seriously as the printing press and to make the composition, mimeographing, binding, and distribution of a novel into a single cottage industry. For the majority of novelists the financial rewards of their art are nugatory, and only a strong devotion to the form for its own sake can drive them to the building of an *oeuvre*. The subsidies provided by university sinecures sustain a fair number of major American novelists; others, in most countries, subvent their art by practicing various kinds of subliterature—journalism, film scripts, textbooks, even pseudonymous pornography. Few novelists write novels and novels only.

Awards, patronage, and ancillary activities

There are certain marginal windfalls, and the hope of gaining one of these tempers the average novelist's chronic desperation. America has its National Book Award as well as its book club choices; France has a great variety of prizes; there are also international bestowals; above all, there glows the rarest and richest of all accolades—the Nobel Prize for Literature. Quite often the Nobel prizeman needs the money as much as the fame, and his

election to the honour is not necessarily a reflection of a universal esteem which, even for geniuses like Samuel Beckett, means large sales and rich royalties. When Sinclair Lewis received the award in 1930, wealth and fame were added to wealth and fame already sufficiently large; when William Faulkner was chosen in 1949, most of his novels had been long out of print in America (if not in Sweden).

Prizes come so rarely, and often seem to be bestowed so capriciously, that few novelists build major hopes on them. They build even fewer hopes on patronage: Harriet Shaw Weaver, James Joyce's patroness, was probably the last of a breed that, from Maecenas on, once intermittently flourished; state patronage—as represented, for instance, by the annual awards of the Arts Council of Great Britain—can provide little more than a temporary palliative for the novelist's indigence. Novelists have more reasonable hopes from the world of the film or the stage, where adaptations can be profitable and even salvatory. The long struggles of the British novelist T.H. White came to an end when his Arthurian sequence *The Once and Future King* (1958) was translated into a stage musical called *Camelot*, though by treating the lump sum paid to him as a single year's income instead of a reward for decades of struggle, nearly all the windfall would have gone for taxes if White had not taken his money into low-tax exile. Such writers as Graham Greene, nearly all of whose novels have been filmed, must be tempted to regard mere book sales as an inconsiderable aspect of the rewards of creative writing. There are few novelists who have not received welcome and unexpected advances on film options, and sometimes the hope of film adaptation has influenced the novelist's style. In certain countries, such as Great Britain but not the United States, television adaptation of published fiction is common, though it pays the author less well than commercial cinema.

When a novelist becomes involved in film-script writing —either in the adaptation of his own work or that of others—the tendency is for him to become subtly corrupted by what seems to him an easier as well as more lucrative technique than that of the novel. Most novelists write dialogue with ease, and their contribution to a film is mostly dialogue: the real problem in novel writing lies in the management of the *récit*. A number of potentially fine novelists, like Terry Southern and Frederic Raphael, have virtually abandoned the literary craft because of their continued success with script writing. In 70-odd years the British novelist Richard Hughes produced only three novels, the excellence of which has been universally recognized; fiction lovers have been deprived of more because of the claims of the film world on Hughes's talent. This kind of situation finds no counterpart in any other period of literary history, except perhaps in the Elizabethan, when the commercial lure of the drama made some good poets write poor plays.

The importance of the paperback

The majority of professional novelists must look primarily to book sales for their income, and they must look decreasingly to hard-cover sales. The novel in its traditional format, firmly stitched and sturdily clothbound, is bought either by libraries or by readers who take fiction seriously enough to wish to acquire a novel as soon as it appears: if they wait 12 months or so they can buy the novel in paper covers for at most a fifth of its original price. The paperback edition of a novel has become, for the vast majority of fiction readers, the form in which they first meet it, and the novelist who does not achieve paperback publication is missing a vast potential audience. He may not repine at this, since the quantitative approach to literary communication may safely be disregarded: the legend on a paperback cover—FIVE MILLION COPIES SOLD—says nothing about the worth of the book within. Nevertheless, the advance he will receive from his hard-cover publisher is geared to eventual paperback expectations, and the "package deal" has become the rule in negotiations between publisher and author's agent. The agent, incidentally, has become important to both publisher and author to an extent that writers like Daniel Defoe and Samuel Richardson would, if resurrected, find hard to understand.

The novelist may reasonably expect to augment his income through the sale of foreign rights in his work, though the rewards accruing from translation are always uncertain. The translator himself is usually a professional man and demands a reasonable reward for his labours, more indeed than the original author may expect: the reputations of some translators are higher than those of some authors, and even their names may be better known. Moreover, the author who earns most from publication in his own language will usually earn most in translation, since it is the high initial home sales that attract foreign publishers to a book. The more "literary" a novel is, the more it exploits the resources of the author's own language, the less likely is it to achieve either popularity at home or publication abroad. Best-selling novels like Mario Puzo's *Godfather* (1969) or Arthur Hailey's *Airport* (1968) are easy to read and easy to translate, so they win all round. It occasionally happens that an author is more popular abroad than he is at home: the best-selling novels of the Scottish physician-novelist A.J. Cronin are perhaps no longer highly regarded in England and America, as they were in the 1930s and 1940s, but they still sell by the million in the U.S.S.R. But a novelist is wisest to expect most from his own country and to regard foreign popularity as an inexplicable bonus.

Problems of obscenity and libel

As though his financial problems were not enough, the novelist frequently has to encounter those dragons unleashed by public morality or by the law. The struggles of Flaubert, Zola, and Joyce, denounced for attempting to advance the frontiers of literary candour, are well known and still vicariously painful, but lesser novelists, working in a more permissive age, can record cognate agonies. Generally speaking, any novelist writing after the publication in the 1960s of Hubert Selby's *Last Exit to Brooklyn* or Gore Vidal's *Myra Breckenridge* can expect little objection, on the part of either publisher or police, to language or subject matter totally unacceptable, under the obscenity laws then operating, in 1922, when *Ulysses* was first published. This is certainly true of America, if not of Ireland or Malta. But many serious novelists fear an eventual reaction against literary permissiveness as a result of the exploitation by cynical obscenity mongers or hard-core pornographers of the existing liberal situation.

In some countries, particularly Great Britain, the law of libel presents insuperable problems to novelists who, innocent of libellous intent, are nevertheless sometimes charged with defamation by persons who claim to be the models for characters in works of fiction. Disclaimers to the effect that "resemblances to real-life people are wholly coincidental" have no validity in law, which upholds the right of a plaintiff to base his charge on the corroboration of "reasonable people." Many such libel cases are settled before they come to trial, and publishers will, for the sake of peace and in the interests of economy, make a cash payment to the plaintiff without considering the author's side. They will also, and herein lies the serious blow to the author, withdraw copies of the allegedly offensive book and pulp the balance of a whole edition. Novelists are seriously hampered in their endeavours to show, in a traditional spirit of artistic honesty, corruption in public life; they have to tread carefully even in depicting purely imaginary characters and situations, since the chance collocation of a name, a profession, and a locality may produce a libellous situation.

EVALUATION AND STUDY

It has been only in comparatively recent times that the novel has been taken sufficiently seriously by critics for the generation of aesthetic appraisal and the formulation of fictional theories. The first critics of the novel developed their craft not in full-length books but in reviews published in periodicals: much of this writing—in the late 18th and early 19th centuries—was of an occasional nature, and not a little of it casual and desultory; nor, at first, did critics of fiction find it easy to separate a kind of moral judgment of the subject matter from an aesthetic judgement of the style. Such fragmentary observations on the novel as those made by Dr. Johnson in conversation

or by Jane Austen in her letters, or, in France, by Gustave Flaubert during the actual process of artistic gestation, have the charm and freshness of insight rather than the weight of true aesthetic judgment. It is perhaps not until the beginning of the 20th century, when Henry James wrote his authoritative prefaces to his own collected novels, that a true criteriology of fiction can be said to have come into existence. The academic study of the novel presupposes some general body of theory, like that provided by Percy Lubbock's *Craft of Fiction* (1921) or E.M. Forster's *Aspects of the Novel* (1927) or the subsequent writings of the critics Edmund Wilson and F.R. Leavis. Since World War II it may be said that university courses in the evaluation of fiction have attained the dignity traditionally monopolized by poetry and the drama.

The distinction between reviewing and criticism

A clear line should be drawn between the craft of fiction criticism and the journeyman work of fiction reviewing. Reviews are mainly intended to provide immediate information about new novels: they are done quickly and are subject to the limitations of space; they not infrequently make hasty judgments that are later regretted. The qualifications sought in a reviewer are not formidable—smartness, panache, waspishness, qualities that often draw the attention of the reader to the personality of the reviewer rather than the work under review—will always be more attractive to circulation-hunting editors than a less spectacular concern with balanced judgment. A thoughtful editor will sometimes put the reviewing of novels into the hands of a practicing novelist, who—knowing the labour that goes into even the meanest book—will be inclined to sympathy more than to flamboyant condemnation. The best critics of fiction are probably novelists *manqués*, men who have attempted the art and, if not exactly failed, not succeeded as well as they could have wished. Novelists who achieve very large success are possibly not to be trusted as critics: obsessed by their own individual aims and attainments, shorn of self-doubt by the literary world's acclaim or their royalty statements, they bring to other men's novels a kind of magisterial blindness.

Novelists can be elated by good reviews and depressed by bad ones, but it is rare that a novelist's practice is much affected by what he reads about himself in the literary columns. Genuine criticism is a very different matter, and a writer's approach to his art can be radically modified by the arguments and summations of a critic he respects or fears. As the hen is unable to judge of the quality of the egg it lays, so the novelist is rarely able to explain or evaluate his work. He relies on the professional critic for the elucidation of the patterns in his novels, for an account of their subliminal symbolism, for a reasoned exposition of their stylistic faults. As for the novel reader, he will often learn enthusiasm for particular novelists through the writings of critics rather than from direct confrontation with the novels themselves. The essays in Edmund Wilsons' *Axel's Castle* (1931) aroused an interest in the Symbolist movement which the movement was not easily able to arouse by itself; the essay on *Finnegans Wake*, collected in Wilson's *Wound and the Bow* (1941), eased the way into a very difficult book in a manner that no grim work of solid exegesis could have achieved. The essence of the finest criticism derives from wisdom and humanity more than from mere expert knowledge. Great literature and great criticism possess in common a sort of penumbra of wide but unsystematic learning, a devotion to civilized values, an awareness of tradition, and a willingness to rely occasionally on the irrational and intuitive.

All this probably means that the criticism of fiction can never, despite the efforts of aestheticians schooled in modern linguistics, become an exact science. A novel must be evaluated in terms of a firmly held literary philosophy, but such a philosophy is, in the final analysis, based on the irrational and subjective. If the major premises on which F.R. Leavis bases his judgments of George Eliot, Mark Twain, and D.H. Lawrence are accepted, then an acceptance of the judgments themselves is inescapable. But many students of fiction who are skeptical of Leavis will read him in order that judgments of their own may emerge out of a purely negative rejection of his. In reading criticism a kind of dialectic is involved, but no synthesis is ever final. The process of revaluation goes on for ever. One of the sure tests of a novel's worth is its capacity for engendering critical dialectic: no novel is beyond criticism, but many are beneath it.

THE FUTURE OF THE NOVEL

It is apparent that neither law nor public morality nor the public's neglect nor the critic's scorn has ever seriously deflected the dedicated novelist from his self-imposed task of interpreting the real world or inventing alternative worlds. Statistics since World War II have shown a steady increase in the number of novels published annually, and beneath the iceberg tip of published fiction lies a whole submarine Everest of unpublished work. It has been said that every person has at least one novel in him, and the near-universal literacy of the West has produced dreams of authorship in social ranks traditionally deprived of literature. Some of these dreams come true, and taxi drivers, pugilists, criminals, and film stars have competed, often successfully, in a field which once belonged to professional writers alone. It is significant that the amateur who dreams of literary success almost invariably chooses the novel form, not the poem, essay, or autobiography. Fiction requires no special training and it can be readable, even absorbing, when it breaks the most elementary rules of style. It tolerates a literary incompetence unthinkable in the poem. If all professional novelists withdrew their labour, the form would not languish: amateurs would fill the market with first and only novels, all of which would find readership somewhere.

Amateur novelists

But the future of any art lies with its professionals. Here a distinction has to be made between the Joyces, Henry Jameses, and Conrads on the one hand, and the more ephemeral Mickey Spillanes, Harold Robbinses, and Irving Wallaces on the other. Of the skill of the latter class of novelists there can be no doubt, but it is a skill employed for limited ends, chiefly the making of money, and through it the novel can never advance as art. The literary professionals, however, are dedicated to the discovery of new means of expressing, through the experiential immediacies that are the very stuff of fiction, the nature of man and society. In the symbiosis of publishing, the best seller will probably continue to finance genuine fictional art. Despite the competition from other art media, and the agonies and the indigence, there are indications that the serious novel will flourish in the future.

It will flourish because it is the one literary form capable of absorbing all the others. The technique of the stage drama or the film can be employed in the novel (as in *Ulysses* and *Giles Goat-Boy*), as can the devices of poetry (as in Philip Toynbee's *Pantaloon* and the novels of Wilson Harris and Janet Frame). In France, as Michel Butor has pointed out, the new novel is increasingly performing some of the tasks of the old essay; in America, as Capote's *In Cold Blood* and Mailer's *Armies of the Night* have shown, the documentary report can gain strength from its presentation as fictional narrative. There are few limits on what the novel can do, there are many experimental paths still to be trod, and there is never any shortage of subject matter.

For all this, periods of decline and inanition may be expected, though not everywhere at once. The strength of the American novel in the period after World War II had something to do with the national atmosphere of breakdown and change: political and social urgencies promoted a quality of urgency in the works of such writers as Mailer, Bellow, Ellison, Heller, and Philip Roth. In the same period, England, having shed its empire and erected a welfare state, robbed its novelists of anything larger to write about than temporary indentations in the class system, suburban adultery, and manners. An achieved or static society does not easily produce great art. France, which has known much social and ideological turmoil, has generated a new aesthetic of the novel as well as a philosophy that, as Sartre and Camus have shown, is very suitable for fictional expression. A state on which intellectual quietism or a political philosophy of art is imposed

by the ruling party can, as the Soviet Union and China show, succeed only in thwarting literary greatness, but the examples of Pasternak and Solzhenitsyn are reminders that repression can, with rare artistic spirits, act as an agonizing stimulus.

Every art in every country is subject to a cyclical process; during a period of decline it is necessary to keep the communication lines open, producing minor art so that it may some day, unexpectedly, turn into major art. Wherever the novel seems to be dying it is probably settling into sleep; elsewhere it will be alive and vigorous enough. It is important to believe that the novel has a future, though not everywhere at once.

BIBLIOGRAPHY. The following works deal in general terms with the reader's approach to the novel: WALTER ALLEN, *Reading a Novel,* rev. ed. (1963); VAN METER AMES, *Aesthetics of the Novel* (1928, reprinted 1966); CLEANTH BROOKS and R.P. WARREN (eds.), *Understanding Fiction,* 2nd ed. (1959); ALEXANDER COMFORT, *The Novel and Our Time* (1948); PELHAM EDGAR, *The Art of the Novel* (1933, reprinted 1966); WILSON FOLLETT, *The Modern Novel: A Study of the Purpose and Meaning of Fiction,* rev. ed. (1923); E.M. FORSTER, *Aspects of the Novel* (1927, many reprintings); PERCY LUBBOCK, *The Craft of Fiction,* new ed. (1957).

The following are concerned with the problems of writing fiction and are all the work of novelists: PHYLLIS BENTLEY, *Some Observations on the Art of Narrative* (1946); *Conrad's Prefaces to His Works,* with an essay by EDWARD GARNETT (1937); HENRY JAMES, *The Art of Fiction and Other Essays,* ed. by MORRIS ROBERTS (1948), and *The Art of the Novel,* introduction by R.P. BLACKMUR (1934); EDITH WHARTON, *The Writing of Fiction* (1925); THOMAS WOLFE, *The Story of a Novel* (1936).

The various elements of the novel are dealt with in the following: BONAMY DOBREE, *Modern Prose Style,* 2nd ed. (1964); MAREN ELWOOD, *Characters Make Your Story* (1942); MANUEL KOMROFF, *How to Write a Novel* (1950); W. VAN O'CONNOR (ed.), *Forms of Modern Fiction* (1948); GEORGE G. WILLIAMS (ed.), *Readings for Creative Writers* (1938).

The following studies deal with the style and philosophy of the novel in the wider sense: DAVID DAICHES, *The Novel and the Modern World,* rev. ed. (1960); AGNES HANSEN, *Twentieth Century Forces in European Fiction* (1934); ALFRED KAZIN, *On Native Grounds* (1942); Y. KRIKORIAN (ed.), *Naturalism and the Human Spirit* (1944); GEORGE LUKACS, *Studies in European Realism* (1950); H.J. MULLER, *Modern Fiction* (1937).

(An.B.)

Novosibirsk

One of the most important cities of the Soviet Union, Novosibirsk is the administrative centre of Novosibirsk *oblast* (province) of the Russian Soviet Federated Socialist Republic and in many respects the capital of all Siberia. Its nodal position on the Trans-Siberian Railroad has caused the town to expand very greatly and rapidly, in step with the economic development of Siberia as a whole. With a population of 1,286,000 in 1976, Novosibirsk is the eighth largest city of the Soviet Union and outstanding both for industrial production and for its educational and scientific research facilities. The town stands on the Ob River, here about half a mile wide, where it is bridged by the Trans-Siberian Railroad just below the confluence of the Inya River. The relatively flat plateau on which Novosibirsk is built is deeply cut by ravines, the continued growth of which presents the city with a major problem; already ravines occupy nearly 15,000 acres (6,000 hectares) of the city territory. Novosibirsk *oblast* covers 68,800 square miles (178,200 square kilometres) and in 1975 had a population of 2,543,000; apart from Novosibirsk itself, it is mainly an agricultural area, with small local urban centres.

Climate

Novosibirsk has an extreme continental climate with a very severe winter. The mean temperature in January is 3.2° F (−16° C), in July 65.8° F (18.8° C). The highest summer temperatures reach 98.6° F (37° C), and in winter the temperature has been known to drop to −58° F (−50° C). In early winter, strong and bitterly cold winds are common. Annual precipitation averages 19.5 inches (495 millimetres), but it is extremely variable from year to year; more than half the total falls in summer, frequently in the form of torrential thunderstorms and hailstorms.

HISTORY

Novosibirsk might well be described as an offspring of the railway. In 1891 the engineer N.G. Garin-Mikhaylovsky, surveying the route for the Trans-Siberian Railroad, chose the small village of Krivoshchekovo on the Ob as the site for the bridging of the river. In 1893 work on the bridge was begun, and beside it a small settlement grew up. By 1897 the bridge was completed, and the settlement had developed into a township of 7,832 people. A landing was built on the Ob, and transshipment and river trade developed briskly. The new settlement was known variously as Gusevka or Aleksandrovsky, but in 1895 it was renamed Novonikolayevsk in honour of the accession of Tsar Nicholas II.

The next two decades saw steady growth, based chiefly on movement of freight by rail and river and on the locomotive depot. A number of small industrial enterprises were established, using local products and including two saw mills, seven flour mills, a tannery, a distillery, a soap works, and two iron-casting works—the beginnings of what was to become a major metallurgical and engineering industry. Shops, warehouses, a small hospital, and a cathedral were built, and in 1903 formal town status was conferred. By the time of the Revolution of 1917 the population had reached 69,000 within the old city boundaries. During the subsequent civil war, Novonikolayevsk was occupied in turn by the independent Czech brigade, the White Army of Adm. A.V. Kolchak and finally, at the end of 1919, by the Red Army. In 1925 the town was renamed Novosibirsk (meaning New Siberia).

The opening up of the mineral and timber wealth of Siberia in the Soviet period, especially after the beginning of the five-year plans in 1928, brought about a remarkably rapid growth of the city. In particular the development of the Kuznetsk (Kuzbass) Basin coalfield about a hundred miles to the east of the town, where mining had begun in the late 19th century, and the establishment there of a large-scale metallurgical industry, fostered the importance of Novosibirsk as a transit point. In 1934 a railway was completed directly linking the town to Novokuznetsk (then called Stalinsk), the principal town on the coalfield. A second factor in the growth of Novosibirsk was the completion in the early 1930s of the Turkistan–Siberian (Turksib) Railway, providing a direct link with Tashkent and Central Asia. Large new factories were established, and the town grew to 120,100 population in 1926 and to 405,600 in 1939, an average annual increment of almost 22,000, mostly by in-migration from other parts of the country.

Kuzbass

World War II, with its eastward evacuation of people and industrial plant from the war zone, still further stimulated growth. By 1959 the population was 886,000, and in 1963 the city passed the million mark. Although growth in the postwar years has been continuous, its rate has slackened; average growth between 1959 and 1970 was 3 percent per year, compared with 6 percent between 1939 and 1959 and 18 percent between 1926 and 1939. In terms of absolute increment there has been relatively little change, 24,000 annually between 1939 and 1959 and 25,000 between 1959 and 1970. To some extent Novosibirsk is affected by the high rate of labour turnover and out-migration that has been characteristic of most of Siberia since about 1960, although less acutely so than elsewhere; many migrants to the city are young and unskilled and leave after only a short stay, giving rise to problems of establishing a stable and skilled labour force.

THE CONTEMPORARY CITY

Administration and setting. Novosibirsk is designated an *oblast* town; that is, its town council is subordinated to the Novosibirsk *oblast* council, although both administrations are located in the town itself. The city limits enclose an area of 184 square miles (477 square kilometres), the third largest city territory in the Soviet Union, although this includes considerable areas of farmland and forest beyond the built-up area. Originally only on the right bank of the Ob, the town now extends to either side of the river and comprises nine administrative wards. Central Ward, on the right bank, is the oldest part

The opera and ballet theatre, Novosibirsk.
Sovfoto

of Novosibirsk and contains most of the principal public buildings, many constructed in the ponderous architectural style of the Stalin era, with much use of locally quarried gray granite. These include the offices of the city council and the local Communist Party organization and various branches of all-union and republican ministries, concerned with the economic development of West Siberia as a whole. In this central area some buildings of the pre-Revolutionary period survive, but, because of its relative youth, the city lacks buildings of historic interest. Most of the theatres and cultural establishments and the largest shops are in the centre. Krasny Prospekt (Red Avenue), the principal street, bisects the centre, running north from the Ob to the local airport and linking the town's three largest squares, Kalinin and Lenin squares and Square of the Soviets. As in most Soviet cities, multistory apartment buildings with two- and three-room flats account for high population densities even in the most central parts.

Although industry is found in all parts of the city, the principal industrial sectors are in the east along the Trans-Siberian line, and on the opposite, left bank of the Ob, where are located the largest engineering and metallurgical factories. An imposing road bridge over the Ob links Kirov Ward to the rest of the city. North of the centre is found a large range of light industrial and food-processing plants. Housing everywhere is in very large apartment buildings. Those in central areas are usually five to seven stories high and generally constructed of stone or brick; those in outer parts of the built-up area, prefabricated buildings constructed in the later 1960s and '70s, are frequently 12 to 20 stories high, giving population densities as great as in the centre. The apartment buildings are grouped in "micro-regions," or neighbourhood units, each of which is provided with basic services such as shops, schools, a polyclinic, and a cinema.

Economy. Novosibirsk is one of the most important manufacturing centres in the Soviet Union. Although there is a wide range of industries, metallurgy and engineering predominate, employing two-thirds of all industrial workers in Novosibirsk. The old, pre-Revolutionary iron industry has been transformed into the modern Kuzmin steelworks, producing hot and cold rolled steel, cold rolled sheet steel, and steel tubes. The tin smelter is the largest in the country, utilizing tin concentrates from distant sources in northern Yakutiya, Transbaikalia, the Far East, and Kirgiziya. A highly specialized metalworking plant is the gold refinery, which refines all gold mined in the Soviet Union. In the range of engineering works, pride of place is taken by the Yefremov heavy machinery and hydraulic press factory and by the large electrical generator plant. Other plants manufacture electrothermal equipment in the form of steel furnaces of 180 tons' capacity, ore concentrating and mining machinery, and agricultural machinery (including disk harrows, seeding and husking machines, and tractor spares). Precision and light engineering plants make machine tools, instruments, radios, and automatic looms. Also part of this vast engineering complex are the big locomotive repair and servicing shops and the ship repair yards on the Ob.

Industries

During the 1960s and '70s the chemical industry developed rapidly, producing synthetic resin, plastics, and pharmaceutical goods. Other industries are primarily concerned with supplying the city and surrounding area with construction materials and with consumer goods—furniture, pianos, shoes, textiles, knitwear, and foodstuffs. The early-established flour-milling industry continues.

Power for the industries is supplied by a 220-kilovolt line from the Kuzbass coalfield and by the Novosibirsk hydroelectric station, constructed between 1951 and 1959 on the Ob some 15 miles (24 kilometres) upstream from the city. The barrage is three miles long and impounds a reservoir that has flooded more than 400 square miles (1,000 square kilometres) and that extends 110 miles upstream. The station has seven generators with a total capacity of 400,400 kilowatts. There are also two thermal power stations in the city itself. Novosibirsk is on the trunk petroleum pipeline that crosses Siberia from the Urals to Angarsk.

Electricity

Communications and transport. In addition to the trunk railway services via the Trans-Siberian, Kuzbass, and Turksib lines, which have played such a major part in the city's growth, local electric commuter trains link the suburbs to the city centre. Up to 70,000 passengers daily use Novosibirsk Main station alone. In First of May Ward, in the southeast sector of the town, is one of the largest marshalling yards in the country. There are two airports, a smaller one serving local air connections and a large main airport with direct flights to Moscow and other major cities of the Soviet Union. The river port has been enlarged and modernized with electric gantry cranes, and in the ice-free season it handles bulk freight moving on the Ob; this is made up chiefly of timber and building materials, with lesser quantities of coal, cement, fuel oil, and grain. There are also passenger services along the Ob, and small craft, including hydrofoils, link the parts of the town along the river. Transportation within the city is by bus, streetcar, and trolleybus.

Culture and education. Novosibirsk is the principal cultural and educational centre in Siberia. Its six theatres

include an imposing opera and ballet theatre, begun in 1931 and opened in 1945, seating 2,000 persons, the Red Torch Drama Theatre, a children's theatre, and a circus. There are botanical gardens, an art gallery, and four museums, as well as a symphony orchestra. The city has a television and radio centre, linked to Moscow (1,752 miles away) by communications satellite. It is an important publishing centre, with the Siberian branch of the Nauka (Science) Publishing House and the West Siberian Publishing House. Several local newspapers and journals are published. The State Public Scientific and Technical Library of the Academy of Sciences of the U.S.S.R. is one of the copyright libraries of the Soviet Union, with a collection of more than 5,000,000 books and reading rooms that can seat 1,000 persons.

Novosibirsk has some 220 general schools and 37 specialist schools—music, ballet, languages, science, and technology. There are 14 higher educational institutions, headed by the Novosibirsk State University, founded in 1959; other higher educational establishments include railway engineering, electrotechnical, medical, agricultural, and teacher-training institutes. The university and a number of these institutes are located in one of the most remarkable developments of Novosibirsk. This is the satellite town of Akademgorodok (Academic Town), which lies 15 miles south of the city centre on the shores of the Novosibirsk reservoir but within the city limits in Soviet Ward. Construction of this suburb was begun in the 1950s to house the Siberian Department of the Academy of Sciences of the U.S.S.R., established in 1957, initially under the energetic directorship of M.A. Lavrentyev. Since then a very large concentration of research and educational institutions has grown up in Akademgorodok. Among these are most of the 22 specialist research institutes that the Academy of Sciences maintains in Novosibirsk, with staffs totalling about 11,500 persons, of whom more than 3,000 are research workers. In Akademgorodok also are the Siberian branches of the All-Union Agricultural Academy, constructed in the early 1970s, and the Academy of Medical Sciences. This concentration of scientific research workers and higher educational teachers has attracted a great deal of attention, both within the Soviet Union and internationally. A feature of the town is the special scientific boarding school for specially gifted children. Akademgorodok is regarded as a prototype for further academic settlements planned for Leningrad, Vladivostok, and other Soviet cities.

Akademgorodok

Health and recreation. Each micro-region of the town has a polyclinic, providing day to day medical, dental, and public health services. The city is adequately supplied with hospitals, including a large unit of 1,500 beds, built in the 1970s. There are several parks, the principal one of which lies along the right bank of the Ob in the north of the city. In the outer rural fringe are children's holiday camps, while the reservoir and its shores are much used for recreation, with boating and swimming facilities. (R.A.F.)

Nuclear Fission

Under certain circumstances it is possible for the atomic nucleus (*i.e.*, the heavy mass at the centre of an atom) to split into two pieces of almost equal mass with the simultaneous liberation of a large amount of energy: this phenomenon is called nuclear fission. Each of the two nuclear fragments will consist of protons (nuclear particles of positive charge) and neutrons (nuclear particles of no charge) that comprised the original nucleus. Normally, only heavy nuclei, such as uranium, can be induced to fission.

GENERAL CONSIDERATIONS

Stages of nuclear fission. The sequence of events that occurs in fission is shown in six stages, for convenience, in Figure 1. The heavy nucleus receives excitation energy as the result of a nuclear reaction having taken place, such as the absorption of a neutron from outside the atom (Stage 1). The excess energy added by the neutron causes the nucleus to undergo rapid changes of shape, oscillating from one form to another until finally it assumes the elongated shape called the transition state nucleus, which is a critical part of the fission process (Stage 2).

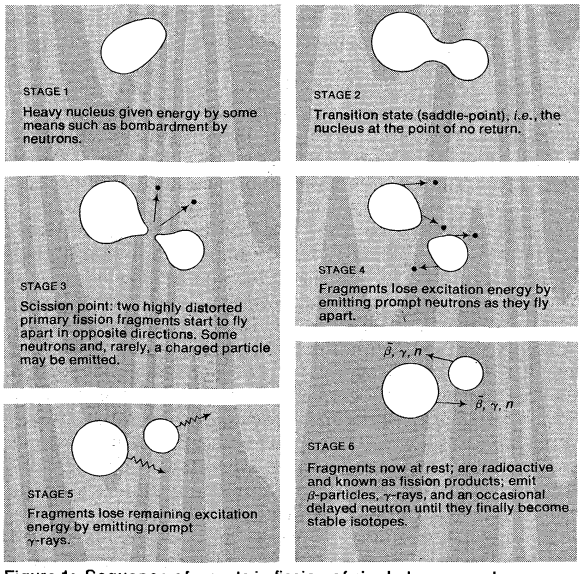


Figure 1: Sequence of events in fission of single heavy nucleus. Time (in seconds) between stages: 1–2, 10^{-15}; 2–3, 10^{-20}; 3–4, 10^{-15}; 4–5, 10^{-11}; 5–6, 10^{-1} up to 10^{7} or more.

At this point of its history, also referred to as the saddle point, the various forces within the nucleus are momentarily in balance. If the nucleus contracts slightly, it will return to its oscillations and will probably rid itself of its excess energy by emitting gamma rays (electromagnetic radiation of high energy). On the other hand, if it stretches a little more, the internal disruptive forces gain control and the nucleus is irrevocably committed to fission and will follow through the rest of the stages in the figure.

At the splitting or scission point (Stage 3) it will finally break into two large pieces, called fission fragments, and it may also emit several neutrons. Because both fission fragments are positively charged, and thus have a mutual electrostatic repulsion, they will fly apart in opposite directions in much the same way as the like poles of two magnets repel each other. Besides their kinetic energy (energy due to their motion), the fragments will also have a large excess of internal excitation energy (analogous to heat energy), most of which they will next lose, first by boiling off prompt neutrons (Stage 4), and after this by emitting prompt gamma rays (Stage 5). The term prompt is used to differentiate these rays from those of delayed emissions that occur at a later stage.

The process from the original excitation of the nucleus to the emission of the gamma rays will take about 10^{-11} second. These fragments have now come to rest and will have become radioactive elements listed in the middle of the periodic table, known as fission products (Stage 6). These then proceed to emit gamma rays, X-rays, and beta rays (consisting of negatively charged electrons that are produced when neutrons turn into protons) and occasionally a delayed neutron over a period of time that may be only a few seconds or may be many years, until they finally have no more excess energy left to disintegrate further, and turn into ordinary stable isotopes. (An isotope is a stable or radioactive atom of specified mass of a given element; an element may have several isotopes. A nuclide, on the other hand, is a general term referring to an atom of any mass and atomic number.)

Fission products

The importance of nuclear fission. The supreme importance of the fission reaction lies in the enormous amount of energy that it releases from the atom, about 2×10^{8} electron volts for each fission, equivalent to around 600,000 kilowatt-hours from the fission of one ounce of uranium-235, the uranium isotope with an atom-

ic mass number of 235. The object of nuclear weapon design is to create an uncontrolled chain reaction (see below) so as to cause as many atoms as possible to undergo fission in a short space of time. The tremendous energy from the fission of one nucleus is thus multiplied a large number of times, and it is this sudden surge of energy released almost instantaneously that constitutes the explosion. A fission explosion is about 400,000 times as powerful as the explosion of an equivalent weight of the chemical explosive TNT. The bombs exploded at Hiroshima and Nagasaki in World War II were fission bombs of this kind, but in 1952 the U.S. fired a nuclear weapon of another type, called a hydrogen or thermonuclear bomb. Whereas in fission the energy comes from the breakup of a nucleus, in this kind of reaction it comes as a result of the fusing together of two light nuclei to form a heavier one (see NUCLEAR FUSION). Nuclear explosives may be used for peaceful purposes, and both the United States and the Soviet Union have research programs that are investigating their possible advantages for such operations as excavation of harbours and canals and for other large earth-moving projects.

The enormous energy of fission may also be employed constructively in the generation of electric power, by harnessing the heat of the reaction to drive turbines. To do this it is necessary that the chain reaction proceed slowly and in a completely controlled manner, a situation that is achieved in a nuclear reactor. In many parts of the world reactors today can supply electricity more cheaply than any other fuel, and they are generating a larger fraction of the world's domestic and industrial power every year (see NUCLEAR REACTOR).

Nuclear power for submarines

Nuclear reactors are also used for propulsion of ships. Many nuclear-powered submarines are in service in the navies of the world, their supremacy over conventional types arising from the fact that they do not run on batteries that require recharging. Since they need only refuel with fissionable material at intervals of several months, they can remain submerged for weeks at a time when necessary. For economic reasons, nuclear-powered surface ships have been less successful, but several countries including the United States, the Soviet Union, and Japan have built experimental ones.

The other important application of nuclear reactors is in the production of copious numbers of neutrons. The majority of radioactive isotopes used extensively in medical treatment, in research, and in industry (see RADIOISOTOPES, APPLICATIONS OF) are made by bombarding nonradioactive substances with neutrons in reactors. Neutrons from nuclear reactors also serve an invaluable purpose in research and development work, which is designed to increase the economic potential of nuclear power.

History of fission research and technology. The discovery of fission was made by two German chemists, Otto Hahn and Fritz Strassmann, in 1938. They and other scientists had been trying to make transuranic elements (*i.e.*, elements that are heavier than uranium, then the heaviest known element in the periodic table) by bombarding uranium with neutrons, when, in a series of experiments, they proved conclusively that the products of their bombardments were not transuranic elements, but elements from the middle of the periodic table. Papers published in 1939 by the theoretical physicists Yakov Frenkel of the Soviet Union, Niels Bohr of Denmark, and John A. Wheeler of the United States provided a description of fission that to this day remains the basis of its theoretical treatment.

Enrico Fermi, an Italian physicist, then suggested that neutrons might be emitted in the fission reaction and that they in turn could induce fission reactions; and if this were so, it should be possible to sustain a chain reaction in uranium that, under certain conditions, might multiply fast enough to cause a nuclear explosion.

First nuclear reactor

A controlled chain reaction was first successfully initiated in the world's first nuclear reactor, set up by Fermi at the University of Chicago in 1942.

Bohr and Wheeler had suggested that only a small amount of the uranium in the natural element was undergoing fission, specifically, the isotope uranium-235, which is present only to the extent of one part in 140. This was confirmed by experiment and it was also discovered that the other isotope, uranium-238, when bombarded by neutrons, gave rise to the transuranic element neptunium, which, upon disintegrating radioactively, produced still another transuranic element, also fissionable or fissile like uranium, which was then named plutonium. The chain of nuclear reactions leading to its formation can be expressed as follows:

$$^{238}\mathrm{U} + n(\text{neutron}) \longrightarrow {}^{239}\mathrm{U} \longrightarrow {}^{239}\mathrm{Np}(\text{neptunium}) \longrightarrow {}^{239}\mathrm{Pu}(\text{plutonium}).$$

(^{239}U: emits a β^- particle; ^{239}Np: emits a β^- particle)

From the experiment it was clear that whereas natural uranium was satisfactory for the production of a slow chain reaction, either uranium-235 or else plutonium-239 would be needed, and in kilogram amounts, to make a bomb. Both these isotopes were produced. The uranium-235 was used for the bomb exploded over Hiroshima and the plutonium-239 for the one exploded over Nagasaki.

After the war major efforts in fission research continued in many countries and resulted in such activities as the development of the fusion bomb, economic nuclear power stations, and isotopes for medical research. Current research is concerned largely with the continued development of reactor systems and thermonuclear bombs. It is also concerned with finding more uses for radioactive materials, and with ever more complex experiments designed to elucidate the complicated mechanism of the basic fission process.

PHENOMENON OF NUCLEAR FISSION

General principles. *Types of fission and why it occurs.* It is a fundamental principle of nature that, if there is nothing to prevent it, every system will arrange itself in such a way that its potential energy will be reduced as much as possible. A simple example of this is water flowing down a hill. The potential energy of the system is high when the water is at the top of the hill but, in moving down, energy is released in the form of kinetic energy, so that when the water reaches the bottom of the hill the potential energy of the system has been greatly reduced.

Atomic nuclei behave in an analogous way. Albert Einstein in 1905 showed that mass and energy are merely different aspects of the same thing and that when mass is converted into energy, the energy is equal to the product of the mass and the velocity of light squared (*i.e.*, by the equation $E = mc^2$, in which E is the energy, m is the transformed mass, and c is the velocity of light). The masses of atomic nuclei have been carefully measured, and reference to these measurements will show that if any heavy nucleus breaks into two pieces of comparable size, the total mass of those pieces will be less than that of the original nucleus. This is to say, the difference in mass will be converted to energy that will be released in the breakup, and that the system after fission will be at a lower energy level than it was before. As a matter of fact, application of this mass test shows that all nuclei heavier than iron in the periodic table are potentially able to undergo fission in this way. All nuclei do not undergo fission at once because there is in the nucleus a potential hump, the equivalent of a dam that prevents the water from flowing down the hill. Called the fission barrier, it exists because of the very different nature of the two kinds of force in the nucleus, the nuclear force, and the electrostatic or coulomb force. The nuclear force is attractive and acts between both types of nucleons—neutrons and protons—but is short ranged so that a nucleon can act only on its immediate nucleon neighbours; the coulomb force, on the other hand, acts between the positively charged protons only, and extends throughout and beyond the whole nucleus (see NUCLEUS, ATOMIC). The lower the fission barrier, the less energy the nucleus requires for it to be overcome, and hence for fission to occur. Since the barrier gets lower as one moves up the periodic table, the trend is for nuclei to be more easily fissioned the heavier they are.

There are two ways, in fact, in which the fission barrier can be overcome. Carrying on the simple analogy, a little energy could be expended in lifting water confined by a dam up over the top of the dam, or perhaps the water might escape through a hole in the side of the dam. In induced fission, the nucleus receives energy from an external source: it might, for example, absorb a neutron, in which case it takes over both the rest mass energy and the kinetic energy of the neutron, and this extra excitation energy may be sufficient to lift it over the barrier and start the sequence of events shown in Figure 1. Fission can be induced in this way, not only by a neutron but also by other particles such as protons, alpha particles (consisting of two protons and two neutrons, identical to a helium nucleus), and gamma rays.

Spontaneous fission

In spontaneous fission a quantum mechanical effect, just like a hole in the dam (see MECHANICS, QUANTUM), allows the fission barrier of the nucleus to be penetrated, which offers another way that the fission sequence takes place, but this time without the outside intervention of a neutron as in induced fission. Spontaneous fission is not too likely, except in some artificially produced heavy isotopes: for example, the half-life (the time required by a large number of atoms for half their number to disintegrate; half-life is different for every nuclide) of the common uranium isotope uranium-238 for spontaneous fission is about 10^{16} years compared to its half-life for alpha decay of only just over 10^{9} years.

Fission cross section

Besides the fission barrier, there is one other important parameter of the fission reaction that must be understood: this is the cross section. In a situation in which a neutron moves toward a heavy nucleus, the cross section for fission is in effect simply the probability that fission will, in fact, take place. The nucleus may be thought of as a target with a certain effective area over which the fission will certainly occur if the neutron strikes within it. The area is not usually the same as the actual geometrical area of the nucleus; indeed, it may be hundreds of times greater or smaller, but it is determined by a variety of complex considerations that are outside the scope of this article. It is measured in units called barns, one barn being equal to 10^{-24} square centimetre.

Fission fragments. Examination of Figure 1 will show that fission of a nucleus can result in the production of fission fragments, charged particles, gamma rays, and neutrons. Since it is not possible to observe the fission process actually occurring, information about it has to be gleaned by experimenting on these diverse products. The fission fragments themselves have proved to be the most fruitful source of information, partly because they are the most important fraction of the debris, and partly because they are the easiest to examine.

There are two quite different approaches to the study of fission fragments. The most direct of these is to catch the fragments in flight and measure either the distribution of their kinetic energies or their angular distribution; that is, to measure the relative numbers of fragments for each increment of kinetic energy or the relative numbers for each direction. This sort of experiment is called on-line (from computer terminology) because the measurements are made during the actual bombardment of the fission target immediately after the emission of the prompt gamma rays. The measurements are taken earlier in the fission process than in the case of the off-line experiment, described below, in which radiochemistry is employed. The kinetic energy distributions give information about the way in which excitation energy is used in the fissioning nucleus, the excitation energy of the fragments themselves, the neutron emission of the fragments, and the way in which the total mass of the nucleus is split up. Angular distributions tell something about the possible energy levels (see NUCLEUS, ATOMIC) of the fissioning nucleus while in its transition stage (Stage 2).

In radiochemical off-line experiments the target is bombarded until sufficient fission has occurred, when the bombardment is stopped, and the target investigated. The fission process will thus have entered its final stage, in which the fragments have become fission products. When the nucleus is torn apart the fragments lose many of their atomic electrons (see ATOMIC STRUCTURE), but these are restored to them as the nuclei come to rest. The chemical properties that characterize the elements of the periodic table depend on these electrons, and so they have now become isotopes of middle-mass elements that can be identified by their radioactive properties and, if required, separated from each other by radiochemical means. They still have not quite reached their lowest possible energy state, however, and so they continue to lose energy by emitting beta particles, gamma rays, and occasionally a delayed neutron. Each time a negative beta particle is ejected, a rearrangement occurs within the nucleus itself and, whereas the mass changes only by a very small amount, the positive charge increases by one unit, and thus the atomic number increases by one and the fission product nucleus moves up one place in the periodic table. This process of beta decay continues until all the excess energy is lost, at which time the nucleus is termed stable. The half-lives of the nuclei that disintegrate or decay by beta emission in these decay chains range from about one second to many years and tend to get longer and longer towards the end of the chain, thus making it easier to do the radiochemistry on the decay products. This sort of experiment is less direct than the on-line type but it permits a greater variety of measurements and leads to positive identification of the isotopes that are produced.

Decay chains

Examples of the shape of the mass distributions of the fission products for two different kinds of fission are given in Figure 2. One curve represents the yield for radium-226 and the other for uranium-235. These distributions

From (top) I.F. Croall, Atomic Energy Research Establishment, Harwell, Berkshire, report R5086 (1967), (bottom) J. Milton and J. Fraser, *Proceedings of the Symposium of the Physics and Chemistry of Fission* (1965)

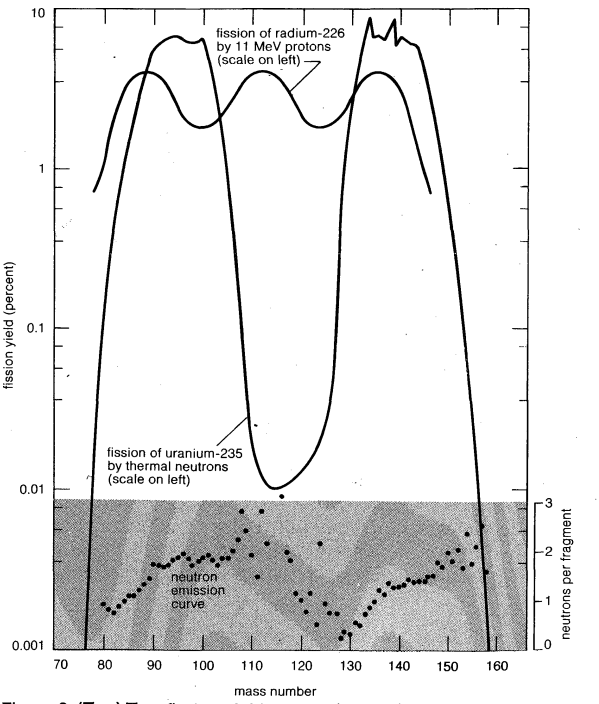


Figure 2: (Top) Two fission yield curves. (Bottom) Shaded area shows prompt neutron emission from uranium-235 thermal neutron fission.

show the percentage of fissions that result in the formation of fission products having any particular mass number, plotted against the mass numbers. These percentages are called fission yields. A glance at the curve for uranium-235 will show, for example, that about 6 percent of all the fissions in a quantity of uranium-235 will give rise to fission products of mass 99 and mass 133, or for 1 percent, masses of 84 and 149, or 105 and 128 (the total must be less than 235). The shapes of the curves reveal a great deal of information about the fission process, and a useful theory of fission must be able to explain the considerable differences in their shapes.

Another distribution that can be examined by radiochemical means, but only in a limited way by on-line

experiments, is the charge distribution. The nuclear charge of the different products along any fission decay chain increases by one unit with each beta decay. The charge distribution gives the probabilities of formation of all the members of a chain in the original fission act; and hence it gives clues as to what actually happened during this act.

Fission neutrons, gamma-rays, and other particles. Prompt neutrons, after fragments, are the most important product of fission because it is they that link successive fissions into the chain that is fundamental to the use of atomic energy in any form. Of these neutrons, 80 to 90 percent come off in Stage 4 of Figure 1 as part of the process that removes the excitation energy from the fission fragments, whereas the remainder are released slightly earlier, in Stage 3.

The first essential piece of information that is needed in a study of prompt particles emitted is the average number of prompt neutrons that are emitted per fission (always designated by the symbol $\bar{\nu}$). When neutrons are emitted, each carries away, on the average, about 7×10^6 electron volts of the excitation energy of the fragment, of which about 2×10^6 electron volts is kinetic energy and 5×10^6 electron volts is binding energy (this is an average, as energies can range from 0 to 17 million electron volts). The easiest way to understand the meaning of binding energy is to imagine the reverse process. When a neutron is absorbed by a nucleus, the mass of the latter after absorption is less than the total mass of nucleus and neutron before absorption. The difference in mass, given off in the form of energy in accordance with Einstein's equation already mentioned, is called the binding energy of the neutron in that particular nucleus because it is equal to the amount of energy that must be put into the nucleus again in order to "unbind" or expel the neutron. Since one neutron removes about 7×10^6 electron volts of energy from the fragment, the number of neutrons released or emitted per fission depends on the total amount of energy the fragments must lose, and this in turn depends on how much excitation energy was given to the nucleus in Stage 1. For fission of uranium-235 by thermal neutrons (a thermal neutron is a neutron in thermal equilibrium with matter at room temperature and has an energy of 1/40 electron volt), $\bar{\nu}$ has a value of about 2.5 neutrons per fission; but when fission is caused by neutrons of say 14.8×10^6 electron volts, $\bar{\nu}$ rises to nearly five neutrons per fission because the extra energy of the bombarding neutrons is added to the excitation energy of the fragments. Something else important about prompt neutrons is the way in which the average number emitted depends on the mass of the fragments. This dependence, for the case of thermal neutron fission of uranium-235, is shown by the series of points superimposed on the mass distribution curves of Figure 2.

Delayed neutrons

Besides the prompt neutrons, delayed neutrons are also found in fission, as part of the beta decay process of the fission products in Stage 6. The half-life of delayed neutron emission ranges from about 0.2 second to about 56 seconds. The Table summarizes the various pieces into which the nucleus of Stage 1 may break up in the fission of uranium-235.

Fragment Distribution in Uranium-235 Fission

fragment	approximate average energy (MeV)	approximate number per fission	stage
Fission fragments	166 (total for pair)	2.5	3
Prompt neutrons	2 (kinetic energy only)	2.5	3, 4
Alpha particles (^{4}He)	15	0.002	3
Tritons (^{3}H)	8	0.0001	3
Prompt γ-rays	1.1	7.5	5
β^--particles	1.2	6.5	6
Delayed γ-rays	0.9	many	6
Delayed neutrons	0.5 (kinetic energy only)	0.016	6

The energy of fission. The energy released in fission is carried away by the fission fragments, neutrons and gamma rays. Eventually these products are halted by matter and their energy is converted almost entirely to heat. It is important to utilize as much as possible of this heat if fission is to be employed to the best advantage and, since the ease with which the various products can be stopped varies enormously, it is important to know how the energy is distributed among them.

The so-called energy release in fission is found by subtracting the sum of the masses of the primary fission fragments (Stage 3 in Figure 1) from the original mass (Stage 1) and converting it into energy units by means of Einstein's equation. This is the energy that appears in Stages 3, 4, and 5. The delayed emission of particles in Stage 6 releases additional energy, some of which is bound up in the kinetic energy of antineutrinos, which accompany the beta decays. Antineutrinos are antiparticles of neutrinos, particles of zero charge and negligible mass that are virtually unstoppable by matter and therefore do not share their energy.

The energy arising from the fission of uranium-235 by thermal neutrons is distributed as follows:

Energy released, prompt	
Kinetic energy of fission fragments	166 MeV
Kinetic energy of prompt neutrons	5
Binding energy of prompt neutrons	12
Energy carried by prompt gamma rays	8
Energy released, delayed	
Energy of beta particles	8
Energy of antineutrinos	12
Energy of delayed gamma rays	7
	218 MeV

Most of this energy is converted into heat within an operating nuclear reactor and so can be used to generate electrical power. The energy available from the binding of the prompt neutrons, however, depends on the way in which they are finally stopped; whereas some of the energy of the delayed gamma rays and beta particles is lost because the fuel has to be removed from the reactor before they are emitted. All of the antineutrino energy is lost because these elusive particles are unstoppable and thus cannot be contained.

Fission chain reactions and their control. *The principle.* Although fission can be initiated by any process that imparts sufficient energy to the heavy nucleus, the neutron provides by far the most efficient way of doing it. The reason for this is because a neutron has no electrical charge and so can approach the nucleus without having to overcome the ever-increasing repulsion between itself and the protons in the nucleus, which a positively charged particle like a proton experiences. As a result of this, even thermal neutrons, that is, neutrons that have kinetic energies the same as the energy of the molecules in the surrounding matter, can enter a nucleus: these energies are less than 0.1 electron volt. In contrast, a proton must have a kinetic energy of about 13×10^6 electron volts, a factor of over 1.3×10^8, if it is to be able to penetrate a uranium-235 nucleus.

The neutron then can be an effective means of giving a nucleus the excitation energy it needs to enable it to pass through the sequence of events shown in Figure 1; but the reason why fission can be used as a source of atomic power is because of the emission of prompt fission neutrons in Stages 3 and 4. Each of these neutrons can enter another fissile nucleus and start the sequence again so that the process can continue indefinitely in what is called a chain reaction. Since there are on the average about 2.5 neutrons produced in each thermal neutron fission of uranium-235, it is easy to see that the number of fissions occurring can multiply enormously. For example, if every one of these neutrons were to cause another fission, there would be 2.5 fissions in the second cycle, 6.25 ($=2.5^2$) in the third, 15.63 ($=2.5^3$) in the fourth, and so on, each cycle taking only about 10^{-15} second. This type of increase is called a geometrical progression and is so rapid that if no other factors intervened, the result would be a nuclear explosion. In fact, many of the prompt neutrons will not cause fission at all, but will be absorbed, either by accident or design, in other types of nuclear reaction; but so long as the average number of neutrons

Chain reaction

per fission that do initiate another fission is greater than one, the total number of fissions will increase. This is called a diverging chain reaction. If the number is less than one, a converging chain reaction results and the number of fissions will decrease. The ratio of the number of fissions occurring in one cycle to the number in the next is called the multiplication factor.

Critical mass. If a fission bomb is to be made, the multiplication factor must be as large as possible; whereas if the requirement is for a controlled nuclear reactor, the multiplication factor must be somewhere near one. The important point is that for both purposes one must be able to achieve values of the factor greater than unity. On the face of it, this would seem to be an easy thing to do. Because about 2.5 prompt neutrons are emitted in the fission of one atom of uranium-235, it should be simple to ensure that at least one of them is available to cause the fission of another atom. It is, in fact, easy to attain a multiplication factor greater than one for uranium-235 but much harder for natural uranium, which has only about one part of this isotope in 140, most of the remainder being uranium-238. The latter isotope captures the fast prompt neutrons quite easily, eventually producing plutonium-239 as mentioned earlier; but it is not nearly as easily fissioned as uranium-235, and so it uses up the neutrons needed to raise the multiplication factor above one.

In most of the reactors in use today the fissile material is mixed with a moderator substance, the purpose of which is to slow down the fission neutrons to thermal energies. It does this by a process called elastic scattering in which neutrons collide with moderator nuclei and give up some of their energy to them in much the same way as a billiard ball loses energy if it is in collision with another one. The lighter the moderator nuclei are, the greater will be the average amount of energy lost by the neutrons in each collision. This is why hydrogenous materials such as water are good moderators, although carbon in the form of graphite has many advantages and is also frequently used.

Besides being captured by some nucleus that does not fission, the neutrons may fail to cause a fission for the much simpler reason that they escape from the piece of material without actually striking another nucleus at all. The smaller the piece of material is, the greater the chance that this will happen, and there must therefore be some critical mass or point of criticality below which so many neutrons escape that the multiplication factor falls below one and a chain reaction cannot occur at all. In a piece of uranium-235 metal the critical mass is only a few pounds, but in natural uranium so many neutrons are absorbed by the uranium-238 that it is not possible for a single piece of metal of any size to become critical. The shape of the fissile material must also be considered since the greater the ratio of its surface area to its volume, the greater is the chance of neutrons escaping. Thus the critical mass of a spherical piece of material will be less than if the same material is made into a cylinder. It is for this reason that vessels used for holding solutions of fissile isotopes in processing plants often have a narrow cylindrical shape. Equipment of this type is termed critically safe.

The control of chain reactions. If fission is to be used in reactors to produce power, or for other purposes, it is essential that one be able to control the chain reaction. It is immediately obvious that means are required whereby the value of the multiplication factor can be smoothly increased or decreased for normal operation, or instantly reduced to well below unity if it becomes necessary to shut down the reactor suddenly. To do this, a way must be found to change the number of neutrons that are absorbed and so to alter the fission rate.

The picture is complicated by the fact that the fission cross section for any isotope is not constant, but depends strongly on the kinetic energy of the neutrons: for example, the uranium-235 fission cross section for thermal neutrons is about 580 barns, whereas for the fast neutrons in a typical reactor it is only about 1.5 barns. The control of the reactor is usually accomplished by means of movable mechanical rods made of some material that has a high absorbing power for the neutrons; that is, it has a high neutron capture cross section; boron is commonly used in thermal reactors for this purpose because its isotope boron-10 has a thermal neutron absorption cross section of about 4000 barns, which drops off smoothly as the neutron energy rises. When the reactor is shut down the rods are inserted into the reactor fuel core; to start it up they are gradually withdrawn until the multiplication factor rises to the desired value. The rate at which the number of neutrons produced increases depends on the length of time between successive cycles of neutron generation in the chain reaction as well as on the value of the multiplication factor. This time is increased by the presence of the delayed fission neutrons and this factor helps to ease the problem of control of the reactor. One final factor that affects control is brought about by the gradual accumulation of fission products in the fuel. Some of these products have high neutron-capture cross sections and so produce a progressive reduction in the maximum value of the multiplication factor that can be achieved in the reactor. Eventually, the effect of these fission-product poisons is to reduce power and make the reactor too expensive to run. It is then necessary to change to fresh fuel.

THEORETICAL ASPECTS OF NUCLEAR FISSION

Nuclear models. An atomic nucleus is composed of a grouping of neutrons and protons. Its stability derives from a balance between the two main kinds of force acting on these nucleons; *i.e.,* the repulsive force between the positively charged protons and the attractive nuclear force between all the nucleons. A complete theoretical treatment of a nuclear event such as fission would involve calculating the effect of these forces on all the nucleons at every stage of the fission process, thus providing detailed predictions of all the phenomena resulting from it. Unfortunately, it is impossible to perform these calculations at present and it is likely that this will remain the case for a long time to come, because the nature of the nuclear force is not yet properly understood, and even if it were, it would be impossible to carry out the incredibly complex computations required.

The solution of this dilemma is to resort to the use of models. The idea is to choose some system that can be regarded as being comparable to the nucleus for some aspects of the fission process but that is itself much simpler than the fissioning nucleus. Calculations are then performed on this system in the hope that the results can be applied to the real nucleus itself. Classification of the various models is not easy because there are so many variants that they tend to shade into one another; but the most important are the liquid-drop model, the adiabatic model, and the statistical model, which are described below.

Liquid-drop model. The liquid-drop model of the nucleus was first used by Bohr and Wheeler in 1939, following a suggestion made shortly after the discovery of fission. The model likens the nucleus to a drop of an incompressible liquid having a uniform electric charge. The nucleons inside the drop all attract each other because of the nuclear force, but the attraction is one-sided at the surface of the drop and this results in a so-called surface tension, which is rather like a stretchable skin surrounding the drop that tries to retract to a minimum area. The positive charge of the protons produces a repulsive force acting in the direction opposite to that of the surface tension. In a stable nucleus the surface forces are stronger than this repulsive force and so the drop is squeezed into a spherical shape. If energy is now given to this drop, for example by its absorption of a bombarding neutron, it will oscillate through a whole variety of different shapes about the spherical. If the nucleus starts to stretch out during one of these contortions, the surface tension will at first try to pull it back to the equilibrium shape. The surface tension, however, falls off very rapidly in comparison to the repulsive force, and so a point may be reached beyond which the latter overcomes it and there is then nothing to prevent the nucleus from breaking up. The point at which the two forces are just in balance is

Surface tension

the transition state, Stage 2 in Figure 1. The further a nucleus can be stretched before it reaches the transition state, the more difficult it is to make it fission. Bohr and Wheeler introduced a fissionability parameter, ranging from zero to one, to provide a measure of the stability of a nucleus toward fission, the smaller its value the greater being the stability. Its approximate value is equal to 1/50 of the square of the atomic number divided by the nuclear mass. It is equal to $1/50\ (Z^2/A^2)$, in which Z is the atomic number and A is the mass number of the nucleus. For a nucleus whose fissionability parameter is equal to 1, the transition state shape is a sphere; but a sphere is also the equilibrium shape for a nucleus before any energy is added, and so such a nucleus is always unstable toward fission and hence cannot exist at all. On the other hand, the transition-state shape of a nucleus having a fissionability parameter equal to zero would be two spheres in contact and such a nucleus would be extremely stable against fission. As an example of a typical real nucleus, one may take uranium-235, whose fissionability parameter has a value of 0.72. The liquid-drop model, although simple in concept, has not proved as successful as at first expected. In its original form it predicted that the mass distributions for the heavy fissile nuclides would be symmetrical, whereas experimental results show that the distributions have two pronounced asymmetrical peaks, as Figure 2 will show. Modern variants of the model have overcome this difficulty, but the additional complications that these variants introduce, together with the complexity of calculations on very distorted drop shapes, have led to the development of other models that may be used either in place of it or to supplement it.

Adiabatic models. In a nucleus, the nucleons have different energy states that may be arranged in a series of energy levels, each representing a different energy and capable of holding a limited number of nucleons. Adiabatic is a term used in thermodynamics to describe a system in which a gas is made to expand or contract sufficiently rapidly so that it does not gain or lose heat from its surroundings, but compensates for the change in volume by raising or lowering its temperature. In adiabatic nuclear models the overall motions of the whole nucleus (collective motions) are slow when compared with the motions of the individual nucleons (particle motions), and when the nucleus is excited and starts its contortions, there is thus time for the particle motions to adjust themselves to the shape changes and there is no exchange of energy between the various levels; that is, the nucleus is behaving adiabatically.

The liquid-drop model posits the nucleus as a blob of matter in which all the nucleons interact strongly with each other. In rival groups of models, called independent-particle models, of which the shell model is the best known, an opposite assumption is made, namely, that there is little or no interaction between individual nucleons, but each moves in an orbit in a force field (the nuclear potential) generated by all the nucleons. In the liquid-drop model the nuclear properties result from the behaviour of the whole nucleus, whereas in the independent-particle models they are the result of the behaviour of just a few of the nucleons. Adiabatic fission models combine some features of both types; that is, calculations are made on the basis of the individual particle interactions, modified by the collective motions of the nucleus, and it is assumed that the nucleus behaves adiabatically as well. Considerable success has resulted in predicting those properties of fissioning nuclei that are dependent on events occurring up to the transition state, such as the angular distributions of the fission fragments (Stage 2, Figure 1), but have been less successful with the scission stage (Stage 3), at which such features as the mass and charge distributions of the fission fragments are probably determined.

Shell model

Statistical models. As originally conceived, the statistical model possessed the great advantage that it contained no arbitrary assumptions put in simply to make the model fit the experimental facts. It assumed that there are forces at work in the nucleus that are not adiabatic in nature and that the effect of these is to establish thermodynamic equilibrium between the different possible conditions of the nucleus at the scission point. The meaning of this can be explained in the following way: a large number of nuclei, all at the scission point, can have any one of a number of different configurations, each of which leads to a different pair of fission fragments. If equilibrium exists, the number of nuclei in each of the various configurations will depend only on the properties of the configurations and will not be influenced in any way by events that have gone before, simply because there has been enough time for the equilibrium to be established.

Non-adiabatic model

In detail, it was assumed that the probability of a nucleus fissioning so as to produce any one of the possible pairs of fission fragments depends only on the closeness of the energy levels of those fragments. Since the fragments must have a definite energy level, the more possible levels each fragment pair can have the better is the chance of being able to produce these particular fragments. It was also assumed that there is no nuclear interaction between the fragments after scission. The mass distribution of the fission fragments of uranium-235 thermal neutron fission was accurately predicted, but it was soon clear that predictions of the mass distribution for plutonium-239 thermal neutron fission fragments and for the kinetic energy distributions of fission fragments were completely wrong.

Limitations of statistical model

The difficulty with the statistical model as it was originally conceived is that while it uses only properly defined physical properties as the basis of its calculations, these properties relate mainly to the fission fragments at the moment of scission. Such fragments exist only for a very brief moment of time and their properties are virtually unmeasurable. This means that the required data must be obtained by extrapolation from measurements on the behaviour of ordinary nuclei, and such long extrapolations may be very inaccurate. Since the original statistical model was proposed innumerable modifications have been made, and these have tended to erode somewhat its independence from arbitrary assumptions. In compensation for this, however, modern statistical models predict accurately many mass distributions and other fission properties and it seems certain that improvements will continue to be made.

The present position. Some important new ideas concerning the stability of nuclei have emerged in the last few years as a direct result of the discovery in 1962 by the Soviet physicist S.M. Polikanov and others of fission isomers. This term can best be explained by means of an example. The normal isotope americium-241 has a spontaneous fission half-life of about 2×10^{14} years, but there is another isotope, also americium-241, which has a spontaneous fission half-life of just over one-thousandth of a second. This is the fission isomer americium-241. Although it seems very short, the half-life of americium-241 is far too long to be explained in terms of existing models. V.M. Strutinskii, a physicist of the Soviet Union, and others have developed a model in which the average properties of nuclei are described by liquid-drop ideas but then have small variations superimposed on them by the action of individual nucleons.

One effect of these developments has been to change the shape of the fission barrier, mentioned earlier. Instead of a single smooth hump, at the top of which the nucleus is in the transition state, there may be two humps. A fission isomer is a nucleus that is stretched out so that its potential energy has fallen to the value represented by the bottom of the hollow between the two humps. It therefore has to surmount or penetrate the second hump in order to fission and so it has rather more stability than would be expected.

The Strutinskii model also changes the earlier predictions concerning the stability of nuclei as the atomic number and mass number increase. It now seems probable that super transuranic nuclei having high atomic numbers around 114 and 126 will be comparatively stable against spontaneous fission and that some of them may therefore have half-lives of the order of many years. Theoretical predictions of the nuclear properties of these super-heavy

Long-lived super transuranics

nuclides suggest that they will have a high fission energy release and may emit as many as ten prompt neutrons per fission. Scientists throughout the world are trying to prepare samples of these nuclides by particle bombardment, or to find them in a natural state, and it seems clear that there are many interesting developments still to come in the field of nuclear fission.

BIBLIOGRAPHY. The classical papers of O. HAHN and F. STRASSMANN, in which the original fission discovery was reported, may be found in *Naturwissenschaften*, vol. 27 (1939). N. BOHR and J.A. WHEELER, "The Mechanism of Nuclear Fission," *Phys. Rev.*, 56:426–450 (1939), is the classical paper that proposed the liquid-drop model, still of basic theoretical importance in fission. A popular type article well worth reading is R.B. LEACHMAN, "Nuclear Fission," *Scient. Am.*, 213:49–59 (1965). More advanced references include: L. WILETS, *Theories of Nuclear Fission* (1964), an excellent work but not easy reading; I. HALPERN, "Nuclear Fission," *A. Rev. Nucl. Sci.*, 9:245–342 (1959); J.R. HUIZENGA and R. VANDENBOSCH, "Nuclear Fission," in *Nuclear Reactions*, ed. by P.M. ENDT and P.B. SMITH, vol. 2, pp. 42–112 (1962); E.K. HYDE, *Nuclear Properties of the Heavy Elements*, vol. 3, *Fission Phenomena* (1964); J.S. FRASER and J.C.D. MILTON, "Nuclear Fission," *A. Rev. Nucl. Sci.*, 16:379–444 (1966); and J.E. GINDLER and J.R. HUIZENGA, "Nuclear Fission," in *Nuclear Chemistry*, ed. by L. YAFFE, vol. 2, pp. 1–183 (1968). For a discussion of superheavy elements see G.T. SEABORG and J.L. BLOOM, *Scient. Am.*, p. 57 (April 1969); G.T. SEABORG, *Isotopes and Radiation Technology*, 73:251 (1970).

(J.G.C.)

Nuclear Fusion

Nuclear fusion refers to the phenomenon in which two or more relatively light atomic nuclei combine to form a heavier atomic nucleus. If the interacting nuclei belong to elements with low atomic numbers, the reactions are exothermic; that is, they release energy. This is due to the fact that, as the number of subatomic particles, or nucleons (neutrons and protons), packed into the nuclei of the lighter elements increases, the tightness of their packing, expressed as binding energy, also increases. The term denoting the binding energy, however, has a minus sign when the total energy of the atomic nucleus is computed from the rest energies of the free nucleons. Since an unstable system will release energy if thereby it becomes more stable, fusion reactions in substances composed of light nuclides may, under certain conditions, occur spontaneously.

DISCOVERY OF FUSION

The scientific and technical problems relating to nuclear fusion have been the subject of research since the 1930s, when exothermic fusion reactions were first discovered. In 1939, the U.S. physicist Hans A. Bethe proposed that most of the energy of stars, including that of the sun, derives from fusion reactions proceeding at very high temperatures and densities. This hypothesis is almost universally accepted today. Soon after World War II, the attempt began to create a weapon of tremendous power by utilizing fusion reactions in a mixture of deuterium and tritium (the heavy isotopes of hydrogen), and in 1952–53 the so-called hydrogen bomb became a reality.

Hydrogen bomb

Early in the 1950s separate research efforts in the Soviet Union, in the United States, and in Great Britain began to seek ways of harnessing fusion reactions for peaceful needs. Each of these nations pursued its objectives in complete secrecy until 1956, when some research results were first made public by the Soviet Union. The second international Atoms-For-Peace Conference in Geneva in 1958 was characterized by broad exchanges of information on the principal trends of research in the control of nuclear fusion. The barriers of secrecy having been removed, further work on the problem produced fruitful cooperation among physicists of many countries. Nevertheless, all attempts to find a shortcut to the construction of a fusion reactor proved unsuccessful, and progress in this direction became slower than had been anticipated.

THE FUSION PROCESS

Two kinds of reactions. It is necessary to distinguish between two important types of fusion reactions. One of these types consists merely of a regrouping of nucleons into new nuclei, without involving conversions of protons to neutrons or vice versa. An example of such a reaction is the fusion of two deuterons:

$$d + d \rightarrow \begin{cases} {}^3\mathrm{He} + n \\ t + p \end{cases}, \qquad (1A)$$

in which d represents a deuteron (proton plus neutron), t a triton (proton plus two neutrons), p a proton, n a neutron, and ^{3}He, a nucleus of helium consisting of two protons plus one neutron. The two arrows indicate that this reaction may proceed in either of two directions: the formation of a nucleus of ^{3}He, a light isotope of helium, and neutrons, or the formation of tritons and protons. Both reaction routes are about equally probable. The second important type of fusion reaction does involve conversions between neutrons and protons. The simplest example of such a reaction is the synthesis of deuterium from hydrogen:

$$p + p \rightarrow d + e^+ + \nu. \qquad (2)$$

In this reaction one of the two protons is converted into a neutron with the simultaneous creation of a positron e^+ (a positive electron), and a neutrino ν (a particle without charge, having negligible mass and great penetrating power); that is, the process of fusion involves a radioactive decay with the emission of two light particles classified as leptons. Each type of fusion reaction has two characteristics: the energy equivalent (defined as the amount of energy released in a single elementary event), and the intensity of the reaction (being a measure of the number of elementary events occurring in a unit of volume of a mixture of reactants per unit time).

Energy yield. If the kinetic energy of the reacting nuclei is relatively small (as is true for all cases of real interest), the energy equivalent may be assumed to equal the sum of the kinetic energies of the products of the reaction. In very accurate calculations, however, the sum of the kinetic energies of the nuclei before the reaction occurs must be taken into account. In atomic and nuclear physics the energy released in elementary events is usually measured in electron volts. (One electron volt is the energy acquired by an electron, or by some other singly charged particle, on being accelerated by a potential difference of 1 volt; 1 electron volt $= 1.6 \times 10^{-12}$ erg $= 1.6 \times 10^{-19}$ joule.) In reaction 1A the release of energy for the route indicated by the upper arrow is 3,300,000 electron volts, while for the route indicated by the lower arrow it is 4,000,000 electron volts. To appreciate the significance of these quantities, it should be realized that if all the atoms in a single gram of deuterium were to react according to 1A, the energy released would amount to approximately 8×10^{10} joules. If, moreover, one takes into account the fact that the fusion of deuterium produces tritons and ^{3}He nuclei, which may enter with deuterons into further highly exothermal reactions, and if consideration is given to the energy released when the freed neutrons are captured in the materials of the container, the total amount of energy liberated is found to be five times larger than this value, reaching approximately 4×10^{11} joules per gram.

Fusion by nuclear bombardment. Fusion reactions can be observed in the laboratory by a standard technique in which a substance is bombarded with a stream of fast particles issuing from an accelerator. The reaction products are registered by the conventional detectors of nuclear radiation, such as Geiger counters, cloud chambers, and photographic emulsions. The reaction yield for any given reaction depends on sigma (σ), a quantity called "the effective cross section," which is defined as the probability that a reaction will occur when a single fast particle passes through a layer of the substance containing a single nuclear target per unit area. Thus, if n_0 fast particles pass through a thin layer of a substance in which there are N nuclear targets per square centimetre, the total number of fusion reactions would be $n_0 N\sigma$. The effective cross section for a given reaction depends on the kinetic energy, W, of the bombarding particle, and it is possible to establish experimentally the precise relationship between these two quantities by counting the number of elementary reaction events for different values of W.

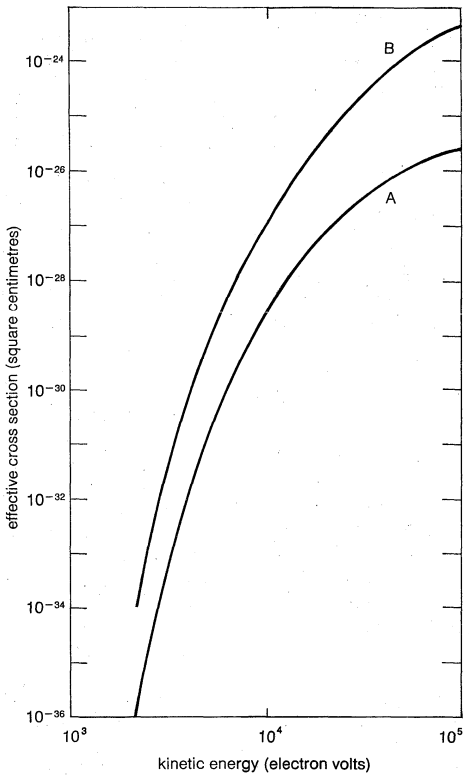


Figure 1: Experimental relationship between cross section of fusion reaction and the kinetic energy of mutual approach of nuclear particles. Curve A refers to reaction 1A for deuterium target; Curve B refers to reaction 1B for tritium target.

Effective cross sections. The results of such experiments are summarized in Figure 1. Curve A relates to reaction 1A, and here σ stands for the *overall* effective cross section, taking into account both possible routes of this reaction. Curve B relates to another reaction:

$$d + t \rightarrow {}^{4}\mathrm{He} + n \qquad (1B)$$

which is of practical interest. The energy equivalent (17,600,000 electron volts) for this reaction is quite high, and its effective cross section is, relatively speaking, very large. It is apparent from Figure 1 that, as the energy of the bombarding particle increases, the value of the effective cross section rises initially quite steeply; but at higher values of W the corresponding rise in σ is drastically retarded, if not stopped altogether. This behaviour may be accounted for in the following manner. A fusion reaction can occur only if two nuclei have succeeded in approaching one another within a distance of the order of

Nuclear forces

10^{-13} centimetre, because only when the distance between the nuclei is this small is it possible for the nuclear forces of attraction to overcome the electrostatic forces of repulsion due to the presence of positive charges on both nuclei. The probability that a bombarding nucleus will penetrate the potential barrier of electrostatic repulsion as it approaches a target nucleus rises very quickly with the increase in the energy of the relative motion of approach. This is expressed in the initial steepness of the rise of the value of σ, and holds strictly only as long as the kinetic energy of approach for the two nuclei remains well below the height of the potential barrier. But as the energy of the relative motion of approach becomes comparable to or exceeds the magnitude of the potential barrier, the increasing insignificance of the potential barrier as an obstacle to nuclear interaction is compensated for by other factors, such as, for example, the greatly reduced time span available for interaction while the two nuclei are separated by still no more than the extremely short distance in which the nuclear force is effective. Because the effect of electrostatic repulsion increases with the magnitude of the charges on the nuclei, the only fusion reactions occurring with appreciable probabilities when two nuclei approach each other with not very high kinetic energies are reactions between the lightest among nuclei; specifically, the nuclei of the various isotopes of hydrogen and helium.

The statements made concerning the dependence of σ on the kinetic energy of approach are, to some extent, also valid for reactions of type 2 accompanied by the emission of leptons. In this case, however, in order to react, not only must the nuclei involved have succeeded in approaching each other to within a distance of the order of 10^{-13} centimetre, but, in addition, one of the fundamental nuclear particles must undergo a radioactive conversion in the negligible time interval during which the nuclei are separated by no more than that distance—a very improbable event. For comparable values of W, therefore, the effective cross sections of reactions involving the conversion of fundamental nuclear particles are expected to be smaller by many orders of magnitude than the effective cross sections of reactions involving merely a regrouping of protons and neutrons. According to theoretical calculations for reaction 2, for example, the value of σ corresponding to values of W not exceeding 1,000,000 electron volts is less than 10^{-46} cm^2. This explains why a fusion reaction even as simple as reaction 2, representing the synthesis of deuterium from hydrogen, cannot in practice be observed in the laboratory.

Low intensities. Although it is quite possible to obtain useful information about elementary fusion events by bombarding a substance with fast nuclei, such bombardments are ineffective as a method of producing fusion reactions with appreciable intensities. The charged, fast particles, when passing through the substance, are greatly retarded by numerous collisions with atomic electrons. In such collisions an initially fast particle rapidly loses relatively large amounts of energy, and its free path in the substance becomes, therefore, very short. While covering its short, free path, the charged particle has only a very small chance of passing an atomic nucleus at a distance close enough for a fusion reaction to be possible. The probability for the occurrence of reaction 1A, for example, when a deuteron with a kinetic energy of the order of 10,000 electron volts is directed at a target of deuterium, is as low as approximately 10^{-11}. Even if W is increased to 1,000,000 electron volts, the probability for the occurrence of this reaction still remains negligible.

Requirements for intensive fusion reactions. From the point of view of usefulness as a source of power, the attainable intensity of fusion reactions and its control are of critical concern. Nuclear fusion processes may develop in a substance spontaneously, that is, without bombardment by fast particles from an external source, if the temperature of the substance is sufficiently high. Such spontaneous reactions occurring because of the high temperature are known as thermonuclear reactions. The higher the temperature of the substance the greater is the kinetic energy of its atomic nuclei in their chaotic motion. The fast-moving nuclei tend to collide with one another, and if in some such collision their kinetic energy of approach is sufficiently high by comparison with the potential barrier of electrostatic repulsion, a fusion reaction may be initiated. The total number of elementary fusion events occurring in a unit of volume of the substance per unit time is proportional to the square of the concentration of nuclei and rises very steeply with the temperature. Consider, for example, a rigid and closed container filled with deuterium and maintained at a pressure of one atmosphere while the temperature is gradually allowed to rise. Until the temperature has reached several hundred thousand degrees Kelvin, a detector will register no sign that a fusion reaction has yet occurred. But at a temperature of 1,000,000° the detector will register as many as about 10^8 elementary fusion events per second for each cubic centimetre of volume enclosed by the vessel, while at a temperature of 100,000,000° the yield will have increased to some 10^{21} fusion events per cubic centimetre per second, implying that in only 0.1 second all deuterons will have participated in reactions with release of nuclear energy.

Spontaneous fusion

Temperature dependence. At any given temperature, T, the atomic nuclei of any substance have a wide dis-

tribution of energies, which, for stationary conditions, is described by Maxwell's law. As long as the temperature is relatively low and the mean kinetic energy of the nuclei is no more than a fraction of the height of the potential barrier of electrostatic repulsion, fusion reactions can be initiated only by the relatively few nuclei whose energies correspond to the remote "tail-end" of the distribution and which, furthermore, happen to encounter one another while approaching from nearly opposite directions, because it is under these conditions that their kinetic energy of approach is greatest. In the region of moderately high temperatures, the yield, or the number of reaction events, increases extremely quickly as the temperature continues to rise. Eventually, with the approach to very high temperatures, nuclei belonging to an increasingly wider part of the energy distribution participate in the fusion process, but the growth of the yield declines. The mean value of the kinetic energies of all nuclei in the substance equals $\frac{3}{2}\,kT$ with k, the Boltzmann constant, equalling 1.38×10^{-16} erg/degree. When the mean kinetic energy reaches the same order of magnitude as the height of the potential barrier of electrostatic repulsion, the growth of the reaction yield practically ceases. At the very high temperatures needed to initiate thermonuclear reactions, nearly all the atoms of the substance are practically fully ionized; that is, they are stripped of all their electrons, resulting in a homogeneous mixture of charged particles, such as free electrons and bare atomic nuclei—all in rapid and randomly directed motion. In this state the substance is called a plasma, and high-temperature plasmas are the only natural medium for the generation of thermonuclear reactions. The fast-moving atomic nuclei of a high-temperature plasma are not retarded in their collisions with free electrons, because the mean kinetic energy is here the same for both types of particles.

Plasma

Predicting yield. It is possible to predict theoretically the yield of thermonuclear reactions for the fusion processes of most general interest. In conformity with the above arguments, the number g of elementary fusion events per second in one cubic centimetre of plasma would be expected to be $g = n_1 n_2 f(T)$, where n_1 and n_2 are the numbers of atoms per cubic centimetre of plasma of each of the two reacting nuclides respectively, while the factor $f(T)$ is some function of the temperature. In the case of a fusion reaction between nuclei of the same nuclide (as, for example, deuterium in reaction 1A), the product $n_1 n_2$ is to be replaced by $\frac{1}{2}n^2$. Figure 2 shows

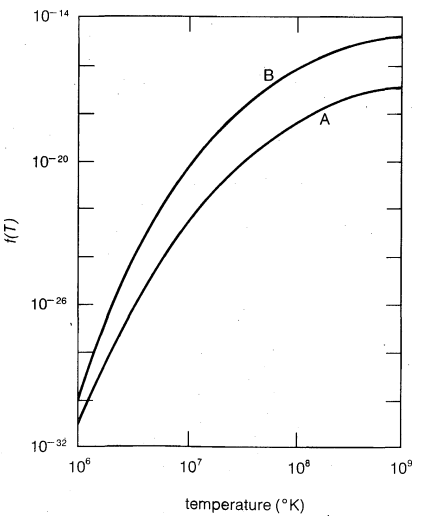


Figure 2: Theoretical relationship between intensity of thermonuclear reaction and temperature. Curve A refers to reaction 1A taking place in pure deuterium; curve B refers to reaction 1B in a mixture of deuterium and tritium.

the behaviour of $f(T)$ for two cases that are of greatest practical interest. Curve A applies to reactions taking place in pure deuterium, while curve B applies to a mixture of deuterium and tritium in which only reactions between different nuclides, as in 1B, are being considered. The effective cross section for reaction 1B is, within a broad range of collision energies, a hundred times greater than that for reaction 1A. It is this fact that accounts for the substantial difference in the behaviour of $f(T)$ for these two reactions.

SOURCES OF THERMONUCLEAR REACTIONS

Thermonuclear reactions and the energy output of stars. According to notions first advanced by Bethe, the energy emitted by the so-called normal stars, of which the sun is an example, derives mainly from sequences of thermonuclear reactions in which hydrogen is converted into helium, with four nuclei of hydrogen (four protons), contributing to the synthesis of a single nucleus of ^{4}He. The end result may be arrived at by several routes, the most important of which are two sequences of nuclear reactions known respectively as the hydrogen cycle and the carbon cycle. These two cycles together contribute most to the energy generated by a star at a stage of its evolution when it consisted mostly of hydrogen.

$$2(p + p \rightarrow d + e^+ + \nu)$$
$$2(d + p \rightarrow {}^3\mathrm{He} + \gamma)$$

(in which ν denotes a neutrino and γ denotes a gamma photon)

$$^3\mathrm{He} + {}^3\mathrm{He} \rightarrow {}^4\mathrm{He} + 2p$$

The hydrogen cycle. The hydrogen cycle begins with reaction 2, which produces deuterium. Two such reactions are required. On reacting with more hydrogen, each of the two deuteriums turns into ^{3}He accompanied by the emission of gamma rays. And, finally, the ^{3}He reacts with more ^{3}He to produce ^{4}He and more hydrogen. The intermediate products, ^{3}He and d, have disappeared, and the net effect of the complete cycle of three reactions is the fusion of four protons to form one nucleus of the heavier of the stable isotopes of helium. The last stage of this cycle requires two nuclei of ^{3}He, each of which had been produced in an independent reaction requiring a deuteron; and each of the two deuterons, in turn, had been produced in an independent reaction from two protons with the simultaneous creation of positrons. When calculating the energy balance of this cycle, therefore, one must also take into account the energy released in the annihilation of two positrons by the two extra electrons remaining after the conversion of hydrogen to helium. The total energy generated by the hydrogen cycle with the creation of a single nucleus of ^{4}He is 26,700,000 electron volts, and nearly all of this energy is initially converted into heat in the immediate vicinity of the points where the reactions occur. These reactions take place well within the deep interior of the star, from whence the thermal energy is slowly transported through the immense volume of the star to its surface by the relatively inefficient processes of thermal conduction. On reaching the surface, the thermal energy is converted to electromagnetic energy, which is radiated into space. Of the total amount of energy generated in a star by thermonuclear processes only about 2 percent is not involved in its heat balance. This energy is dissipated into space by neutrinos—particles that, because of their enormous penetration capacity, are not retained within the star for significant amounts of time. According to theoretical calculations, the hydrogen cycle contributes the greater share of the energy generated in the interior of stars when the temperatures are below 15,000,000 degrees.

Neutrinos

The carbon cycle. At higher temperatures the more rapid conversion of hydrogen to helium is accomplished mostly through the carbon cycle:

$^{12}\mathrm{C} + p \rightarrow {}^{13}\mathrm{N} + \gamma,$
$^{13}\mathrm{N} \rightarrow {}^{13}\mathrm{C} + e^+ + \nu$ (radioactive decay of the unstable ^{13}N nucleus),
$^{13}\mathrm{C} + p \rightarrow {}^{14}\mathrm{N} + \gamma,$
$^{14}\mathrm{N} + p \rightarrow {}^{15}\mathrm{O} + \gamma,$
$^{15}\mathrm{O} \rightarrow {}^{15}\mathrm{N} + e^+ + \nu$ (radioactive decay of the unstable ^{15}O nucleus),
$^{15}\mathrm{N} + p \rightarrow {}^4\mathrm{He} + {}^{12}\mathrm{C}.$

The nucleus of ^{12}C acts merely as a catalyst, and the net

effect is the fusion of four protons to form a single nucleus of ^{4}He, just as in the case of the hydrogen cycle. As expected, therefore, the total energy generated by the carbon cycle is found to be the same as for the hydrogen cycle. But here the part of the energy (6 percent) carried away by neutrinos is somewhat larger. Besides these two main processes of converting hydrogen to helium, there are others that will not be dealt with here. In old stars of high density, in which much of the hydrogen has already been converted to helium, fusion reactions producing heavier elements, such as carbon from helium, become increasingly more important. These processes occur at considerably higher temperatures than the carbon cycle, and the generation of energy per unit mass is lower.

Stellar energy. The rates at which energy is released by thermonuclear fusion processes in stars are characteristically low. In the sun, for example, the rate of heat generation per gram is a mere 2×10^{-7} watt. This is less by many orders of magnitude than the power generated in burning one gram of an organic substance, such as coal, in the furnace of an old-fashioned thermal power plant. Yet, despite this negligible rate at which nuclear energy is being released, the temperatures prevailing in the interior of stars are extremely high, ranging from 10,000,000° to 20,000,000°. This is due to the fact that, because of the huge dimensions of stars, the hot zones in their deep interior, where most of the fusion reactions occur, are very effectively thermally insulated.

It ought to be borne in mind that modern notions concerning nuclear fusion reactions in the interior of stars are speculative. Nevertheless, the thermonuclear fusion hypothesis is the only hypothesis capable of accounting for the origin of stellar energy within the framework of present knowledge of atomic and nuclear processes. In principle, at least, it is possible to put this hypothesis to an experimental test that would consist of registering the flux of neutrinos issuing from the nearest star, the sun. Neutrino fluxes from the sun can be intercepted by means of specialized, extremely bulky, and expensive apparatus in laboratories situated deep underground where they cannot be reached by any of the various types of other particles arriving at the surface of the earth from outer space. Such experiments have just begun, but their outcome will be critical. If it is indeed possible to detect neutrinos born in the interior of the sun, the experimental results will have corroborated one of the boldest physical hypotheses ever advanced to account for cosmic phenomena. Conversely, the proved absence of neutrino fluxes of the anticipated intensity would be an even more intriguing result, because it would force the postulation of entirely new types of processes involving the release of energy, and this could entail the revision of some of the most fundamental ideas of physics.

Hypothesis

The hydrogen bomb. The hydrogen bomb, or H-bomb as it has come to be known, is an instrument of destruction in which thermonuclear reactions of tremendous power are artificially stimulated. The explosion of such bombs is known to liberate energies of up to 10^{17} joules. Its design is, in principle at least, quite simple. The initial explosion of an atomic detonating device embedded in a substance containing deuterium, tritium, and ^{6}Li (the lighter isotope of lithium) raises the temperature to about 10,000,000° within microseconds. This practically instantaneous heating of the mixture of the three light nuclides gives rise to exothermic nuclear fusion reactions of considerable intensity that cause a further increase in the temperature. As a consequence, the reaction rate is sharply accelerated, so that, in an extremely short time interval, all of the many nuclei making up the bulk of the bomb have managed to react and thus to contribute to the release of the surplus nuclear energy contained in the bomb.

Atomic detonator

The time required for a considerable fraction of the total number of nuclei to participate in the fusion process is, at a given temperature, inversely proportional to the concentration of nuclei, or the density of the substance. It is therefore essential that the density be sufficiently high, implying that the explosive mixture must be in a condensed state. Otherwise, the nuclear explosive, after being ignited by the atomic detonator, would be scattered long before each atom could have an opportunity to participate in a nuclear reaction. The main nuclear processes occurring in the H-bomb are reactions between two deuterons (1A), reactions between deuterium and tritium (1B), and reactions involving the disintegration of lithium according to $^6\text{Li} + n \rightarrow {}^4\text{He} + t$, which compensate for the rapid depletion of tritium through reaction 1B. The disintegration of lithium is likewise exothermic. Its energy equivalent is 4,800,000 electron volts. The neutrons for this reaction derive initially from the fission of the heavy uranium nuclei in the explosion of the detonator, but subsequently they are supplied by reactions 1A and 1B.

The generation of energy by the explosion of a thermonuclear bomb is considerably enhanced if the bomb is encased in a shell of ordinary uranium. When this is so, the fast neutrons produced by reactions 1A and 1B will give rise to nuclear fission processes in the uranium envelope, which are accompanied by the release of energy at a rate of about 200,000,000 electron volts for each elementary fission event. Thus, thermonuclear bombs of this type operate through a complex sequence of reactions initiated by nuclear fission, continued through nuclear fusion, and terminated by fission once more. The additional nuclear fuel used in this case is an inexpensive isotope of uranium that is unsuitable for use with an ordinary atomic bomb.

Thermonuclear reactor. The explosion of a thermonuclear bomb, representing a sudden uncontrolled release of energy, cannot in practice be utilized directly to provide power for industry. Hence, the prospects for harnessing the vast amounts of energy, which are so readily available through nuclear reactions involving the light elements, assume critical importance. Because ordinary water contains about one deuteron for every 6,000 protons, and the separation of deuterium from water is a sufficiently simple matter, the oceans may be considered a practically inexhaustible source of nuclear energy. The problem then reduces to one of finding ways of exercising sufficient control over the rates or progress of thermonuclear reactions in high-temperature plasmas, so that the release of nuclear energy would be gradual and could be maintained at a level at which it may be conveniently transformed into electrical energy with maximum efficiency. Any system capable of accomplishing this task is called a thermonuclear reactor.

Inexhaustible source of fuel

BASIC REQUIREMENTS IN THERMONUCLEAR REACTORS

Two conditions. If a high-temperature plasma is to serve as fuel in a thermonuclear reactor, two principal conditions have to be met. The first condition is that the amount of energy released in the fusion reactions must exceed the plasma's losses of energy by radiation, which result from the emission of X-rays that accompanies the collisions of fast electrons with atomic nuclei. Because the plasmas in which nuclear fusion rates could be controlled most readily have relatively low densities, X-rays could easily escape. The energy carried off by X-rays is roughly proportional to T, in which T is the plasma temperature; but the energy released by fusion reactions increases with T according to its $\frac{5}{2}$ power and begins to exceed the radiation losses only above a certain definite value of T. A thermonuclear reactor can produce surplus energy only at a temperature higher than this threshold value, which for pure deuterium is about 3×10^8 degrees, while for a mixture of equal parts of deuterium and tritium, it is about 4×10^7 degrees.

Operating temperatures

The second condition that has to be satisfied in a thermonuclear reactor follows from the fact that energy is removed from the plasma not only by electromagnetic radiation but also through thermal conduction, as well as by streams of fast particles. The combined effect of all these factors on the energy balance of the plasma is taken into account by introducing the quantity τ, defined as the "mean lifetime" of a particle in the plasma. Because the thermal energy of the plasma is made up of the energy of motion of individual particles, τ is also a measure of the interval of time for which the plasma's thermal energy is preserved. A thermonuclear reactor will produce surplus

Mean lifetime

energy only if there is a sufficiently high probability that an atomic nucleus in the plasma during the time interval τ will experience a collision leading to a fusion reaction. The probability for such a collision is proportional to both τ and n, the concentration of particles in the plasma. Thus, the product $n\tau$ must be required to exceed a certain limit. The limiting value of $n\tau$ for pure deuterium is about 10^{16} particle-sec/cm^3, while for a mixture of equal amounts of deuterium and tritium it is only about 10^{14} particle-sec/cm^3—that is, two orders of magnitude smaller. Thus, a mixture of deuterium and tritium would be a much more effective nuclear fuel than pure deuterium. Tritium, however, is not available from natural sources, though small quantities of it could be prepared in ordinary atomic reactors. For this reason, the use of tritium as a component of a nuclear fuel would be feasible only if the thermonuclear reactor happened to be a breeder of that isotope. In principle at least, the breeding of tritium can be realized through a neutron-multiplication reaction (in which a highly energetic neutron on interacting with a nucleus causes the emission of two neutrons of lower energy), as well as through a reaction involving the capture of neutrons by ^{6}Li. But in practice it is deuterium that appears destined to play the major role as a fuel in thermonuclear power plants of the future.

Deuterium fuel

Problem of control. For the technical solution of the problem of effective control over nuclear fusion processes it is first necessary to find ways of heating a plasma to a temperature of the order of 10^8 degrees and, having once reached such a high temperature, to maintain this temperature for a sufficiently long time interval. These requirements represent the major obstacles to the design and construction of a practical thermonuclear reactor. A high-temperature plasma is an extremely efficient conductor of heat. Its thermal conductivity at a temperature of about 10^8 degrees is a million times greater than the thermal conductivity of ordinary metals at room temperature. In a thermonuclear reactor, therefore, the plasma must be effectively insulated from the walls of its container. This insulation is of critical importance because, in the absence of such insulation, all the energy imparted to the plasma as it is being heated will immediately be transmitted to the walls by the fast-moving particles of the plasma. It is possible to insulate a high-temperature plasma by means of a strong magnetic field whose lines of force permeate the plasma and surround it on all sides. In a magnetic field, the electrons and ions of a plasma are relatively free to move only in directions parallel to the lines of force. A charged particle moving perpendicular to the direction of a magnetic field is compelled to move in a circular path, the radius of which becomes smaller as the strength of the magnetic field increases. (For the sake of simplicity, only those portions of the plasma are being considered here in which the magnetic field is very nearly homogeneous.) In general, therefore, the trajectory of a charged particle would be a helix appearing to be wound over a magnetic line of force (Figure 3). It follows that, wherever

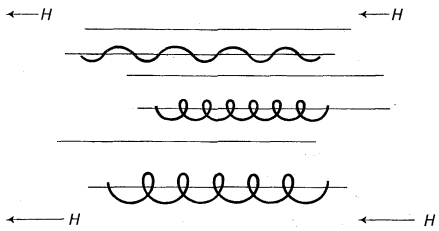


Figure 3: Trajectories of charged particles in homogeneous magnetic fields. Parallel straight lines are lines of force of a magnetic field with strength, H, constant in direction, but increasing toward left. Screwlike curves are paths of given charged particle in various regions of magnetic field (see text).

the lines of force run parallel to the boundaries of the plasma, the magnetic field acts as a barrier preventing the escape of the charged particles. This phenomenon is known as "magnetic thermal insulation." It can also be explained in macroscopic terms. Plasmas are strongly diamagnetic, implying that the strength of the magnetic field over the region occupied by the plasma is greatly reduced. Because of the displacement of the magnetic field by the plasma, the magnetic pressure outside the region occupied by the plasma exceeds the magnetic pressure inside the plasma, and it is this difference in magnetic pressure that prevents the plasma from spreading. During the early stages in the study of controlled fusion, it was anticipated that the principle of magnetic thermal insulation would rapidly lead to the realization of the objective—the construction of a thermonuclear generator. Soon, however, it became clear that, because of certain complicating phenomena in the behaviour of plasmas in magnetic fields, the practical application of magnetic thermal insulation would not be easy to accomplish. It was then realized that the solution of the technical aspects of the problem of controlling nuclear fusion would have to wait until the physical behaviour of high-temperature plasmas is more fully understood.

Magnetic thermal insulation

HIGH-TEMPERATURE PLASMAS

The efforts directed toward achieving control over nuclear fusion, as well as basic studies of properties of high-temperature plasmas per se, require the generation of plasmas with sufficiently high concentrations, n, of particles at extremely high temperatures, and the preservation of the thermal energy of such plasmas for considerable intervals of time τ. The separate values of τ and n are not as critical as the product $n\tau$.

Trends in research. *Mirror trap.* Several avenues of research are directed toward this common objective. One approach is based on the use of systems known as magnetic "mirror traps." The containment of the plasma in such devices depends on the motion of charged particles in nonuniform magnetic fields. A charged particle moving along a line of force in the direction of increasing field strength is compelled to slow down. But while its longitudinal velocity decreases, its speed of rotation about the line force increases, so that the total speed in its helical path remains constant. Therefore, the only change is in the angle between the direction of the velocity and the direction of the magnetic lines of force. If the velocity of a particle is initially directed at a sufficiently large angle to a line of force, the particle will, on approaching a stronger field region, rebound from it as though reflected from a mirror. It should thus be possible to confine all plasma particles between two "magnetic mirrors" in a system in which the magnetic field strength increases at either side of a relatively weak central region. The simplest example of a system designed to incorporate this principle is the device shown in Figure 4. Here, the

"Magnetic mirrors"

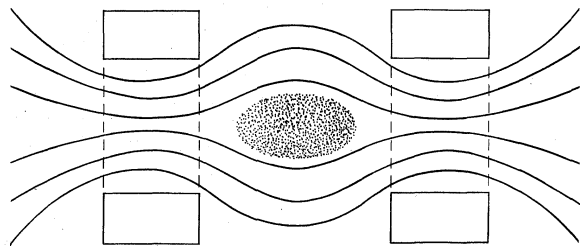


Figure 4: *Plasma trap of the open type employing a pair of magnetic mirrors.*
Four rectangles represent a pair of coils acting as magnetic mirrors, curved lines of force show configuration of magnetic field, and dotted area is region occupied by plasma.

two magnetic mirrors confining the plasma to the space between them are created by a pair of coils that are arranged so that the current through them flows in the same direction. Such a system also offers a considerable choice in the method of generating the high-temperature plasma that it is designed to contain. The plasma within the magnetic trap may be created by such methods as injecting streams of fast particles, capturing streams of plasma, or high-frequency heating.

Closed magnetic trap. The second major trend of research is concerned with ways of generating and main-

taining a circular ring of plasma within the magnetic field of a toroid, an enclosure shaped like the inflated inner tube of a tire. In this case the bulk of the plasma would be free to flow along the strictly concentric magnetic lines of force which are wholly confined to the space within the toroid, as shown in Figure 5. Such devices are referred to as closed magnetic traps, as distinguished from the open magnetic (mirror) traps described above.

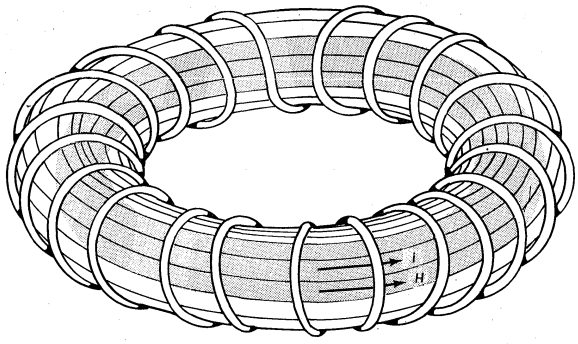


Figure 5: Closed-type plasma trap. Dotted area represents ring-shaped mass of plasma. Current *I*, circulating in plasma ring itself, produces a strong magnetic field of toroidal configuration (not shown). Current flowing in windings of toroid enveloping plasma creates a relatively weak ring-shaped magnetic field *H*, which helps suppress plasma's instabilities.

In the two types of magnetic trap mentioned so far, the bulk of the plasma must be in equilibrium with the magnetic forces acting on it, and all processes associated with the plasma must, therefore, be quasi-stationary. From the macroscopic point of view this means that in order for these traps, whether open or closed, to function properly, no dynamic processes such as would give rise to inertia forces may be associated with the plasma. In microscopic terms, the quasi-stationary behaviour of the plasma implies that the fast, charged particles, while they are still confined by the magnetic system, will perform thousands or tens of thousand of oscillations between the boundaries of the region occupied by the plasma. In traps for the quasi-stationary confinement of plasma, the magnetic fields either are constant or are changing only slowly with time.

Pinch principle. Apart from systems for the quasi-stationary confinement of plasma, much effort is also directed toward developing methods in which the plasma is heated very rapidly in the process of being compressed by a quickly rising magnetic field. This so-called fast-pinch principle is capable of leading to very high concentrations of energy in small volumes of space over very short intervals of time. The magnetic fields in these rapid-compression devices are of the impulse type, implying that they reach their maximum strength within a few microseconds. There are several approaches to achieving the rapid heating of a plasma in a rising magnetic field. The scheme that is most developed at present involves the so-called theta-pinch, or Θ-pinch, system. In such a system the plasma is generated by ionizing a gas in a tube placed inside a coil of low inductance.

When a powerfully charged capacitor is connected to the coil, the very quickly mounting current within the completed circuit gives rise to a magnetic field pulse inside the coil. This field compresses the plasma while simultaneously heating it to a high temperature within a time interval of no more than a few microseconds.

State of progress and prospects for the future. *Traps employing magnetic mirrors.* Initial research was concentrated chiefly on systems of the simplest type with two magnetic mirrors as shown in Figure 4. This method, however, failed to achieve a sufficiently long containment of a dense enough plasma at high temperatures. The lifetime of a plasma in such a trap is limited to tens of microseconds because the strength of the magnetic field, while increasing along the lines of force, is attenuated in radial directions, and the plasma, being diamagnetic, tends to move toward the weaker regions of the magnetic field. If a slight bulge accidently develops somewhere along the lateral surface of the plasma in the central part of the trap, the bulge will tend to grow in a radial direction. Thus, any minor chance disturbances are likely to produce tongues of plasma extending further and further into the surrounding space. In almost no time the plasma will spread throughout the volume of the container and disappear on contact with the container walls.

Plasma tongues

This illustrates the main problem encountered in nearly all experiments aimed at preserving very hot plasmas for significant intervals of time. So capricious is this substance that it seeks to cast off the shackles of magnetic thermal insulation and wriggle its way out of the magnetic trap. The main effort in research is now directed toward dealing with the various instabilities of plasmas.

The large-scale instabilities observed in open magnetic traps of the simplest type, for example, could be reduced by changing the configuration of the magnetic field. But a truly stable confinement of plasma can be achieved only by having the field strength increase in any and all directions away from the region occupied by the plasma. When this is the case, the plasma is said to be in a magnetic well. One of the more plausible methods of effecting a system that would satisfy this requirement is shown schematically in Figure 6. The magnetic field is produced not only by the two coils but also by linear current-carrying conductors that are disposed symmetrically with respect to the common axis of the coils. The current flowing in the linear conductors alternates direction and, if the current intensity is sufficiently high, the magnetic field strength increases in all directions away from the centre of such a system.

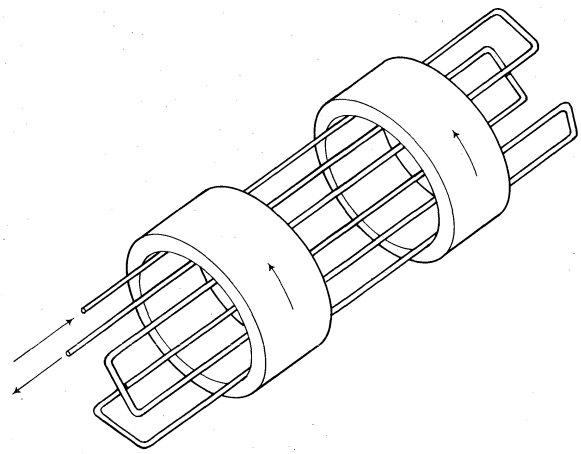


Figure 6: Advanced type of plasma trap, designed so that magnetic field strength increases in all directions away from central region. Short tubes represent a pair of coils wound in the same direction and are responsible for longitudinal increase in strength of magnetic field. Linear conductors, which carry current that alternates in direction, are responsible for radial increase in magnetic field strength. Arrows indicate direction of currents in coils and linear conductors, respectively.

The earliest experiments with such magnetic systems were carried out in the Soviet Union between 1961 and 1963. For the first time, it had become possible to maintain stable configurations of plasmas at temperatures of several tens of millions of degrees with concentrations of up to 10^{10} particles/cm^3, and it was then generally recognized that open traps with magnetic mirrors are feasible only in conjunction with a magnetic well design. All subsequent installations built in the U.S.S.R., the U.S., and Great Britain fulfill this requirement. In these installations it is now possible to produce stable plasmas with a temperature of about 10^7 degrees and with densities of up to 10^{13} particles/cm^3. The maximum values of the product $n\tau$ attainable in present-day mirror traps with strong magnetic fields range from 10^8 particle-sec/cm^3 to 10^9 particle-sec/cm^3. This, however, is still smaller, by as many as five orders of magnitude, than the minimum value of $n\tau$ necessary for the operation of a thermonuclear generator fuelled with a mixture of deuterium and tritium. The temperature also is still too low.

The earliest experiments

Realizing that only a few years ago the value of $n\tau$ in such systems was barely of the order of 10^4 particles-sec/cm^3, and considering the logarithm of $n\tau$ a measure of progress, it appears that half the route toward the ultimate goal via mirror-trap development has just about been covered. From the standpoint of future industrial utilization, however, traps with magnetic mirrors have a crucial drawback. Even if all types of plasma instability could be completely taken care of, such systems would suffer from a loss of particles along the lines of force, a disadvantage that cannot be eliminated in principle. In a trap with magnetic mirrors, it is possible to confine only those particles whose velocity vectors do not form too small an angle with the lines of force. Hence, if the direction of such a plasma particle's velocity has greatly changed as a result of its collisions with other particles, it may now be able to escape from the trap. The amount of energy carried away by the stream of particles from the trap because of collisions with other particles in the plasma would be likely to play a very important role in the energy balance of a future thermonuclear reactor. Rough estimates show that the reactor would have a positive energy yield only if a considerable part of the energy losses due to the escape of particles through the mirrors could be recovered. Because it is uncertain at present whether this is feasible, the future of thermonuclear reactors employing traps with magnetic mirrors also remains uncertain. That the principle of magnetic mirrors will eventually lead to the realization of the first thermonuclear reactor cannot yet be completely ruled out.

Closed magnetic traps. Systems employing closed magnetic traps fall into two groups. In the first group, the plasma is kept generally stable by the toroidal magnetic field due to a current circulating in the ring of plasma itself, while the more dangerous instabilities are supressed by a very strong externally generated field concentric with the ring of plasma. This latter field is produced by coils wound over the toroidal chamber that houses the plasma, as shown schematically in Figure 5. This is the method of plasma containment used in the Tokamak installations of the U.S.S.R. The heating of the plasma in such systems has been accomplished so far exclusively by the joule heat generated by the current flowing through the plasma. But, in principle, other methods of heating could also be used. With longitudinal magnetic field strengths of up to 35 kilogauss, it has now been possible, at the Tokamak installations, to obtain plasmas without appreciable signs of instability at electron temperatures in excess of 10^7 degrees, deuteron temperatures of about 5×10^6 degrees and densities of approximately 5×10^{13} particles/cm^3. With the mean lifetimes of particles reaching into the hundredths of a second, the product $n\tau$ becomes approximately 10^{12} particle-sec/cm^3. In the experiments at Tokamak, the stable plasma ring produces a long-lived neutron radiation that appears to be of thermonuclear origin.

The Tokamak installation

The other group of systems employing closed magnetic traps are the so-called stellarators. In stellarators the ring of plasma is maintained within the toroidal chamber by only an external field that has, however, a complex structure. The design is such that the lines of force are compelled to undergo what is known as a rotational transformation, amounting to having the lines of force wind continuously around the annular axis of the toroidal chamber as one moves along it. The investigation of the properties of plasmas in stellarators is a key element in the research program on controlled nuclear fusion in the United States. The complex geometry of the magnetic field, however, has made it very difficult to suppress the instabilities of the ring of plasma and, for this reason, it has not yet been possible to achieve in such systems the containment of plasmas for sufficiently long intervals of time.

Stellarator systems

In any event, the construction of thermonuclear power plants based on any one of the quasi-stationary closed-trap type of reactors would face tremendous technical difficulties because such high magnetic field strengths are required in conjunction with such large installations.

Impulse compression of plasma. The greatest interest in this line of research has for some years centred on systems of the Θ-pinch type, in which the plasma is compressed by a rapidly increasing external magnetic field. The study of such systems originated in the United States and then spread to a number of laboratories in Great Britain and the Federal Republic of Germany. With magnetic fields rising to 1×10^5 gauss within a few microseconds, it is now possible to produce plasmas at temperatures ranging from 1×10^7 degrees to 3×10^7 degrees and densities ranging from 10^{16} particles/cm^3 to 10^{17} particles/cm^3. The time interval for which such high temperatures can be maintained depends on the dimensions of the apparatus, that is, on the length of the coil in which the process of compression takes place. With coil lengths of about one metre, τ ranges from five to ten microseconds, and the product $n\tau$ reaches a value of about 10^{11} particle-sec/cm^3. The high temperatures, together with the high particle densities in these Θ-pinch systems, produce conditions that give rise to neutron radiation. If it were possible to overcome some of the difficulties of converting the open Θ-pinch system to a closed ring design, a manifold increase would be achieved in the duration of plasma confinement and this would make such a system highly promising.

The compression and simultaneous heating of plasmas can be accomplished by means other than that of an externally produced magnetic field. In the earlier stages of nuclear fusion studies, much interest was expressed in the possibility of producing high temperatures by the rapid compression of plasmas in what came to be known as linear pinch systems. In this case a filament of plasma is compressed in the course of a very short time interval by the magnetic field accompanying a strong current that flows through the plasma filament itself. It was found, however, that because of the instability of a plasma filament carrying a strong current, the values of τ in linear pinch systems tended to be very small. According to the criteria discussed above, therefore, a positive energy yield for a thermonuclear reactor based on the linear pinch system could be brought about only by creating a very dense plasma and passing a gigantic current through it. This, in turn, would imply the practically instantaneous concentration within the plasma of an enormous amount of energy, followed by its practically instantaneous liberation; that is, the process would assume the character of an intense explosion. Despite such discouraging implications, however, continued study has shown that, if the strictly cylindrical geometry of the system is replaced by a noncylindrical design, it is possible to obtain the formation of a plasma focus in the very small volume of approximately 0.01 cm^3 and to achieve within this focus a density of roughly 10^{19} particles/cm^3 at temperatures reaching some 10^7 degrees. Such a plasma focus can be maintained for about 10^{-7} seconds, so that the product $n\tau$ attains a value of 10^{12} particle-sec/cm^3. The plasma focus is an impulse source of intensive neutron radiation. Currently, experiments with plasma foci are being conducted in the U.S.S.R., the U.S., Great Britain, and Italy.

Linear pinch systems

Technological developments of recent years have opened up new avenues for producing high temperatures in plasmas by processes of extremely short duration. The impulse heating of plasmas to very high temperatures can be achieved, for example, by means of laser beams (see LASER AND MASER). In principle, at least, this offers the possibility of generating thermonuclear reactions of considerable intensity without the necessity for thermally insulating the plasma by magnetic fields. Modern laser technology makes it possible to produce well-focussed beams of light carrying large amounts of energy for time intervals shorter than 10^{-9} second. In such rapid heating, the energy losses of the plasma through thermal conduction and diffusion cannot play an appreciable role. Nevertheless, it can be anticipated that the explosive character of energy generation in any system employing this principle would lead to technical difficulties comparable to those expected in the case of cylindrical linear pinch systems.

Heating by laser beams

When can a solution to the problem of controlled fusion be expected? Perhaps only when mankind's need for new sources of power becomes so acute that all possible approaches are investigated.

BIBLIOGRAPHY. L.H. ALLER and D.B. MCLAUGHLIN (eds.), *Stellar Structure* (1965), a monograph on all contemporary aspects of the structure of stars, including the question of the dissipations of stellar energy, extensive bibliography; REPUBLIC OF INDIA, MINISTRY OF INFORMATION AND BROADCASTING, *Nuclear Explosions and Their Effects*, 2nd ed. (1958), popularized exposition of the physical bases of atomic and thermonuclear explosions, and of their destructive effect; L.A. ARTSIMOVICH, *Controlled Thermonuclear Reactions* (1964; pub. orig. in Russian, 1963), a monograph containing an exposition of the theoretical and experimental bases of physics of high-temperature plasma and its relationship to the problem of controlled thermonuclear syntheses; H. MARK and S. FERNBACH (eds.), *Properties of Matter Under Unusual Conditions* (1969).

(L.A.)

Nuclear Reactor

A nuclear reactor is a device designed to permit self-sustaining and controlled nuclear fissions, with the object of generating heat, producing radioactive isotopes or plutonium, developing an intense field of nuclear radiation, or serving some other useful purpose. Nuclear fission (*q.v.*) is the phenomenon in which an atom of fissionable material disintegrates when struck by a neutron, producing two entirely different atoms and generating a large amount of heat. In the fission process, neutrons are also given off, and these neutrons can cause more atoms to fission, thus leading to the possibility of a chain reaction. In the atom bomb this chain reaction is uncontrolled; in a nuclear reactor it is very carefully controlled.

The amount of heat generated in the fission process is very large. If all the atoms in a pound (0.45 kilogram) of uranium-235 were to undergo fission, the heat produced would be equivalent to burning 1,500 tons of coal. This phenomenon is the basis of the major application of nuclear reactors: the production of large amounts of heat for electrical power generation.

Whenever nuclear fission occurs, the two atoms that are produced are radioactive, often intensely so. They may spontaneously release highly penetrating gamma rays (X-ray-like radiation) and generally less penetrating beta rays (electrons). A nuclear reactor, therefore, is potentially a source of heat, neutrons, and radiation.

This article is divided into the following sections:

NUCLEAR-REACTOR PRINCIPLES

The chain reaction. For a chain reaction to take place, it is necessary for at least one of the neutrons released during each fission to cause another fission to occur. This condition requires the release of a large number of neutrons because many escape the chain-reaction system altogether, and some, after collisions with atoms, become absorbed without causing fission.

The only atom found in nature capable of supporting a nuclear chain reaction is a rare isotope of uranium, uranium-235. Most of the atoms of which normal uranium is composed contain 238 particles in their nuclei. The 235 variety, which exists in the proportion of approximately one out of every 140 atoms, generally splits easily when a neutron collides with it. It is more likely to split if the neutron is moving at a relatively slow pace; faster neutrons tend to fly directly by without any effect. Neutrons released during fission have high speeds, so that one way to start a chain reaction is to find some means of slowing them down. A chain reaction can also be made more likely to occur by increasing the proportion of uranium-235 atoms in the fuel, a process known as enrichment. With sufficient enrichment, the chain reaction can occur without the slowing down of neutrons. Still another way to facilitate the chain reaction is to make the reactor core large enough to minimize the number of neutrons escaping its boundaries. For any particular reactor core there is a critical size below which the chain reaction cannot occur. This critical size decreases with the increasing enrichment in uranium-235. It may also be reduced by surrounding the core with a reflector, a material that reflects neutrons back into the core.

The critical core size

Finally, a chain reaction is more likely to be sustained if the number of neutrons captured by nonfissionable atoms is minimized. Certain kinds of atoms readily capture neutrons, particularly the slow ones. Materials composed of these atoms may have to be avoided insofar as possible. The nonfissionable portion of uranium, uranium-238, can absorb neutrons, particularly those at intermediate speeds, without producing fission. Fissions can occur with fast neutrons; nevertheless, this probability is so low that the number of neutrons supplied in this manner provide only a small contribution to the requirements of the chain reaction.

Moderation. The process by which neutrons are slowed down is called moderation. In the process, a moderator composed of light atoms with low-absorption cross sections is employed. The cross section may be thought of as the area surrounding the nucleus of the atom through which the neutron must pass to interact with the nucleus. When neutrons are released into the moderating material, they bounce from one atom to another, losing speed with each collision in a billiard-ball fashion. This action occurs only when the atoms with which they collide are small and do not absorb the neutrons. Deuterium, for example, is an excellent moderator because its atomic weight is low, only twice that of a neutron, and its absorption cross section is also low. Hydrogen is better from the standpoint of its mass, close to that of a neutron, but its absorption, or capture, cross section is somewhat higher than that of deuterium.

Slowing of neutrons. If uranium-238 is to be contained in the core of the reactor, the designer must consider its relatively high-absorption cross section at intermediate neutron speeds. For this reason the cores of reactors often consist of a lattice arrangement in which the uranium is spaced at regular intervals in the form of rods embedded in a moderator. This approach has led to the heterogeneous-reactor type in which the fuel is in the form of discrete rods or plates, allowing the neutrons released at high speeds to escape the uranium quickly and then bounce about in the moderator before re-entering the fuel rods. Neutron speeds are thus reduced to the point at which their chances of capture by a uranium-238 atom are reduced, and their chances of capture by a uranium-235 atom are increased.

Reactor control. There would be an excessive release of heat if the chain reaction were to proceed at too great a rate. To control the reaction, a material that has a high neutron-absorption cross section, such as cadmium or boron, is inserted in the core, often in the form of rods. The reactivity of the core increases when the rods, called control rods, are withdrawn. When the control rods are inserted as far as possible, the reaction is completely shut down, and any further chain reaction is prevented. Another method of control requires varying the amount of fuel, but in this case withdrawing the fuel rods reduces reactivity. Control of the reaction may also be achieved by varying the amount of a reflector or the amount of a moderator.

Control rods

Heat removal. Nuclear fission releases energy in the form of heat. If the power of the reactor and the rate at which fissions occur are kept low, the core may become only slightly warm, and no special provisions will be required to remove the heat. On the other hand, high power levels require elaborate heat removal systems inside the core. In these systems cooling fluids, such as water, gases, or liquid metals, are circulated through the core. The fluids must be composed of materials that have suitable nuclear properties so that not too many neutrons will be absorbed, and they must also have suitable flow and heat-transfer properties.

Shielding. To protect personnel in the vicinity of the reactor from radiation, a shield is installed around the core to absorb escaping neutrons and gamma rays from the fission and the fission fragments. The amount of shielding required depends on the power level at which the reactor is operated and the length of time during which the radioactive fission fragments accumulate. Some shielding is provided to protect various materials, the physical properties of which are affected by radiation.

NUCLEAR-REACTOR DEVELOPMENT

The first reactors. *Two atomic-bomb designs.* Even before the World War II atomic bomb project was organized in the United States as the Manhattan Project, two different ways to achieve a nuclear chain reaction had been considered. Later, these provided the basis for two different ways to make an atomic bomb, and both methods worked.

In one design, a chain reaction is achieved by separating the fissionable uranium-235 from the more abundant uranium-238. With a sufficient amount of enriched uranium (the critical mass), a chain reaction occurs if a moderator is not used (*i.e.*, without slowing down the neutrons to a point such that the probability of neutron-causing fission becomes much higher). By concentrating the fissionable material, they believed that they could accept the lower fission cross section, which allows neutrons to bypass atoms more easily without causing fission. When the chain reaction proceeds so rapidly that all or most of the atoms undergo fission before being blown apart, a nuclear weapon results. (See also NUCLEAR WEAPONS.)

The problems involved in separating uranium-235 are formidable. Their solution became one of the major accomplishments of the Manhattan Project. Success came first at a large gaseous diffusion plant at Oak Ridge, Tennessee. The separated uranium-235 was incorporated in the atomic bomb that was exploded at Hiroshima on August 6, 1945.

Another method for producing a chain reaction is the use of natural uranium with a moderator that takes full advantage of the high fission cross section of uranium-235 for slow neutrons. It was not recognized at first that this method could also lead to an atomic bomb because the pertinent discovery had not yet been made. Scientists knew that uranium-238 has a relatively high capture cross section for neutrons with intermediate speed and that the lattice design of fuel elements embedded in a moderator would minimize losses of neutrons into uranium-238. What had not been established was that the absorption of a neutron into uranium-238 leads to the formation of a new type of fissionable atom, one not normally found in nature. When a uranium-238 atom absorbs a neutron, it becomes uranium-239, which is unstable and emits an electron. It thus transmutes itself from uranium to another element, called neptunium, with an atomic number of 93 instead of 92. Neptunium, in turn, is also unstable, leading to a similar transmutation to plutonium, with an atomic number of 94. Because both reactions occur rapidly, the plutonium exists within a few days after the neutron has been absorbed by uranium-238. The plutonium isotope produced, plutonium-239, is relatively stable and fissionable. It releases approximately the same number of neutrons as does uranium-235; *i.e.*, between two and three on the average for each fission, and its fission cross section for slow neutrons is even higher than that for uranium-235.

With these discoveries it became apparent that the production of plutonium could be a substantial benefit to be gained through the capture of neutrons by uranium-238. For this purpose it would be necessary to design a reactor to favour such captures, but without using up so many neutrons that the chain reaction could not proceed.

Though a natural uranium chain reaction is a possible source of heat and radiation, it is too bulky to serve as the basis for an atomic bomb. It may, however, be used to produce plutonium, which then can be fabricated into a bomb without having to rely on uranium-235. The design and construction of a large plutonium-producing nuclear reactor, therefore, had a high priority during the early days of the Manhattan Project; it was plutonium, in fact that was employed in the first nuclear explosion in a nuclear-weapons test at Alamogordo, New Mexico, on July 16, 1945.

The Chicago pile. The world's first nuclear reactor was constructed in the United States at the University of Chicago under the direction of Enrico Fermi. It achieved criticality (a self-sustaining chain reaction) on December 2, 1942. The reactor consisted of 400 tons of graphite, six tons of uranium metal, and 50 tons of uranium oxide, with control rods made of cadmium. Instruments placed inside the pile measured neutron intensity. When the cadmium rods were withdrawn gradually, at a certain point the neutron intensity began to increase rapidly, signalling the start of a self-sustaining nuclear chain reaction. The cadmium rods were reinserted before any appreciable amount of heat developed and before the neutron and radiation levels became hazardous. It was later in the day that the historic news was transmitted by telephone from Arthur Holly Compton in Chicago to James B. Conant at Harvard University with the following guarded language: "Jim, you will be interested to know that the Italian navigator has just landed in the New World."

The world's first reactor

The experiment confirmed what had been expected: criticality was reached even more easily than anticipated. Plans were formulated for the construction of four other reactors in the U.S., and the work of harnessing the atom for peaceful purposes was underway.

Reactor types. *Thermal and fast reactors.* The long-run advantages of a reactor that operated with fast neutrons, without a moderator, had been recognized by 1946. With sufficient enrichment, criticality could be achieved despite the lower cross section for fission with fast neutrons. Because fissions produced by fast neutrons supply slightly more neutrons on the average, they provide extra neutrons beyond those required to keep the chain reaction going. Calculations showed that with sufficient uranium-238 in or near the core, the supply of neutrons that resulted from occasional fissions of uranium-238 would increase slightly.

More important, there would be a sufficent supply of the extra neutrons captured by the uranium-238 to produce plutonium faster than the uranium-235 was consumed. In other words, using one neutron per fission to cause the next fission, with allowances for escaping neutrons or losses caused by capture in structural materials, there would still be more than one neutron on the average remaining to be captured by uranium-238, which would then convert to plutonium-239. It thus became theoretically possible to consume uranium completely, thus tremendously increasing the value of uranium as a new source of energy.

To test these ideas, an experimental breeder reactor, the EBR-I, was built in Idaho by the Argonne National Laboratory. The reactor core was composed of fuel rods of highly enriched uranium-235, surrounded by a blanket of natural uranium to provide the fertile material in which the plutonium could be formed.

The breeder reactor

Operating successfully for the first time during 1951, the reactors demonstrated not only the feasibility of breeding but also the feasibility of generating electricity with nuclear power. Reactor heat was transferred with a liquid metal (a mixture of sodium and potassium, called NaK) to a steam generator that provided power for a small electric turbine.

The two reactor concepts, one making use of slow neu-

trons (generally referred to as thermal neutrons), and the breeder concept, utilizing fast neutrons, were forerunners of two fundamental classes of reactors: thermal reactors and fast reactors. Most of the wide variety of reactor types that have since been developed are of these two classes.

Thermal power reactors (Figure 1) have been developed to the point that they are commercially attractive as energy sources for large power plants, but, in the early 1970s, fast reactors had not yet reached that stage of development.

By courtesy of the Atomic Energy Commission

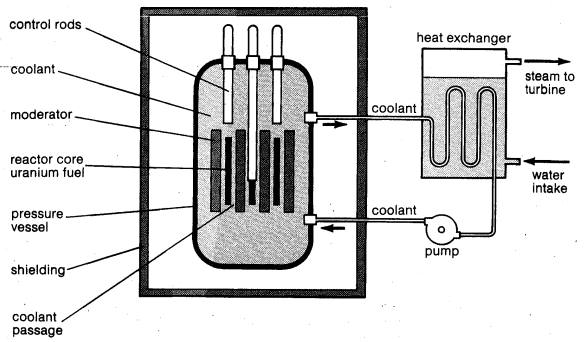


Figure 1: Components of a power reactor designed to generate high-pressure steam.

Plutonium-production reactors. The production of plutonium for military purposes was expanded in the United States after World War II with the construction of additional graphite-moderated reactors and a new series of large heavy-water reactors (HWR). (Heavy water is water in which the conventional hydrogen atoms have been replaced by deuterium, the hydrogen-2 isotope.) The heavy-water reactors were also used to convert lithium into tritium (the hydrogen-3 isotope) for use in thermonuclear weapons.

Materials-testing reactor. One of the early needs in reactor development was a materials-testing reactor that could supply a high flux of neutrons to test their effects on materials and thus provide necessary information for the design of other reactors. Neutron flux is the number of neutrons crossing a unit area each second.

Pressurized water was selected to serve as both moderator and coolant in the first materials-testing reactor. The fuel was highly enriched uranium embedded in aluminum, arranged in the form of parallel plates. Water was pumped through at high speeds to carry heat away, thus allowing the reactor to operate at a high power density with a correspondingly high neutron flux. The reactor was placed in full-scale operation in 1952.

Aircraft-propulsion reactor. The use of nuclear power for aircraft propulsion was proposed immediately after the end of World War II. The U.S. Air Force began a project known as Nuclear Energy for the Propulsion of Aircraft (NEPA); but controversy over the potential hazard began at once, and, as the complexities of the problem became more evident, the project was discontinued in 1951.

Responsibility for the design of an aircraft engine was shifted to the General Electric Company and the National Reactor Testing Station in Idaho where test facilities could be built. But in the end enthusiasm for the project was overcome by doubts about feasibility and safety. The goal of planes that could fly for long periods without refuelling did not seem to justify the development expense nor the hazards that would inevitably be associated with carrying nuclear reactors in airplanes, and the project was abandoned in 1961.

Reactors for a nuclear navy. The use of nuclear reactors for the propulsion of naval vessels has been highly successful. Although it was not directly involved in nuclear matters during World War II, the U.S. Navy was interested in nuclear energy before the establishment of the Manhattan Project as the result of research undertaken at the Naval Research Laboratory.

Nuclear energy for submarines

Studies of the use of nuclear reactors as a source of energy for submarine propulsion began in 1946. The potential advantages over oil as a fuel were obvious. With nuclear fuel, it would no longer be necessary to surface to recharge batteries, and the energy from uranium for a given bulk would be far higher than that from oil, thus making possible almost indefinite operation without refuelling.

Two submarine-propulsion projects were formulated, the submarine intermediate reactor and the submarine thermal reactor. The submarine-intermediate-reactor concept involved the use of enriched-uranium pins as fuel with liquid sodium as a coolant and with intermediate-energy neutrons. This concept was thought to have a long-range potential for breeding and electric-power generation as well as for submarine propulsion. It soon became clear, however, that the requirements for submarine propulsion are different from those for electric-power generation, and military requirements taking precedence, the reactor was designed solely for submarine propulsion. The "Seawolf," a submarine possessing the submarine-intermediate-reactor type of propulsion, was operated successfully from 1956 until it was shut down in 1959.

The concept was not adopted by the Navy because of the spectacular success of the water-cooled and -moderated reactors that had been achieved in the meantime by the submarine-thermal-reactor concept. Use of pressurized water both as coolant and moderator was advantageous because the Navy was thoroughly familiar with the material. Because operation at high pressure was required to achieve reasonable efficiency, new problems arose, centred on control, corrosion, fuel-element fabrication, and shielding. Materials had to be selected that would maintain their integrity under intense radiation by both gamma rays and neutrons and that would not themselves absorb too many neutrons. To provide the necessary compactness, it was necessary to use highly enriched uranium-235.

Homogeneous reactor. Another significant concept introduced immediately after World War II was that of the homogeneous reactor, in which uranium, usually in the form of uranyl sulfate, was mixed with a water moderator to form a slurry that was continuously circulated to remove fission products and plutonium, thus avoiding the problem of fuel fabrication. With pressurization, the system could become hot enough to be used as a source for the generation of electricity. One of the major difficulties was finding materials that could withstand the corrosiveness of the slurry. A small, homogeneous, experimental reactor was built and operated during 1952, and in February 1953, it became the second reactor to demonstrate the technical feasibility of converting nuclear to electrical energy.

Reactor development worldwide. During the early period after World War II, reactor development was also studied in other countries, notably the United Kingdom, Canada, France, and the Soviet Union.

With assistance from French and English scientists, Canada had a small, natural-uranium, heavy-water-moderated reactor operating at Chalk River, Ontario, in 1945. By 1947, another reactor, also using natural uranium and heavy water, was able to supply what were at that time extremely high neutron fluxes.

Later, this reactor was used by Argonne National Laboratory scientists to test materials for large, heavy-water-moderated plutonium- and tritium-production reactors. This second reactor was used to produce experimental quantities of uranium-233 from thorium as well as plutonium from uranium-238.

Reactor development in the United Kingdom

Before 1942, the United Kingdom had discussed with the United States the possibility of making atomic bombs. From that period until the close of the war, however, its role was essentially limited to the Canadian project and to some participation in the work at the Los Alamos Scientific Laboratory. Immediately after the war, the British, assigning a high priority on plutonium production for weapons development, planned to use graphite-moderated, air-cooled reactors. Two small reactors of

this type were operating for research purposes by 1948, and two plutonium-production reactors were operating successfully at Windscale by 1951, when it was decided to combine plutonium and power production in subsequent reactors, cooling them with carbon dioxide instead of air. The resulting series of Calder Hall reactors, as they were known, gave the United Kingdom an early lead in the commercial production of electric power and at the same time enabled it to build a stockpile of plutonium.

By 1948, France had a heavy-water-moderated natural-uranium reactor, similar to Canada's first, for research purposes. A plutonium-production policy similar to that of the United Kingdom was adopted; by 1956 the first of a series of reactors was being operated as a small, dual-purpose plant. This was air-cooled, but subsequent reactors of this type also switched to carbon dioxide.

NUCLEAR POWER DEVELOPMENTS TO 1962

Industrial participation. By 1950, certain industrial leaders in the United States were impatient to participate in the development of nuclear power. Industry proposed that an atomic-power plant be built with private funds, to produce power and plutonium at the same location. This proposal led to an "industrial participation program" in which the Atomic Energy Commission (AEC), the successor of the Manhattan Project, allowed industrial study teams to have access to pertinent classified information in order to "examine the economic and technical aspects of building this (dual-purpose) reactor in the next few years." A variety of concepts was considered, ranging from graphite moderation with sodium cooling, graphite moderation with helium cooling, homogeneous systems with uranium-233 production instead of plutonium, to a fast-breeder reactor. Although the dual-purpose-reactor concept was finally abandoned in the United States as the basis for starting a private nuclear power industry, the studies undertaken laid the foundation for extensive industrial participation in reactor development.

The work of one of the study teams led eventually to the Enrico Fermi fast-breeder reactor at Monroe, Michigan. Another team, proceeding without direct government assistance, purchased a boiling-water reactor (BWR) with an electrical-power capacity of 208,000 kilowatts, at Morris, Illinois. It became, in 1959, one of the earlier power-demonstration plants (see Table 1).

The dual-purpose concept was abandoned as a spur to power production because of the reluctance to combine military and commercial objectives. One reason for the abandonment was the fact that high-priority-plutonium requirements would affect the reactor design and reduce the efficiency for power production.

The British, however, took a different view. They decided to build graphite-moderated, carbon-dioxide-cooled reactors to produce plutonium, for military purposes, while at the same time utilizing the heat to generate electricity. As a result, they became the largest producers of electricity from nuclear fuel, but the plants were not economical when used entirely for the production of electricity.

The dual-purpose concept was revived by the United States in 1958, when the government authorized the design and construction of a new production reactor with plutonium production as its primary purpose, but with the by-product, heat, to be sold to an electric-utility company in the form of steam to operate an 800,000-kilowatt electric-power plant. In this instance, the purpose was to reduce the cost of plutonium production. The reactor first became critical in 1963, and in 1966 the adjacent Washington Public Power Supply System generating station began to produce electricity. In 1971, with a declining need for plutonium, shutdown of the reactor was under serious consideration.

"Atoms for peace." On December 8, 1953, President Eisenhower recommended the establishment of the International Atomic Energy Agency. Shortly thereafter, he recommended that the Atomic Energy Act of 1946 be amended to provide more encouragement to the peaceful application of atomic energy. In 1954 a new Atomic Energy Act was adopted, making it possible for nuclear reactors to be privately owned under suitable licensing arrangements. Nuclear materials, plutonium, uranium-233, and uranium-235, were to be leased for private use. Reactor development was encouraged in other ways for peaceful purposes, and international cooperation in all peaceful applications of atomic energy was likewise encouraged.

The power-reactor development program. During the same year, the Atomic Energy Commission announced a power-reactor development program centred on five reactor concepts: pressurized water, boiling water, sodium–graphite, homogeneous, and fast breeder.

Five reactor concepts

The first of these was a pressurized-water reactor to be built immediately at Shippingport, Pennsylvania, as a plant with 60,000 kilowatts of electrical power capacity with steam to be sold to an electric-power company. The technology was derived largely from experience with the submarine thermal reactor.

The concept of a boiling-water reactor was derived from experiments at the Argonne National Laboratory with a small reactor that had been allowed to boil the water surrounding the core on a continuous basis, thus suggesting the possibility of direct coupling to a steam turbine instead of using hot pressurized water from the reactor to produce steam in a boiler separate from the reactor. The experiment had shown that if the control rods were abruptly removed, thus causing a sudden power expansion, the boiling water became steam so rapidly that the moderating effect of the water was lost and the chain reaction came to a stop long before the fuel elements were damaged.

The homogeneous reactor was to be based upon earlier experience at Oak Ridge. It would be slightly larger, and would hopefully serve as a prototype for a still larger plant.

The sodium–graphite reactor was intended to combine the well-known graphite technology as the moderator with sodium cooling to achieve higher temperatures for more efficient power production, without the inconvenience of the high pressures associated with water cooling.

Finally, the potential advantages of a fast-breeder reactor were well recognized. With the success of the original experimental breeder reactor (EBR-I), it was decided that the Argonne National Laboratory should build a larger version (the EBR-II), also to be in Idaho at the National Reactor Testing Station. It would be scaled up to about 60,000 kilowatts of heat with an electrical production of about 15,000 kilowatts. It would be loaded first with uranium-235 and later with plutonium in order to enable it to produce larger amounts of plutonium in the uranium blanket.

By the end of the decade, each of these five experimental reactors was operating and another had been added. A small reactor moderated with organic material had been tried because of its potential for producing fairly high steam temperatures at relatively low pressures. A major disadvantage, however, was the low heat-transfer properties of the organic material and its tendency to decompose and polymerize (combine two or more small molecules into large ones).

Except for the homogeneous-reactor concept, each of the experimental reactors led to the design and construction of industrial prototypes or demonstration reactors. Some of these plants were large enough to produce a significant amount of power, but there was still no proof that the costs could be made competitive with power plants burning fossil fuels.

The European Atomic Energy Community. Prompted largely by a growing shortage of coal and oil, the six Common Market countries of Europe in March 1957, ratified the establishment of the European Atomic Energy Community (Euratom). Earlier, representatives from West Germany, France, and Italy had visited the United States and had discussed plans for a nuclear power program in Europe. Their report, "A Target for Euratom," noted the growing dependence of Europe on en-

Founding of Euratom

Table 1: Nuclear Electric Power Plant Prototypes
(1962)

name and owner	location	type	power		start-up
			plant kw (*) net	reactor kw (†)	
Shippingport Atomic Power Station (AEC and Duquesne Light Company)	Shippingport, Pa.	pressurized water	60,000	231,000	1957
Dresden Nuclear Power Station (Commonwealth Edison Company)	Morris, Ill.	boiling water	208,000	700,000	1959
Yankee Nuclear Power Station (Yankee Atomic Electric Company)	Rowe, Mass.	pressurized water	161,000	540,000	1960
Indian Point Unit No. 1 (Consolidated Edison Co. of New York, Inc.)	Indian Point, N.Y.	pressurized water	255,000	585,000	1962
Hallam Nuclear Power Facility, Sheldon Station (AEC and Consumers Public Power District)	Hallam, Neb.	sodium–graphite	75,000	240,000	1962
Big Rock Nuclear Power Plant (Consumers Power Company)	Big Rock Point, Mich.	boiling water	47,800	157,000	1962
Elk River Reactor (AEC and Rural Cooperative Power Association)	Elk River, Minn.	boiling water	20,000	58,200	1962
West Germany (Rhine-Westphalia Power Company, RWE)	Kahl-am-Main	boiling water	15,000	60,000	1960
Belgium (Center for the Study of Nuclear Energy, CEN)	Mol	pressurized water	11,500	43,000	1962

*Electrical output. †Thermal output.

ergy imports and the crisis that the threatened closing of the Suez Canal posed to obtaining oil from the Middle East. With rising fuel costs and rising demand for electric power, they were enthusiastic about the advent of nuclear power as a means by which Europe could become less dependent on fuel imports. They proposed that Euratom set a target of 15,000,000 kilowatts of installed nuclear power capacity to be built during the next ten years.

It appeared, therefore, that the conditions for nuclear-power development in Europe were more favourable than those in the United States and indeed that European experience might become a proving ground from which American power reactor development could benefit.

These expectations were not realized, partly because of the increasing availability of coal and the discovery of new oil and gas reserves. France had the beginnings of a nuclear-power industry based upon the dual-purpose graphite-moderated, carbon-dioxide-cooled-reactor approach. Italy opened negotiations to buy plants from both the United States and the United Kingdom and was the first to propose a nuclear-power plant of substantial size to be built under arrangements between Euratom and the United States. These initiatives, however, fell far short of what the Europeans had visualized. The Euratom proposal was dependent upon decisions by utility companies, and the uncertainties in the cost and availability of nuclear fuel, the complexities of nuclear-liability insurance, together with the high capital costs, discouraged the switch from fossil to nuclear fuels.

By 1962, five years after Euratom's founding, two firm Euratom power projects, and a third underway, provided altogether about 700,000 kilowatts of capacity. Obviously the earlier goal was not to be achieved, and there had been only a beginning in the development of an international market for nuclear-power equipment. During the same period, Japan, with its dependence on fuel imports, recognized the potential advantages of nuclear power and began to explore the possibility of purchasing both nuclear technology and equipment from the United States and from the United Kingdom.

Progress in reactor development. By 1962, 233 American-built nuclear reactors were being operated. Of these, 18 were directed toward electric-power production with half of them at the prototype stage. There were 21 test reactors and 116 research or teaching reactors, of which 37 were located in foreign countries. There were 42 naval propulsion reactors, with 51 more under construction. Thirteen large materials-production reactors were operating with plutonium as the primary product.

It was evident that substantial progress in reactor development had been made, but economically competitive nuclear-power plants had not yet been achieved. The nine prototype power reactors are described briefly in Table 1. In addition, experiments at various stages were underway with a variety of reactor concepts.

Boiling reactor experiment no. 5 (Borax-5). This was the fifth in a series of experimental reactors built to study boiling-water reactors and to demonstrate the feasibility of generating superheated steam with a nuclear reactor. In order to obtain the higher efficiencies associated with higher temperatures, saturated steam is produced in one set of fuel elements in the core and then recirculated through a second set in the same core to obtain the superheated steam.

Superheated-steam production

Experimental boiling-water reactor. This reactor was designed to experiment with gradually increasing power outputs. After having been designed to boil water at a rate of 20,000 kilowatts of heat, it was found to be possible to raise the level to more than 100,000 kilowatts. In addition, the reactor was used for experimentation with plutonium as a fuel.

Organic-moderated-reactor experiment. Organic compounds, such as diphenyl, were used both as moderator and coolant. Such compounds have the advantage of a relatively high boiling point, thus permitting higher temperatures without so much pressurization as required by water.

Heavy-water components-test reactor. This experimental pressurized heavy-water reactor was directed toward heavy-water-power-reactor development. Test irradiations were made of fuel elements composed of natural or slightly enriched uranium.

Plutonium-recycle-test reactor. This was a heavy-water-moderated and -cooled reactor with the cooling water circulated through pressure tubes. It had a heat capacity of 70,000 kilowatts, and its primary purpose was to test the use of plutonium as an alternate for uranium-235 as a fuel in power reactors.

NUCLEAR POWER DEVELOPMENTS AFTER 1962

Two years after 1962 forecasts of a growing need for nuclear energy to supplement fossil fuels, economically competitive nuclear power was signalled for the first time by the announcement of plans for the private financing and construction of large-scale nuclear-power plants that would produce electricity at lower costs than would be possible with fossil fuels. The first of these was a boiling-water reactor at Oyster Creek, New Jersey, with an elec-

trical capacity of 515,000 kilowatts. This was followed by many other orders for large plants, both boiling-water and pressurized-water types, some ranging up to the 1,000,000-kilowatt range.

By the end of March 1971, 20 nuclear-power plants with a combined capacity of 8,121,000 kilowatts were licensed for operation in the United States. An additional 55 plants with a combined capacity of 46,221,000 kilowatts were under construction, including one of the high-temperature, gas-cooled type. Applications for the construction of 25 more plants had been filed. In 1964 the U.S. Atomic Energy Act was amended to provide for a gradual transition from government to private ownership of nuclear fuel for power plants. A variety of projects directed toward power reactors with higher ratios of conversion of uranium to plutonium and with higher thermal efficiencies was supported for a while. By the end of the decade, with the growing commercial acceptance of the water reactors, the AEC shifted its top priority in reactor development to fast breeders. Most of the other programs involving different reactor concepts were either abandoned or placed on standby.

Economically competitive nuclear power

Thermal reactors. *Pressurized-water reactors (PWR).* The first reactor of this type was the submarine thermal reactor (STR) built in Idaho in 1953, which had led directly to the first nuclear-powered submarine, the USS "Nautilus." As the name implies, a pressurized-water reactor is both cooled and moderated by water under high pressure, thus permitting high temperatures. As indicated in Figure 2A, the hot water is pumped from a pressurized vessel containing the nuclear core to a steam generator in which heat is exchanged to produce the steam that drives a turbogenerator.

The fuel is arranged in a lattice, usually enriched in commercial reactors to 3 to 4 percent of uranium-235, and fabricated in the form of rods composed of uranium dioxide pellets. Although uranium dioxide has the disadvantage of a relatively low heat conductivity, it was chosen as the preferred chemical form of the fuel because of its resistance to radiation damage, thus increasing the time that the fuel may be left in the reactor. Both stainless steel and a zirconium alloy have been used for fuel cladding (a coating designed to contain fission products released in the fuel). They both do well in keeping fission products within the fuel, even though the accumulation of gaseous fission products causes pressures to build up.

Fuel enrichment depends on various factors, such as the size of the core, the power level at which the reactor is operated, the amount of moderation (particularly in relation to the total amount of fuel), the diameter of fuel elements, the nature of the cladding, the length of time the fuel remains in the reactor, and the fuel-cycle costs, including the credit derived from plutonium production.

The diameter of the rods influences the rate at which heat can flow from the fuel through the cladding to the pressurized water, which in turn determines the power level at which the reactor can safely operate. In determining the rod diameter, consideration is given to the maximum allowable temperature in the centre of the fuel rod (about 4,800° F, 2,650° C) and the temperature gradient through the rod and the cladding to the pressurized water. The space between the fuel and the cladding is filled with helium to facilitate the heat flow. The rod diameter also affects the speed of neutrons within the uranium, and, therefore, the proportion that is absorbed in the uranium-238 and produces plutonium.

Reactor shielding

The shielding of the reactor must keep heat losses and radiation levels external to the reactor down to acceptable levels. In addition to thermal shielding, neutrons and gamma rays must be absorbed. Typically, a "core barrel" also serves to confine the coolant flow within the core region, and is enclosed in a thermal shield, a pressure vessel, a water shield against neutrons, and a blanket of reinforced concrete for gamma-ray absorption; there may be a second concrete blanket, and, finally, the containment wall. In the newer pressurized-water reactors, fuel burnup has been improved to the point that fuel assemblies may be left in the reactor for years before refuelling is required.

Boiling-water reactors (BWR). These reactors have much in common with the pressurized-water reactors (Figure 2B). The main difference is that the intermediate steam generator is omitted, and steam is supplied directly from boiling water in the reactor core. Less pressurization

By courtesy of (A) Westinghouse Electric Corporation; (B) General Electric Company; (C) Gulf General Atomic Company; (D) *Chemical Week*, May 25, 1968

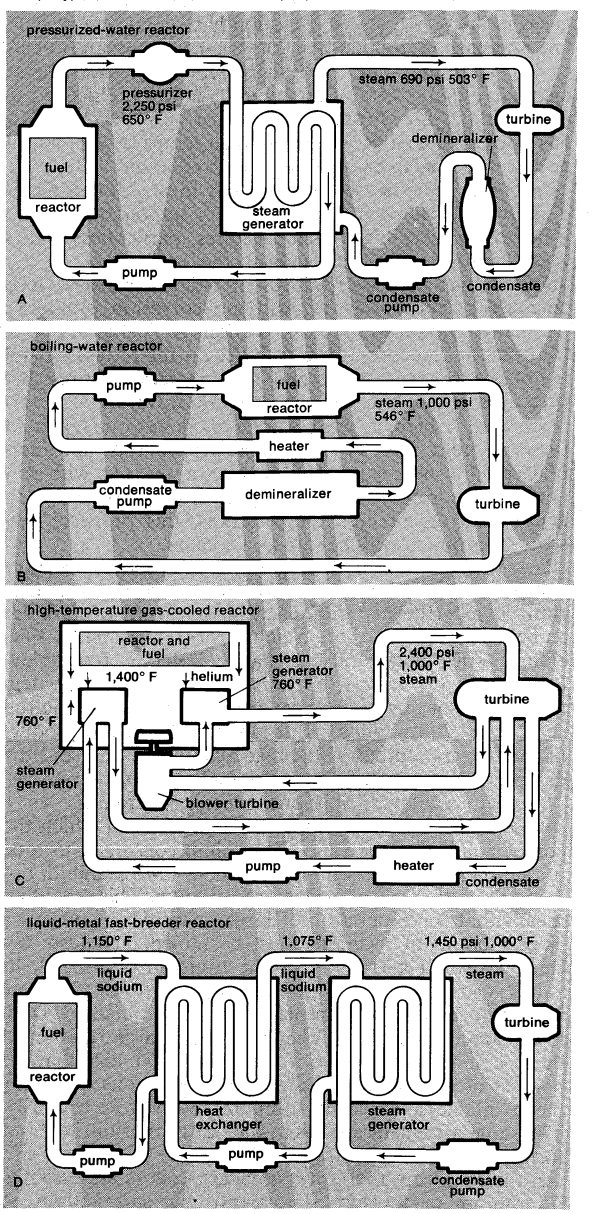


Figure 2: Nuclear-reactor systems of greatest importance for central-station power (see text).

is needed because the water is allowed to boil, and less pumping is needed because of the large amount of heat absorbed by boiling water. An excellent safety feature is the increase in steam production that results from an increase in power level, thus reducing the water volume and lessening its moderating ability, a condition that in turn reduces the reactivity. In other words, an inadvertent power increase tends to be self-correcting.

High-temperature gas-cooled reactors (HTGR). The conversion of heat to electricity can be accomplished more efficiently if the heat can be generated at higher temperatures, thus decreasing the percentage of the total heat that must be rejected. One of the obvious advantages of a nuclear reactor for power production is that it can be brought to as high a temperature as the materials of which it is made will permit. Limiting factors include melting, change of shape, and corrosion. These effects are sometimes exaggerated under irradiation; and if the ma-

terials deteriorate to the point that they interfere with the flow of heat out of the reactor as fast as it is produced, then the temperature of the reactor may increase to unacceptable levels.

The use of gas as a coolant is one method of achieving high temperatures, even though its heat-transfer properties are generally less favourable than liquids (Figure 2C). At higher pressures, the heat-transfer properties of gases improve. Most consideration has been given to helium and carbon dioxide. Air has been tried because of its easy availability, but its heat-transfer properties are so ineffective that excessive amounts of power are needed to circulate the air through the reactor. Furthermore, air has poor chemical properties at high temperatures, and it becomes radioactive under neutron irradiation.

Carbon dioxide is used, particularly in the United Kingdom and in France, in power reactors, although its chemical properties at higher temperatures, particularly under intense radiation, tend to be unsatisfactory.

Helium has the disadvantage of being expensive. It is chemically inert, however, with no corrosion problems; its neutron-absorption cross section is negligible; it is transparent, thus providing visibility during refuelling and maintainance operations; and at high temperatures and pressures its heat-transfer properties improve.

Graphite as a moderator

Graphite, which is used as the moderator in each of these reactors, has good mechanical properties and thermal stability and is a good conductor of heat with low neutron absorption. Although graphite is chemically reactive with air at high temperatures, this property presents no problem with helium as a coolant.

The fuel in the high-temperature gas-cooled reactors consists of highly enriched uranium, together with thorium as a fertile material, each in the forms of carbides embedded in pyrocarbon, a dense form of graphite. Fuel elements of this general type have a potential for achieving very high burnups, in the range, perhaps, equivalent to 500,000 megawatt-days per metric ton (one metric ton = about 2,200 pounds). Burnup is equal to the product of the thermal output of the reactor in megawatts and the days of operation divided by the fuel load in metric tons.

Breeder reactors. Without breeding, uranium as a source of energy is limited. Even in reactors in which the conversion ratio (the ratio of uranium consumed to total uranium employed) is improved, there is no possibility of using more than a few percent of uranium. This condition means that for generation of large quantities of power, a large amount of uranium must be mined, and also, because so small a fraction of the uranium is utilized, its value is so low that it is not economical to use the more abundant but more expensive uranium obtained from low-grade ores. On the other hand, if a large fraction of the uranium is consumed, it becomes so valuable that the poorest ores may be mined economically, thus extending the available and usable uranium reserves enormously. The same argument applies to the use of thorium reserves through conversion to uranium-233.

In the long run, therefore, breeder reactors are essential if uranium or thorium is to contribute substantially to meeting the world's energy needs.

Liquid-metal fast-breeder reactors. The concept that was receiving greatest emphasis in the 1970s is a liquid-metal fast-breeder reactor (LMFBR) operating on the uranium–plutonium cycle. With liquid-metal cooling, higher temperatures and therefore higher efficiencies are obtainable. Sodium has been selected as the coolant because it has favourable heat-transfer properties, a low-absorption cross section for neutrons, and its atoms are heavy enough so as not to slow down the neutrons, thus allowing the reactor to become a fast reactor (Figure 2D).

The fuel is expected to be plutonium diluted by uranium, providing a use, at least for start-up, of the plutonium produced by thermal water reactors. Oxides of uranium and plutonium rather than their metallic forms are used because of their better resistance to radiation damage.

Fast-fuel-test facility

The second experimental breeder reactor (EBR-II), originally intended as an experimental prototype, was used instead as a facility for testing fuels and materials in a fast neutron flux. During the early 1970s a larger fast-fuel-test facility (FFTF) was being built at Hanford, Washington, in order to provide a higher flux capability for fuels and materials testing.

Gas-cooled fast-breeder reactors. This reactor represents an extension of the technology of the high-temperature gas-cooled reactor, but it differs in the absence of graphite as a moderator and in its use of the uranium–plutonium cycle instead of the uranium-233–thorium cycle. The core of the reactor is composed of metallic-clad pins of mixed uranium–plutonium oxide fuel. It has helium cooling with a potential for being used to drive gas turbines directly instead of working through a heat exchanger to supply steam.

The breeding ratio may be high enough so that the doubling time (*i.e.*, the time required to produce an amount of new fuel equivalent to the amount contained in the reactor and its associated fuel cycle at the outset) will turn out to be only about ten years.

One of the questions about this concept centres on the safety problem in avoiding excessive heating in the event that the helium cooling is accidentally interrupted. Even though the chain reaction in a reactor is stopped, the temperature of the fuel elements can increase because of the heat produced through the absorption of intense gamma rays from fission products. Other questions centre on the performance of the metallic cladding of the fuel elements at high temperatures.

The molten-salt reactor. An experiment has been performed as a step toward a molten-salt breeder reactor that would operate on the thorium–uranium-233 cycle and therefore need not be a fast reactor in order to breed. The fuel is a mixture of lithium-7 fluoride, beryllium fluoride, zirconium tetrafluoride, uranium tetrafluoride, and thorium fluoride.

The circulating molten fuel enters the reactor at 1,175° F (635° C) and leaves at 1,225° F (663° C). The molten-salt mixture is stable at these high temperatures and under the intense radiation, and it has a low vapour pressure, thus avoiding pressurization problems. There are, however, some substantial engineering problems associated with the circulation of intensely radioactive fluid at high temperatures.

Power reactors worldwide. The original Euratom target of 15,000,000 kilowatts of electrical capacity in its six member countries by 1967 was not realized. Instead, by 1971 it turned out to be a little more than 3,000,000, with about half of this capacity in France. By this time, the United Kingdom had more than 5,000,000 kilowatts. On a worldwide basis, the total capacity was approximately 20,000,000 kilowatts, with more than half of this total of installed capacity being outside of the United States. Aside from the United States, 48 plants were operating, of which twelve had been built by U.S. companies. Great Britain was the early leader in the development of nuclear-fueled electric-power plants, largely because its shortage of hydroelectric sources and problems with cost and availability of fossil fuels made nuclear power attractive.

In the early 1970s nuclear production of electric power was becoming competitive in many areas and promised certain advantages, such as less air pollution, and a large number of nuclear installations were in the planning or construction stage all over the world. A listing of these power reactors showing their types and capacities in kilowatts (electric) is given in Table 2.

Gas-cooled reactors. The program of the United Kingdom was still the largest in 1971 in terms of total electricity having been produced, with its reliance on gas-cooled graphite-moderated technology using natural uranium as fuel. Two reactors of this type have been sold to other countries, one to Japan and one to Italy.

Capital costs for this reactor type are high, but this disadvantage is partially offset by the building of reactors of high capacity and the achievement of high-load factors. In the early 1970s, advanced gas-cooled reactors were being built with slightly enriched uranium fuel instead of natural uranium, thus permitting a substantial reduction in reactor size. With this enrichment, it was also possible

Table 2: Nuclear Power Reactors in Countries Outside the U.S. (1971)

country and plant	type	capacity (kW)
Belgium		
BR-3	PWR (U.S.)	11,500
Canada		
NPD	HWR	22,000
CANDU (Douglas Pt.)	HWR	208,000
France		
G-2 (Marcoule)	gas/graphite	40,000
G-3 (Marcoule)	gas/graphite	40,000
Chinon-1 (EDF-1)	gas/graphite	84,000
Chinon-2 (EDF-2)	gas/graphite	230,000
Chinon-3 (EDF-3)	gas/graphite	480,000
SENA (Fr.-Belg.)	PWR (U.S.)	266,000
St. Laurent 1	gas/graphite	500,000
West Germany		
VAK (Kahl-am-Main)	BWR (U.S.)	15,000
KRB (Gundreimmingen)	BWR (U.S.)	240,000
HDR (Grosswelzheim)	BWR (Superheat)	25,000
MZFR (Karlsruhe)	HWR	50,000
AVR (Jülich)	HTGR (Pebble Bed)	15,000
KWL (Lingen)	BWR	250,000
KWO (Obrigheim)	PWR	324,000
India		
Tarapur-1	BWR (U.S.)	190,000
Tarapur-2	BWR (U.S.)	190,000
Italy		
Simea (Latina)	gas/graphite	200,000
SENN (Garigliano)	BWR (U.S.)	150,000
SELNI (Trino Vercellese)	PWR (U.S.)	247,000
Japan		
JPDR	BWR (U.S.)	11,000
JAPCO-1	gas/graphite	166,000
JAPCO-2	BWR (U.S.)	322,000
The Netherlands		
Dodewaard (GKN)	BWR	52,000
Spain		
Zorita-1	PWR (U.S.)	153,000
Sweden		
Agesta	HWR	9,000
Switzerland		
Beznau-1 (NOK)	PWR (U.S.)	350,000
U.K.		
Calder Hall	gas/graphite	200,000
Chapelcross	gas/graphite	200,000
Berkeley	gas/graphite	276,000
Bradwell	gas/graphite	300,000
Hunterston A	gas/graphite	320,000
Trawsfynydd	gas/graphite	500,000
Hinkley Point A	gas/graphite	500,000
Dungeness A	gas/graphite	550,000
Sizewell A	gas/graphite	580,000
Oldbury A	gas/graphite	600,000
Wylfa	gas/graphite	1,180,000
Dounreay	liquid-metal fast-breed reactor	14,000
Windscale	advanced gas-cooled	35,000
Winfrith	steam generating heavy water reactor	100,000
U.S.S.R.		
Beloyarsk		
first unit	water/graphite	100,000
second unit	water/graphite	200,000
Novo-Voronezh		
first unit	PWR	240,000
second unit	PWR	365,000
Ulyanovskaya	BWR	50,000

to switch to an oxide form of fuel rather than using natural uranium, with stainless-steel cladding to enclose the fuel elements instead of a magnesium alloy. These switches permit higher temperature operation and higher fuel burnups, both of which help to reduce costs.

The so-called Dragon Project located at Winfrith in the United Kingdom is a cooperative venture of 12 European countries with participation by the United States, Euratom, and the European Nuclear Energy Agency. It produces no electricity, but it has operated successfully at a power level of 20,000 kilowatts of heat.

CANDU reactors

Heavy-water reactors. The Canadian interest in heavy-water reactors (HWR) has led to a Canadian Deuterium Uranium (CANDU) series of power reactors, the first of which was located at Douglas Point, in Canada.

The use of deuterium instead of hydrogen in the moderator has the advantage of a lower neutron-absorption cross section, permitting the use of natural uranium in the oxide form. Furthermore, with a better neutron economy, the lattice of fuel elements may be more widely spaced, thus permitting the use of zirconium-alloy pressure tubes in which heavy water as a coolant is confined. The hot heavy water is pumped through annular channels surrounding the cylindrical fuel elements placed in these tubes.

The use of heavy water as a moderator was also being studied in the 1970s in Sweden and West Germany (using pressure vessels instead of tubes), in France (carbon dioxide cooling through pressure tubes), in the United Kingdom (boiling light water in pressure tubes), and in a Euratom project (organic cooling). Other varieties may be considered, including the use of enriched uranium or plutonium or uranium-233 as fuel.

Light-water reactors. During the early 1970s both pressurized- and boiling-water reactors were being built in about equal numbers in many countries throughout the world. In spite of the disadvantages of requiring enriched fuel, of making inefficient use of the fuel, and of operating at inefficient temperatures, widespread recognition was being given to their economic and environmental advantages over fossil-fuelled plants. In competition with the United States, a manufacturing capability for light-water reactors was developing in France, Western Germany, and the Soviet Union.

Fast-breeder reactors. There is widespread interest in other countries in breeder reactors. Prototypes have been built in the United Kingdom, in the Soviet Union, and in France. Second generation plants are in the Soviet Union, with 150,000 kilowatts of electrical output, and in the United Kingdom with 250,000 kilowatts. Others are being planned in France, West Germany, Japan, and Italy.

NUCLEAR-REACTOR SAFETY PROBLEMS

A nuclear reactor is not a bomb and thus does not present a bomb-type hazard; nevertheless, the problem of safety is critical. It is simplified by certain self-controlling features of reactors; *e.g.*, the fact that any increase in temperature in a boiling-water or pressurized-water reactor reduces the reactivity—that is, tends to shut down the reactor, thus providing an inherent stability.

Nuclear excursions. The possibility of a nuclear excursion, involving a substantial release of radioactive materials, nevertheless remains. A loss of coolant is conceivable. The loss would immediately stop the chain reaction, but the heat residue could melt the fuel, releasing a large volume of radioactivity. Coolant systems are designed to make the probability of such an accident extremely remote, and an emergency coolant is provided in the form of automatically sprayed borated water. In the very remote possibility of both coolant and emergency-coolant systems failing, further engineering safeguards must be provided. These include back-up devices for rendering harmless any fission products released and surrounding the entire reactor, in addition to the normal shielding, with a steel shell strong enough to contain any scale of accidental radiation fallout. The reactor plant is always placed in an area remote enough to further minimize danger to the nearest community.

Emergency safeguards

Coolant purification. During normal power-reactor operation, there may be some leakage of slight amounts of fission products and perhaps of radioactive corrosion products into the coolant circulating through the core. All reactors, therefore, provide for the removal of such materials by a coolant-purification system. Also, in pressurized-water reactors using boron as a burnable poison, there may be some tritium production in the water coolant resulting from the absorption of neutrons in boron. Although all of these materials are normally completely contained in the reactor, it is difficult to completely prevent leakage. In some cases some of the fission products, such as the inert gases krypton and xenon, are vented high into the air through stacks after a holdup period that allows the short-lived isotopes to decay. Similarly, the tritium in pressurized-water reactors, chemically identical to water after combining with oxygen, may be released to a small extent in the external cooling water.

REACTORS IN OTHER APPLICATIONS

Besides the reactors built for electric-power production and to propel nuclear submarines, many reactors throughout the world serve other purposes. In every case, some use is made of either heat or radiation or both.

Research, training, and test reactors. Reactors may be designed and used for research and test purposes, in which case neutrons escaping from the core are used directly. The gamma rays released from fission fragments may be used more conveniently after the fission fragments have been removed from the core.

Research reactors are usually small compared to power reactors. Many of them on university campuses are used for training and educational purposes and to provide neutrons for research in many areas in the physical and the biological sciences. Larger research reactors are found in national laboratories, in which the greater expenses associated with operating at the higher power levels needed to obtain higher neutron fluxes are justified.

Test reactors require relatively high neutron fluxes because their primary purpose is to test the effects of neutrons on materials, particularly those being considered for use within a power reactor. Materials sometimes deteriorate after long exposure to high neutron fluxes. In a test reactor it is possible to study these effects without waiting for the actual reactor operation. These tests can be particularly important in regard to the possible use of various types of nuclear fuel for power reactors.

Small power reactors. In addition to the large installations that provide electric power commercially from central stations, small-scale nuclear electric plants have become practical, particularly in places too inaccessible for convenient refuelling with more conventional fuels.

Nuclear ships. Many naval surface vessels, such as the U.S. aircraft carrier "Enterprise," are nuclear fuelled. The demonstration of nuclear propulsion for merchant ships undertaken in the 1960s by the United-States-built "Savannah," powered by a pressurized-water reactor, proved a technical success but an economic failure, because it was not possible to operate the ship without a subsidy. The demonstration, however, is significant, and other nonmilitary applications of nuclear power at sea have been developed, notably the Soviet ice-breaker "Lenin." West Germany has built the merchant ship "Otto Hahn," equipped with a pressurized-water reactor to provide power for propulsion, and Japan was building its first nuclear-powered merchant ship in the early 1970s.

Desalinization. In a number of places in the world freshwater is scarce and expensive to import, but seawater is at hand. Nuclear power may provide the most economical means for operating desalinization plants. Consideration is being given to dual-purpose plants in which part of the heat could be used for electric-power production.

Rocket propulsion. Small nuclear reactors were being developed in the 1970s to provide auxiliary sources of heat and electricity for space flights; more ambitious was research directed at producing a nuclear engine for rocket propulsion. Major research and development activities were aimed at early applications—for example, for space shuttle transportation—and at advanced fission and fusion concepts. (See also ROCKETS AND MISSILE SYSTEMS.)

NUCLEAR FUEL

Most uranium ores contain less than 1 percent of uranium, usually in the form of uranium oxide. After the mining and milling of the ore, the uranium must be concentrated and then converted to uranium hexafluoride if it is to be put through an enrichment process in a gaseous diffusion plant. Having achieved the desired enrichment, it is then converted to the metal or to the oxide or to whatever chemical form is desired and fabricated into the required shape together with the needed cladding. After installation in the reactor, some portion of the uranium will be consumed, leaving fission products, the remaining uranium-238, and the plutonium that has been formed. After the optimum burnup time, the spent fuel is removed and transported to a chemical-processing plant in which the fission products and the plutonium are removed, leaving unused uranium to be converted back to uranyl hexafluoride for re-enrichment and re-use. The fission products in some cases have value as radiation sources and may be regarded as by-products that can be used as nuclear fuel. If not, they must be stored as radioactive wastes. The plutonium is a by-product. A simplified outline of this total fuel cycle is shown in Figure 3.

Fission products as by-products

Drawing by D. Meighan

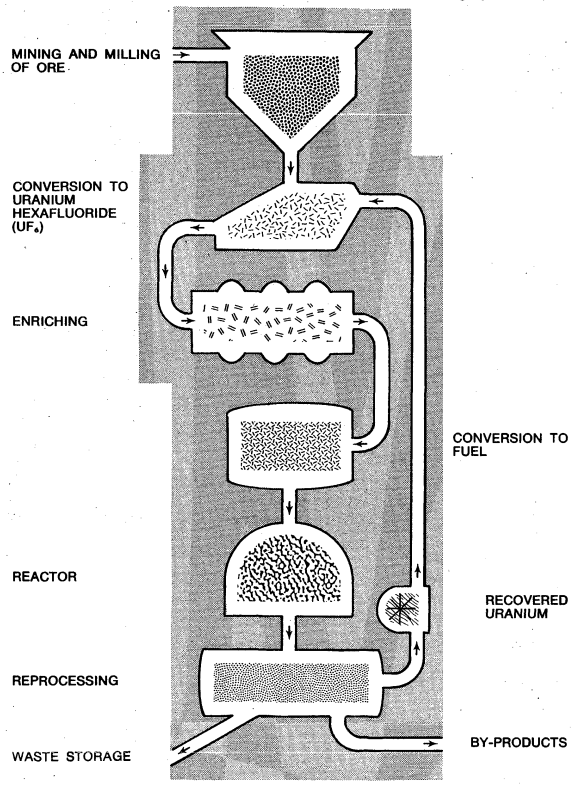


Figure 3: Basic steps in the preparation and processing of fuel for nuclear reactors.

Sources of uranium and thorium. The highest grade uranium ores available to the Western countries, containing from 1 to 4 percent uranium, are found in the Katanga Province of the Republic of the Congo and in Canada. Medium-grade ores in range of 0.1 to 0.5 percent uranium are found in many parts of the world. One of the more significant lower grade ores is found in the South African gold-ore residues. In the United States, extensive phosphate fields, oil shales, and uraniferous lignites contain large amounts of uranium with a concentration of 0.01 percent or less. Uranium is present in seawater in large amounts, but at extremely low concentrations.

Large deposits of thorium ores are found in India, Canada, and Brazil in monazite-bearing sands. Lesser amounts occur in Australia, Madagascar, South Africa, and the United States.

During purification, the uranium oxide is usually converted first to a nitrate and then to uranium dioxide. If the fuel product is to be enriched, it may be converted first to uranium tetrafluoride, commonly referred to as green salt, and then converted further to uranium hexafluoride. The latter is a gas at temperatures above 56.5° C (133.7° F) at atmospheric pressure and may therefore be fed through the gaseous diffusion process.

Monazite ore may contain as much as 1 to 5 percent of thorium dioxide. The extraction and purification of thorium as a nitrate is carried out in a manner analogous to that for uranium. It may then be converted first to thorium tetrafluoride and then reduced to the metal or converted to the oxide.

Enrichment of uranium. Enrichment requires the separation of the isotopes of uranium-238 and uranium-235. They are chemically identical and differ only in that uranium-238 has a slightly greater mass.

The principle of the gaseous diffusion separation process is to allow the uranium in gaseous form (uranium hexafluoride) to diffuse through a porous barrier. On the average, the lighter molecules have a slightly greater speed in order that their kinetic energies may be the same as the heavier molecules. These lighter molecules come in contact with the walls of the container more frequently than the heavier molecules and therefore are more likely to encounter a pore and pass through the barrier. Each barrier has hundreds of millions of pores per square inch with the average pore diameter being about 0.000002 inch (0.00005 millimetre).

The difference in penetration is very small because the masses of the two molecules are so nearly the same. Nevertheless, it is possible to obtain desired enrichments by arranging the barriers in a cascade so that the effect is multiplied over and over again. Thousands of barrier stages are used with high pressure on one side and low pressure on the other. The uranium hexafluoride enters at about the middle of the cascade. As the gas moves through the cascade from high pressure to low pressure in each stage, it becomes more enriched. The gas moving in the other direction contains increasing proportions of uranium-238. A pump is required between each stage; thus the pumping capacity of the plant and the electric-power requirements for the pumps are substantial. Because it has been the major source of enriched uranium since that first developed by the Manhattan Project, gaseous diffusion technology is well developed.

Another process receiving serious consideration and a great deal of experimentation is the use of a high-speed centrifuge. Centrifugal forces cause the heavier atoms to gravitate toward the periphery of a spinning centrifuge, thus producing some enrichment in the gas remaining near the centre. One of the major difficulties is the construction of centrifuges that will themselves withstand the strong centrifugal forces.

Fabrication and preparation of the fuel. For homogeneous reactors, the fuel is in the form of a fluid, and there is no need for fabrication or cladding. It must merely be in the proper chemical form.

Fuel types

For heterogeneous reactors, fuel is used in many different chemical forms and shapes and with different types of cladding. The fuel should have a high thermal conductivity, be resistant to radiation damage, be chemically stable (particularly with respect to the coolant in the event of leakage through the cladding), and it should be easy to fabricate. Uranium has been used in the form of a pure metal, as a constituent of an alloy, in the form of its oxide, its carbide, and in other forms. Each has its advantages and disadvantages with respect to particular reactor types. Uranium metal corrodes in water and changes its shape with irradiation. Uranium dioxide has far better radiation resistance, thus allowing higher burnup, but its thermal conductivity is low. Uranium carbide has high thermal conductivity, favourable mechanical properties, and excellent dimensional stability under irradiation.

Fuel cladding must have a low neutron-absorption cross section if it is used in thermal reactors. It must be mechanically strong (to withstand possible outside pressures or inner pressures caused by the accumulation of gaseous fission products). It must be radiation-resistant and able to withstand attacks by both the fuel and the coolant at the temperatures existing in the reactor. Materials that have been used include aluminum, magnesium, stainless steel, zirconium, graphite, and various carbides. In most cases, some bonding material is used to improve the thermal contact between the fuel and the cladding.

Fuel burnup. The revenue-producing portion of the fuel cycle occurs in the period when the fuel is within the reactor. In general, the longer it can stay inside the reactor, the more the other costs of the cycle are offset. The limits on burnup are determined by changes in the reactivity of the reactor and by the physical condition of the fuel. The accumulation of fission products that absorb neutrons and the reduction in the amount of fissionable material (uranium-235, uranium-233, or plutonium-239), thus diminishing the supply of neutrons, could eventually bring the chain reaction to a stop. The production of plutonium (or uranium-233) would be an offsetting factor helping to maintain reactivity. In most cases, however, radiation damage to the fuel elements would make them unusable before they reached the point at which reactivity is too low. The fuel or cladding may change shape or become corroded or brittle or deteriorate in other ways, so that the heat flow to the coolant is interrupted or fission products may escape into the coolant.

Fuel reprocessing. When the fuel elements are first removed from the reactor, they must be allowed to cool long enough for the radioactivity of the fission products to decay to acceptable levels, a process usually involving their immersion in water for a month or so. The chemical-separation process, even after this cooling-off period, has to be carried out under conditions of intense radioactivity.

After being transported in suitably constructed radiation-resistant containers, the fuel is taken to a chemical-separation plant in which the jacket may first be removed mechanically, or the fuel may be dissolved with its jacketing still in place, depending upon the fuel and the circumstances. Great care is taken to make certain that at no time is there a large enough quantity of fissionable material accumulated to allow a chain reaction to occur. Separation of the uranium and plutonium may be carried out by one of three general methods: (1) aqueous processes with the principal separation being accomplished by solvent extraction techniques; (2) volatility methods depending on the distillation of uranium hexafluoride; and (3) pyrometallurgical (or melt-refining) processes, such as those that provide for removal of impurities as oxides. Other procedures based upon liquid-metal or fused-salt extraction, metal distillation, zone melting, and electrorefining have been studied.

Fuel dejacketing

Radioactive-waste disposal. The disposal of the highly radioactive wastes that remain as liquid residues after removal of the uranium and plutonium and any fission products having practical value represents one of the more difficult problems in the fuel cycle. Some of the radioactive fission products, notably strontium-90 and cesium-137, are long-lived, and it is difficult to store them in such a way that they will not escape into the ground or water and thus become a hazard.

For the short term, large underground tanks have proved satisfactory with mechanisms built into them to carry away the heat caused by the radioactivity; but with the accumulation of ever-increasing amounts of fission products, other methods are being considered. One promising method is to prepare the material for permanent storage in deep salt mines. The material must be reduced to solid form, and the salt mine must be sufficiently deep and meet certain geological criteria that will guarantee no disturbance for thousands of years to come.

Gaseous fission product wastes are also troublesome. Iodine, krypton, and xenon, the most important examples, are liberated when the fuel is dissolved. The last two, being inert, are sometimes released into the atmosphere, but they can also be adsorbed in carbon or silica gel at low temperatures and then stored.

Fuel management. High burnups are generally desirable, and various procedures may be followed that affect not only burnup but the total cost of the fuel cycle. The advantages and disadvantages of various procedures, together with the interactions between various portions of the cycle, are complex, requiring sophisticated computer techniques to find the best procedures.

Instead of loading and unloading the total core as a unit (batch irradiation), it may be desirable to load part of the fuel in a zone toward the outside of the core and then, after a certain period of operation, to use this partial batch to replace a previous partial batch in a central region. In this way, the extra enrichment needed by the fresh fuel to obtain a long burnup is compensated by the fission-product poisons that have accumulated in the old fuel in the central region.

There are some variations in the nuclear-fuel cycle with different reactor types, and in every instance the complex of interactions between its several components must be

considered. In addition to the pattern of moving batches of fuel within the reactor, the enrichment of the fuel may be varied, plutonium may or may not be recycled for use as a fuel, the size of each batch loaded into the reactor may be varied, the power level may be lowered to increase the life of the fuel, and neutron-control poisons may be used in different ways.

BIBLIOGRAPHY. H.D. SMYTH, *A General Account of the Development of Methods of Using Atomic Energy for Military Purposes Under the Auspices of the United States Government, 1940–1945* (1945), the first public disclosure of the technical accomplishments that made it possible to use atomic bombs at the close of World War II (a classic in the history of nuclear technology); RICHARD G. HEWLETT and OSCAR E. ANDERSON, JR., *The New World 1939/1946*, vol. 1 of *A History of the United States Atomic Energy Commission* (1962); and *Atomic Shield 1947/1952*, vol. 2 by RICHARD G. HEWLETT and FRANCIS DUNCAN (1969), two volumes that constitute the early official history of the United States Atomic Energy Commission, containing well-documented, detailed information pertaining not only to wartime developments but also to early efforts to develop a nuclear reactor program and descriptions of the debates and controversies surrounding the early postwar development of nuclear weapons; T. REIS, *L'Énergie nucléaire dans le monde* (1957), a detailed summary of reactor-development interests of countries throughout the world in the mid-1950s; AMERICAN NUCLEAR SOCIETY, *International Conference on the Constructive Uses of Atomic Energy: Proceedings, November 10–15, 1968* (1969), a report on the international status of nuclear power technology during the late 1960s; SAMUEL GLASSTONE and ALEXANDER SESONSKE, *Nuclear Reactor Engineering* (1967), an excellent text on nuclear reactors that is technically sound and understandable to students of engineering. See also the *Annual Report to the Congress of the Atomic Energy Commission.*

(P.N.P.)

Nuclear Weapons

Nuclear weapons derive their enormous explosive force from either the fission or fusion of atomic nuclei. Their significance may best be appreciated by the coining of the words kiloton (1,000 tons) and megaton (1,000,000 tons) of TNT equivalent to describe their blast effect. For example, the first nuclear fission bomb, the one dropped on Hiroshima, Japan, in 1945, released energy equalling 20,000 tons (20 kilotons) of chemical explosive from less than 100 pounds of uranium. Fusion, or thermonuclear, bombs (H-bombs) have given yields ranging up to 60 megatons. The first nuclear weapons were bombs delivered by aircraft; warheads for ballistic missiles have come to be by far the most important nuclear weapons. Smaller tactical, or battlefield, nuclear weapons have been developed; they include artillery projectiles, howitzer shells, and demolition munitions.

The basic principle of nuclear fission weapons involves the assembly of a sufficient amount of the uranium isotope uranium-235 or of the plutonium isotope plutonium-239 to "go supercritical"—that is, for neutrons (which cause fission and are in turn released during fission) to be produced at a faster rate than they can escape from the assembly. There are two ways in which a subcritical assembly of fissionable material can be rendered supercritical and made to explode. The subcritical assembly may consist of two parts, each of which is too small to have a positive multiplication rate; the two parts can be shot together by a gun-type device. Alternatively, a spherical subcritical assembly surrounded by a shell of chemical high explosive may be compressed into a supercritical one by igniting the explosive.

The basic principle of the thermonuclear, or fusion, weapon is to produce ignition conditions in a thermonuclear fuel, such as deuterium, an isotope of hydrogen with double the weight of normal hydrogen, or lithium deuteride. The sun may be considered a thermonuclear device; its main fuel is deuterium, which it consumes in its core at temperatures of 10,000,000° to 20,000,000° C (18,000,000° to 36,000,000° F). To achieve comparable temperatures in a weapon, a fission device is used.

There are no insurmountable scientific or technological difficulties in designing and building either a fission or a fusion weapon. In the early 1970s, five nations, the United States, the Soviet Union, the United Kingdom, France, and China, had tested and stockpiled both types of weapons, and many other nations were capable of producing them, although the Non-Proliferation Treaty of 1970 made the process more difficult for signatory nations.

For a detailed explanation of the basic processes, see NUCLEAR FISSION; NUCLEAR FUSION.

FISSION WEAPONS

Following the discovery of artificial radioactivity in the 1930s, the Italian physicist Enrico Fermi performed a series of experiments, exposing many elements to low-velocity neutrons. He obtained more than 400 new radioactive substances, but nearly all were isotopes of the exposed elements—that is, the same elements chemically but with a different atomic mass. When he exposed thorium and uranium, however, he encountered a striking exception: chemically different radioactive products resulted, indicating that new elements had been formed, rather than merely isotopes of the original elements. Fermi concluded that he had produced elements beyond uranium (element 92), the last element in the periodic table; he called them transuranic elements and named two of them ausenium (element 93) and hesperium (element 94). During the autumn of 1938, however, at the very moment when Fermi was receiving the Nobel Prize for his work, Otto Hahn and Fritz Strassmann of Germany were discovering that one of the "new" elements was actually barium (element 56).

The mystery of uranium fission

The Danish scientist Niels Bohr visited the United States in January 1939 and carried with him the explanation given by the Austrian refugee scientists Lise Meitner and her nephew Otto Frisch. A new process explained Hahn's surprising data. Low-velocity neutrons caused the uranium nucleus to fission, or break apart into two smaller pieces the combined atomic numbers of which equalled that of uranium, for example, the two elements barium and krypton. Much energy was released in the process. This news set off experiments at many laboratories. Bohr worked with John Wheeler at Princeton; they postulated that the uranium isotope uranium-235 was the one undergoing fission; the other isotope, uranium-238, merely absorbed the neutrons. It was discovered that neutrons were produced during the fission process; on the average, each fissioning atom produced more than two neutrons. If the proper amount of material were assembled, these free neutrons might create a chain reaction. Under special conditions, a very fast chain reaction might produce a very large release of energy; in short, a weapon of fantastic power might be feasible.

Development of the fission bomb. The possibility that such a weapon might first be developed by Nazi Germany alarmed many scientists and was drawn to the attention of Pres. Franklin D. Roosevelt by Albert Einstein, then living in the United States. The President appointed an Advisory Committee on Uranium; it reported that a chain reaction in uranium was possible though unproved. Chain-reaction experiments with carbon and uranium were started in New York at Columbia University, and in March 1940 it was confirmed that the isotope uranium-235 was responsible for low-velocity neutron fission in uranium. The Advisory Committee on Uranium increased its support of the Columbia experiments and arranged for a study of possible methods for separating the uranium-235 isotope from the much more abundant uranium-238. (Normal uranium contains approximately 0.7 percent uranium-235, most of the remainder being uranium-238.) The centrifuge process, in which the heavier isotope is spun to the outside, as in a cream separator, at first seemed the leading candidate, but at Columbia a rival process was proposed, diffusion of gaseous uranium hexafluoride through barriers, or filters; more molecules containing the lighter isotope, uranium-235, would pass through the filter than those containing the heavier isotope, slightly enriching the mixture on the far side. It was calculated that a sequence of several thousand stages would be needed to enrich the mixture to 90 percent uranium-235; the total barrier area would be many acres.

During the same summer of 1940, Edwin McMillan and

The discovery of neptunium and plutonium

Philip Abelson of the University of California at Berkeley discovered element 93, named neptunium; they inferred that this element would decay into element 94. The Bohr and Wheeler fission theory suggested that one of the isotopes, mass number 239, of this new element might also fission under low-velocity neutron bombardment. The cyclotron at the University of California in Berkeley was put to work to make enough element 94 for experiments; by mid-1941, element 94 had been firmly identified and named plutonium, and its fission characteristics had been established. Low-velocity neutrons did cause it to undergo fission, and at a rate much higher than that of uranium-235. The Berkeley group, under Ernest Lawrence, was also considering producing large quantities of uranium-235 by turning one of their cyclotrons into a super mass spectrograph. A mass spectrograph employs a magnetic field to bend a current of uranium ions; the heavier ions (uranium-238) bend at a larger radius than the lighter ions (uranium-235), allowing the two separated currents to be collected in separate receivers.

In the spring of 1941 a review committee reported that a nuclear explosive probably could not be available before 1945; a chain reaction in natural uranium was probably 18 months off; it would take at least an additional year to produce enough plutonium for a bomb; it would take three to five years to separate enough uranium-235. Further, it was held that all of these estimates were optimistic. In late June 1941 President Roosevelt established the Office of Scientific Research and Development under the direction of the United States scientist Vannevar Bush; a British committee simultaneously recommended pushing their uranium-235 bomb project at maximum speed.

In the fall of 1941 the Columbia chain-reaction experiment with natural uranium and carbon yielded negative results. A review committee concluded that boron impurities might be poisoning it by absorbing neutrons. It was decided to transfer all such work to the University of Chicago and repeat the experiment there with high-purity carbon. At Berkeley, the cyclotron, converted into a mass spectrograph (now called a calutron), was exceeding expectations in separating uranium-235, and it was enlarged to a ten-calutron system with a total ion current of one ampere, capable of producing a tenth of an ounce of uranium-235 per day.

The move toward actual production

The U.S. entry into World War II in December 1941 was decisive in providing funds for a massive research and production effort for obtaining fissionable materials. In May 1942 the momentous decision was made to proceed simultaneously on all promising production methods. Bush decided that the army should be brought into the production plant construction activities. The Corps of Engineers opened an office in New York City and named it the Manhattan Engineer District Office. After considerable argument over priorities, a workable arrangement was achieved with the formation of a three-man policy board chaired by Bush and the appointment of Gen. Leslie Groves head of the Manhattan Engineer District. Groves arranged contracts for a gaseous diffusion separation plant, a plutonium production facility, and a calutron pilot plant, which might be expanded later. The day before the success of Fermi's chain-reaction experiment on December 2, 1942, Groves signed the construction contract for the production reactors. Many problems were still unsolved: the diffusion barrier had not yet been demonstrated as practical. Berkeley had been successful with its empirically designed calutron, but the Oak Ridge pilot plant contractors were understandably uneasy about the rough specifications available for the massive separation of uranium-235, which was designated Y-12 (for yttrium-12) effort. Plutonium chemistry was almost unknown; in fact, it was not known whether plutonium gave off neutrons during fission, or how many.

Meantime, as part of the reorganization in June 1942, J. Robert Oppenheimer became the director of Project Y, the group that was to design the actual weapon. The effort was spread over several locations; in the fall, Groves and Oppenheimer chose a laboratory site in New Mexico, the former Los Alamos Ranch School some 90 miles (140 kilometres) north of Albuquerque. By July two essential and encouraging pieces of experimental data had been obtained—plutonium did give off neutrons in fission, more than uranium-235; and the neutrons were emitted in a short time compared to that needed to bring the weapon materials into a supercritical assembly. The theorists contributed one discouraging note: their estimate of the critical mass for uranium-235 had risen over threefold, to something between 50 and 100 pounds.

The emphasis during the summer and fall of 1943 was on the gun method of assembly: the projectile, a subcritical piece of uranium-235 (or plutonium-239), would be placed in a gun barrel and fired into the target, another subcritical piece of uranium-235. After the mass was joined (now supercritical), a neutron source would be used to start the chain reaction. A problem developed with the plutonium gun. In manufacturing plutonium-239 from uranium-238 in a reactor, some of the plutonium-239 absorbs a neutron and becomes plutonium-240. This material undergoes spontaneous fission, producing neutrons. Some neutrons will always be present in a plutonium assembly and cause it to begin multiplying as soon as it goes critical, before it reaches supercriticality; it will then explode prematurely and produce comparatively little energy. The gun designers tried to beat this problem by achieving higher projectile speeds, but they lost out in the end to a better idea—the implosion method.

The implosion technique

In April 1943 a Project Y physicist, Seth Neddermeyer, proposed to assemble a supercritical mass from many directions, instead of just two as in the gun. In particular, a number of shaped charges placed on the surface of a sphere would fire many subcritical pieces into one common ball at the centre of the sphere. John von Neumann, a U.S. mathematician who had had experience in shaped-charge, armour-piercing work, supported the implosion method enthusiastically and pointed out that the greater speed of assembly might solve the plutonium-240 problem. U.S. physicist Edward Teller suggested that the converging material might also become compressed, offering the possibility that less material would be needed. By late 1943 the implosion method was being given an increasingly higher priority; by July 1944 it had become clear that the plutonium gun could not be built. The only way to use plutonium in a weapon was by the implosion method.

By 1944 the Manhattan Project was spending money at a rate of over $1,000,000,000 per year. The situation was likened to a nightmarish horse race; no one could say which of the horses (the calutron plant, the diffusion plant, or the plutonium reactors) was likely to win or whether any of them would even finish the race. In July 1944 the first Y-12 calutrons had been running for three months but were operating at less than 50 percent efficiency; the main problem was in recovering the large amounts of material that reached neither the uranium-235 nor uranium-238 boxes and, thus, had to be rerun through the system. The diffusion plant was far from completion, the production of satisfactory barriers remaining the major problem. The first reactor at Hanford, Washington, had been turned on in September, but it had promptly turned itself off. Solving this problem, which proved to be caused by absorption of neutrons by one of the fission products, took several months. These delays meant almost certainly that the war in Europe would be over before the weapon could be ready. The ultimate target was slowly changing from Germany to Japan.

In April 1945, two weeks after he had become president, Harry Truman was briefed on the status of the project: the uranium-235 gun design had been frozen, but sufficient uranium-235 would not be accumulated until around August 1. Enough plutonium-239 would be available for an implosion assembly to be tested in early July; a second would be ready in August. Several B-29s had been modified to carry the weapons; support construction was under way at Tinian, in the Mariana Islands, 1,500 miles south of Japan.

The test of the plutonium weapon was named "Trinity"; it was fired before dawn on July 16, 1945. The theorists' predictions of the energy release ranged from the equiva-

lent of 1,000 tons of TNT to an optimistic 5,000 tons. Instead the test produced an energy, or yield, equivalent to 20,000 tons of TNT.

The bombing of Hiroshima

A single B-29 bomber flew over Hiroshima, Japan, on August 6, 1945, at 8:15 in the morning, local time. The untested uranium-235 gun assembly was air burst 2,000 feet (600 metres) above the city to eliminate local fallout. Hiroshima had been chosen because it was a port of embarkation, a convoy assembly site, and the site of an army headquarters, as well as a manufacturing centre. Two-thirds of the city area was destroyed. The second weapon, a duplicate of the plutonium-239 implosion assembly tested in "Trinity," was scheduled to be dropped on Kokura on August 11; a third was being prepared in the U.S. for possible use in late August or early September.

To avoid bad weather, the schedule was moved up two days to August 9. The B-29 spent 45 minutes over Kokura without sighting its aim point; it then proceeded to the secondary target of Nagasaki. About 50 percent of that city's area was destroyed. The next day, Radio Tokyo broadcast Japan's tentative acceptance of the Potsdam terms of surrender.

Fission weapon programs outside the U.S. Scientists in several countries performed experiments in connection with nuclear reactors and fission weapons during World War II, but no country other than the U.S. carried its projects as far as separating uranium-235 or manufacturing plutonium-239. In Paris, Jean-Frédéric Joliot-Curie and two colleagues had measured the number of neutrons emitted during fission and concluded that a chain reaction was possible. Another French researcher introduced the concept of critical mass and calculated that a sphere of several tons of pure uranium might produce a self-sustaining reaction. During the fall of France in June 1940, Joliot-Curie's two colleagues reached England with the world's entire supply of heavy water (400 pounds) and continued their chain-reaction experiments at Cambridge.

The British weapon project started informally, as in the U.S., among university physicists. In April 1940 a short paper of Professors Frisch and Rudolf Peierls, expanding on the idea of critical mass, estimated that a super weapon could be built using several pounds of pure uranium-235 and that this amount of material might be obtainable from a chain of diffusion tubes. A group known as the MAUD Committee was set up in the Ministry of Aircraft Production. This committee's feasibility report, in July 1941, caused the U.S. to propose a joint production effort, but the British felt they were far ahead and were cool toward the proposal. A year later, when the British wished to merge the two projects, the U.S. was moving rapidly and had lost interest in a joint effort. A number of British scientists did join the weapon design effort at Los Alamos. By 1943 the separate British project was abandoned.

Soviet nuclear beginnings

In the Soviet Union the news of Joliot-Curie's experiments had excited great interest. Igor Kurchatov and others at Leningrad started studies on nuclear reactors and, by late 1940, felt ready to approach the government for the necessary funds to build one. The German attack in July 1941 focussed most physicists' attention on more immediate problems. One of Kurchatov's students observed the gradual disappearance of articles on fission physics from the U.S. journals; in May 1942 he wrote an impassioned plea to several outstanding U.S.S.R. physicists and the State Defense Committee that the chain reaction experiments should continue, that nuclear weapons should not be written off. A government review committee, including leading Soviet scientists, recommended that the work be continued at a high priority with Kurchatov in charge.

A Uranium Institute with a few dozen physicists was established in Moscow; studies were started on nuclear reactors, on the separation of the uranium isotopes by barrier diffusion, and on the ballistic problems of gun assembly. Uranium metal and carbon of sufficient purity, however, did not become available until after World War II. The Soviet government then expanded the effort and set a national goal of highest priority to produce a nuclear weapon. The first Soviet chain-reaction experiment went critical on Christmas 1946; the first nuclear weapon was tested on August 29, 1949. The time between the two events, two and a half years, was the same as that between Fermi's successful experiment in Chicago and the "Trinity" test at Alamogordo.

In Germany the War Office also received an enthusiastic letter, in April 1939, advocating nuclear weapon development. By the time the war had started, Germany had a special office for the military application of nuclear fission; chain-reaction experiments with uranium and carbon were being planned, and ways of separating the uranium isotopes were under study. Some measurements on carbon, later shown to be in error, led the physicist Werner Heisenberg to recommend that heavy water be used, instead, for the moderator. This dependence on scarce heavy water was a major reason why the German experiments never reached a successful conclusion. The isotope separation studies were oriented toward low enrichments (about 1 percent uranium-235) for the chain reaction experiments; they never got past the laboratory apparatus stage, and several times these prototypes were destroyed in bombing attacks. As for the fission weapon itself, it was a rather distant goal, and practically nothing but "back-of-the-envelope" studies were done on it.

German difficulties

The French and Chinese nuclear weapon projects were postwar efforts. Some French scientists had worked in exile during the war in Britain and Canada and returned to work for the Commissariat à l'Energie Atomique. This French effort, however, was aimed at energy production; not until 1954 was a section for military applications formed. The first French plutonium production reactor went on line in 1956; the first nuclear weapon was tested in 1960.

The Chinese program started in 1958 with a slightly enriched uranium reactor, which was part of a technical aid program provided by the Soviet Union. When that aid program was stopped a year or two later, foreign intelligence experts estimated it would be the early 1970s before China could produce enough plutonium to test a weapon. China's first test bomb (1964), however, contained no plutonium at all—it used uranium greatly enriched in uranium-235. China continued to move rapidly. Its second test bomb was dropped from an airplane; its fourth was delivered by a missile. Its sixth test was a multimegaton thermonuclear device. While there were only five known nuclear powers as the 1970s began, there were rumours that several other countries (such as India and Israel) had completed all the necessary research, development, and production of materials and lacked only the final credential, a proof test, to be considered nuclear powers.

THERMONUCLEAR WEAPONS AND THEIR DEPLOYMENT

The United States project. The U.S. Super project started from a conversation in early 1942 between Fermi and Teller. Fermi suggested that the explosion of a fission weapon could be used to start something similar to the reactions in the Sun. Teller undertook to analyze the thermonuclear processes in some detail and presented his work several months later to a summer study group of theoretical physicists in Berkeley. They concluded that a weapon based on the thermonuclear fusion of deuterium was indeed possible. Oppenheimer made a special trip east to arrange for experimental studies, and as the news leaked to physicists at Berkeley and Chicago, several volunteered for the new laboratory at Los Alamos out of interest in the Super project.

The project's first and most obvious requirement, however, was a working fission weapon. It was suggested that the addition of tritium (an isotope of hydrogen with three times the normal weight) could lower the required ignition temperature, because the reaction rate of a deuterium–tritium mixture was many times larger than the rate for pure deuterium. Tritium, then, seemed to be a necessity. Groves made arrangements for a pilot plant operation to manufacture tritium from lithium in the experimental reactor at Oak Ridge (tritium is produced when a neutron is captured by the lithium-6 isotope of lithium). The net effect of a review in 1944, however, was to delay further work on the Super until after the war.

Even after the war, the Super continued to be a low-priority program, behind testing and stockpiling fission weapons. The test of the Soviet Union's first fission weapon on August 29, 1949, provoked a vigorous debate among United States scientists and political and military leaders. The Atomic Energy Commission was divided; the majority recommended against the development of the Super at that time. The General Advisory Committee had concluded that the theoretical studies were incomplete, that success might require large amounts of tritium, but that an imaginative and concerted attack on the problem had a better than even chance of producing a fusion weapon within five years. They observed that such a superweapon could not be confined to military targets and hoped that its development could be avoided.

The Joint Chiefs of Staff also made a study but did not recommend a high-priority program to build a Super, although it did urge a speedy determination of feasibility and some prudent plans for long lead-time production items. The Joint Committee on Atomic Energy of the U.S. Congress was strongly in favour of the Super. On January 31, 1950, President Truman made his decision that work on the Super would proceed.

The production facilities for thermonuclear materials were pushed steadily. By the end of 1950, a pilot plant for the production of heavy water was completed, and a Savannah River (South Carolina) site was selected for building up to six dual reactors capable of producing either tritium or plutonium. The Super theoretical studies went less smoothly.

The H-bomb and the computer

The development of thermonuclear weapons, which because their energy comes from "burning" the deuterium isotope of hydrogen came to be known as hydrogen bombs, or H-bombs, was intimately associated with electronic computers. The calculations needed in early 1950, however, were larger than any existing computer could handle. Drastically simplified calculations, suitable for available computers, showed that Super appeared to be marginal, that much more tritium would be required than previously believed.

In the spring of 1951, a new approach was mapped, and a Pacific test of thermonuclear principles proved successful. By fall Los Alamos had gone on a six-day work week to ensure that the first device test would be completed before the end of 1952. On the morning of November 1, 1952, the test was successful, obliterating a small island and leaving a wave-washed crater more than a mile in diameter.

Control agreements. When a nation becomes a nuclear power, it often issues an extensive statement about its weapon programs; most nuclear powers have been more reticent about their first thermonuclear test. The Soviet Union tested its first thermonuclear device on August 12, 1953; the United Kingdom tested its first on May 15, 1957. The larger yields of the thermonuclear explosions and the increasing pace of nuclear testing made the world aware of the fallout problem—that is, the danger to life of the radioactive products given off during a nuclear explosion, the total of which was building up at an increasing rate.

During 1958 the Soviet Union accepted a U.S. offer to convene a conference of technical experts on the problem of supervising a ban on testing. A temporary moratorium on testing went into effect on October 31, 1958, while the talks continued. It lasted almost three years, until the Soviet Union resumed atmospheric testing on September 1, 1961, conducting over 30 tests, including one with a yield of 60 megatons, before the U.S. resumed testing at Christmas Island some eight months later. The talks themselves had continued, punctuated by the atmospheric testing episode and the Cuban missile crisis of October 1962. In June 1963 Pres. John F. Kennedy announced that three-power talks (the United Kingdom, the U.S., and the U.S.S.R.) would resume in Moscow. The result was the Moscow Treaty, or the partial Nuclear Test-Ban Treaty, which became effective on October 10, 1963. It bans nuclear weapon tests in the atmosphere, in outer space, and under water. The treaty does not specifically prohibit testing in any other environment, provided the explosion does not cause radioactive debris to be present outside the territorial borders of the tester. Any signatory power can withdraw by giving three months notice. France and the People's Republic of China did not sign, China testing its first fission weapon about a year later. Five tests later, on June 17, 1967, it tested its first thermonuclear weapon. France continued its atmospheric tests in the South Pacific and achieved its first thermonuclear weapon on August 24, 1968.

In the spring of 1966, the U.S. Senate passed a resolution in support of negotiations for a non-proliferation treaty to prevent the further spread of nuclear weapons. The negotiations were formally carried out under the auspices of an Eighteen-Nation Disarmament Committee (ENDC). In January 1968, the U.S. and the U.S.S.R. submitted a draft treaty to the ENDC. The treaty, formally called a Treaty on the Non-Proliferation of Nuclear Weapons, was opened for signature on July 1, 1968, at London, Moscow, and Washington. It went into effect on March 5, 1970, when it had been ratified by the three depository nations (the U.K., U.S., U.S.S.R.) and 40 other nations. The nuclear weapon states agreed not to assist any nonnuclear weapon states to obtain or produce nuclear explosives in any way; they would, however, make available the potential benefits of peaceful applications of nuclear explosions for such industrial or civil-engineering uses as the stimulation of gas fields or the construction of harbours and canals. The nonnuclear weapon states agreed not to produce nuclear explosives and to allow inspection of their nuclear reactors and nuclear-materials production facilities to insure that no source or special fissionable material was being diverted for possible weapons use. These safeguard procedures were being worked out in the early 1970s. Several countries, including China, France, India, Israel, Japan, South Africa, and Spain, had not yet ratified the treaty.

The Non-Proliferation Treaty

Deployment of nuclear weapons. The main role played by nuclear weapons has been a strategic one—to support a country's doctrine of deterrence. In the U.S. this doctrine has ranged from so-called massive retaliation in the 1950s to mutual deterrence in the 1970s. Most new nuclear powers have chosen the long-range heavy bomber as their first nuclear weapon carrier. Early carriers were simply modified versions of existing aircraft; later planes were specifically designed for the long ranges, high speeds, and other characteristics thought necessary.

Several developments, such as reliable guidance systems, were needed before the intercontinental ballistic missile (ICBM) became militarily attractive. As missile design improved, with solid propellants replacing the operationally awkward liquid propellants and with guidance that reduced the warhead yield requirement, great progress was also made in producing higher thermonuclear yields per pound of warhead weight, as well as in reducing the overall size. The fleet ballistic missile system of the U.S. Navy, started in 1956, saw the first Polaris submarine, the nuclear-powered "George Washington," become operational on November 15, 1960. The sixth Polaris submarine conducted an operational test, including the detonation of the warhead, near Christmas Island in May 1962. The Minuteman missile system of the U.S. Air Force, started in 1958, saw its first squadron become operational in early 1963. In the early 1970s new versions of the Polaris and Minuteman were deployed, capable of carrying multiple, independently targetable reentry vehicles (MIRV).

The Soviet Union deployed its nuclear weapons in a pattern similar to that of the U.S., except for the early postwar years when it built a large, intermediate-range ballistic missile (IRBM) force, targeted for western Europe. For many years, the U.S.S.R. ICBM force lagged behind the U.S.; strategic parity with the U.S. for ICBM's was attained in 1969. At that time, the U.S. estimated the Soviet ICBM threat at about 300 SS-9s, capable of carrying 25 megatons each, and 700 SS-11s, a Minuteman-like missile using a storable liquid propellant. In the early 1970s, the Soviet Union also had a number of Polaris-like submarines.

The United Kingdom and France have selected a sub-

marine-launched ballistic-missile system as their main nuclear deterrent force. The U.K. deployed four, and France planned to deploy five, Polaris-like submarines in the mid-1970s. In addition, France was emplacing 18 IRBM's in silos in Haute-Provence.

In the early 1970s, China had made no statements on the method it would use in the deployment of its nuclear stockpile. Its fourth nuclear weapon test was delivered by a missile. U.S. estimates were that the Chinese would be capable of producing up to 25 ICBM's by 1975.

Just as there have been efforts to control the testing and proliferation of nuclear weapons, there have been efforts to control nuclear weapon delivery systems. In January 1969, the Soviet Union announced that it stood ready to begin talks with the U.S. on the limitation and reduction of strategic nuclear delivery systems, including defensive systems. Strategic arms limitation talks (SALT) opened in Helsinki, Finland, on November 17, 1969, and continued into the early 1970s.

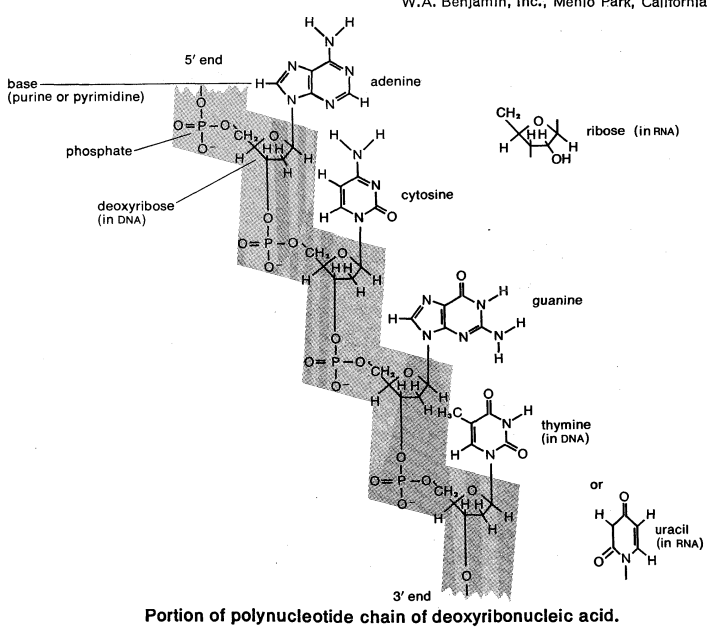
BIBLIOGRAPHY. R.G. HEWLETT and O.E. ANDERSON, *A History of the United States Atomic Energy Commission*, vol. 1, *The New World 1936/1946* (1962); R.G. HEWLETT and F. DUNCAN, *ibid.*, vol. 2, *Atomic Shield 1947/1952* (1969); M.M. GOWING, *Britain and Atomic Energy, 1939–1945*, vol. 1, *An Official History of the U.K. Atomic Energy Project* (1964); D.J.C. IRVING, *The German Atomic Bomb* (1967), a history of the German Atomic Energy Project (1939–46), based on 400 German documents captured by the U.S. Alsos mission; I.N. GOLOVIN, *I.V. Kurchatov: A Socialist-Realist Biography of the Soviet Nuclear Scientist* (1968; orig. pub. in Russian, 1967), a biography of the U.S.S.R. physicist who headed the Soviet Union's fission weapon project.

(W.J.F.)

Nucleic Acid

Nucleic acids are of interest because they provide the genetic material of the cell and, by directing the process of protein synthesis, determine its inherited characteristics (see GENE). Nucleic acids are naturally occurring complex phosphorus compounds, acidic in character, and capable of being broken down chemically to yield phosphoric acid, sugars, and a mixture of organic bases (purines and pyrimidines). About 1868 nuclei isolated from pus cells were found to contain an unusual phosphorus compound, which was named nuclein. In later years, complex phosphorus-containing acid materials were also isolated from a wide variety of cells; these appeared to be chemically similar to nuclein and came to be called nucleic acids.

There are two classes of nucleic acids: ribonucleic acids (RNA) and deoxyribonucleic acids (DNA). Both RNA and DNA are macromolecules (giant molecules) that can be distinguished from each other by their base and sugar contents. DNA, a major constituent of chromosomes in the nuclei of all cells, is also found in other cellular components (*e.g.*, mitochondria, chloroplasts). RNA is present in both the nucleus and cytoplasm of many cells. The bulk of cytoplasmic RNA is associated with ribosomes, small particles composed of RNA and proteins that are the site of protein synthesis.



Portion of polynucleotide chain of deoxyribonucleic acid.

General features of nucleic acids. Nucleic acids are polynucleotides, long chain compounds consisting of repeating structural units called nucleotides. Each nucleotide contains a pentose (or five-carbon) sugar, a purine or pyrimidine base, and a phosphate residue. The pentose sugar is ribose in RNA and 2-deoxyribose in DNA (see CARBOHYDRATE); the major purine and pyrimidine bases are adenine, guanine, cytosine, and thymine (or uracil in RNA). For further information about the chemical properties of purine and pyrimidine bases and nucleotides, see NUCLEOTIDES.

Enzymatic breakdown of nucleic acids results in the release of nucleotide units in which a phosphate residue is attached to the deoxyribose sugar at one of two possible positions on the sugar molecule. These internucleotide linkages are known as phosphodiester bonds. The internucleotide linkage is the same in RNA.

It was the fact that adenine and thymine are present in approximately equal amounts in DNA, as are guanine and cytosine, together with information from X-ray crystallography of DNA that led Nobel Prize winners J.D. Watson and F.H.C. Crick to postulate that the DNA molecule consists of two chains or strands of polynucleotides coiled around each other to form a double helix, the bases of one strand being paired with complementary bases of the other by hydrogen bonds: adenine paired with thymine and cytosine with guanine. RNA has unequal proportions of the four bases and, in addition, contains unusual bases (*e.g.*, pseudouridine, various methylated purines).

Nucleic acids can be separated from other cellular constituents by treating a tissue with cold acid. Most of the contents of the tissue go into solution, but both RNA and DNA are insoluble in cold acid and can be removed from the mixture. RNA and DNA, in turn, can be separated by treatment with alkali; in this case, RNA is broken down by the alkali into nucleotide units, but DNA remains unchanged. If acid is added to this mixture, DNA precipitates (comes out of solution), and the RNA nucleotide units remain in solution.

The DNA found in cell nuclei (as chromosomes) and in other cell components usually is bound to proteins called histones. The protein bound to the DNA in sperm cells is known as protamine. Bacterial DNA is not associated with protein. DNA–protein complexes from animal tissues, known as deoxyribonucleoproteins, are molecules of very high molecular weight (from 10,000,000 to more than 100,000,000 times the weight of a hydrogen atom). Approximately 25 to 50 percent of the dry weight of this complex is nucleic acid.

Characteristics of DNA. *Physical properties.* Many of the physical properties of DNA depend on the methods used to purify it. If, for example, fragmentation of DNA molecules by either mechanical means or enzymes is avoided, DNA preparations of very high molecular weight (*e.g.*, 120,000,000), corresponding to a chromosome or to the entire nucleic acid complement of a virus or a bacterium, can be obtained. Solutions of such preparations are viscous, and almost any manipulation (*e.g.*, stirring) can break the molecules into fragments of lower molecular weight.

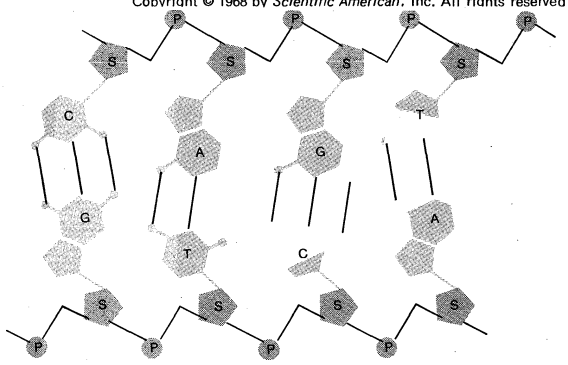
Because of their purine and pyrimidine base content, DNA preparations absorb ultraviolet light. If a solution of DNA is heated to a critical temperature, it absorbs more ultraviolet light and becomes less viscous because the orderly helical structure of the molecules breaks down. DNA molecules become irregularly coiled structures during this process, which is called thermal denaturation of DNA. The melting temperature at which thermal denaturation takes place is dependent on the base composition of the DNA molecule. DNA molecules with a high content of guanine and cytosine, for example, are more stable at high tem-

peratures than those high in adenine and thymine because the hydrogen bonding between guanine and cytosine is stronger. At temperatures slightly above the melting temperature, the two strands of DNA separate. If the DNA solution is cooled slowly, sometimes recombination of the strands and partial restoration of the double helix occurs.

If a concentrated salt solution is spun in a centrifuge at high speed for a long period of time, an equilibrium is eventually attained in which the concentration, and, therefore, the density, of the salt solution increases gradually from the top of the tube to the bottom. If the salt solution also contains DNA, these molecules migrate to the level in the tube in which the density of the salt solution equals their own. This is an important method for determining with great precision the density of nucleic acids and for separating DNA molecules of different densities. The density of DNA molecules with a high proportion of guanine and cytosine is slightly greater than that of molecules with a lower proportion of these bases. Denaturation also causes a slight increase in density. Although DNA usually is double-stranded, there are single-stranded forms (for example, the small bacterial virus designated ϕX 174), which do not show the denaturation phenomena of the double-stranded molecules.

Importance of hydrogen bonds

Structure. The long polynucleotide chains that comprise DNA molecules would form flexible threadlike molecules, instead of coils, were it not for cross-links (hydrogen bonds) between bases of each chain. Measurements indicate that the space between DNA chains agrees with values calculated for hydrogen bond linkage of a purine base to a pyrimidine base (the space is too small for two purines and too large for two pyrimidines). Three cross-links occur between a guanine residue (purine base) and its complementary base, cytosine (a pyrimidine), in two complementary chains; on the other hand, two cross-links occur between adenine (purine base) and its complementary base, thymine (a pyrimidine). Each purine base is linked to its complementary pyrimidine base in the opposite chain by either two or three hydrogen bonds.



DNA structure.
(C) Cytosine. (T) Thymine. (A) Adenine. (G) Guanine.
(P) Phosphate. (S) Deoxyribose (sugar).

The sugar molecules (deoxyribose in DNA) are linked by phosphodiester bonds to form the backbone of each DNA chain. The bases, with their large hydrophobic (water-hating) surfaces, are stacked together on the inside of the molecule like a pile of coins. These two complementary chains coil around each other to form a double helix. In other words, DNA resembles a spiral staircase in which the bases are the steps and the sugar phosphate residues are the bannisters. The DNA molecule derives stability not only from hydrogen bonding between base pairs in complementary chains but also from hydrophobic "base-stacking forces" between neighbouring bases on the same chain. The DNA of mammalian chromosomes and DNA-containing viruses generally consist of long unbranched double helical threads of the type described above. The DNA of many viruses (including ϕX 174) and of mitochondria, however, are circular, and, in some cases, there is evidence that such cyclic double helices are twisted into a supercoiled form. A single break in either chain of the helix releases the molecule, and an untwisted structure results.

Base sequence. The biological function of DNA molecules, whether single- or double-stranded, cyclic or unbranched, is to provide a genetic message encoded in a sequence of purine and pyrimidine bases. Although there are ways to determine the sequences of the units in some macromolecules like proteins (see PROTEIN), it is extremely difficult to do so for DNA molecules, which generally are of much higher molecular weight. In addition, nucleic acids are composed of fewer kinds of units (*i.e.*, the purine and pyrimidine nucleotides) than are found in proteins (*i.e.*, about 20 amino acids). It is also more difficult to obtain a single molecule of pure DNA than it is to purify proteins. Progress has, however, been made in the base sequence determination of the relatively small DNA of the ϕX 174 virus, which contains about 5,500 nucleotide units; thus far it has been possible to obtain only statistical evidence about the order in which these nucleotides are arranged. The usual experimental approach takes advantage of the fact that the purine bases of DNA are removed by acid, while the pyrimidine bases retain their original positions in the molecule. Pyrimidine nucleotides have been shown to occur frequently in clusters.

Laboratory synthesis of DNA. Experiments with living cells have tended to confirm the hypothesis that DNA is synthesized by a process in which the two strands of an existing molecule separate while new complementary strands are synthesized using the old strands as templates, or patterns. For a detailed discussion of the biosynthesis of DNA, see GENE. The product of DNA synthesis in the cell is the formation of two molecules, each containing one strand from the original DNA and a newly synthesized complementary strand.

Enzymes purified from mammalian and bacterial cells catalyze the synthesis of DNA in the test tube. The reaction takes place only in the presence of the nucleoside triphosphates of the four bases (adenine, cytosine, guanine, and thymine) and preformed DNA, which serves as a template. That the product has the same nucleotide sequence as the template has been demonstrated by the technique of "nearest neighbour frequency analysis." One of the four nucleotides contains radioactive phosphorus. During DNA synthesis these radioactive phosphorus residues form internucleotide links with nucleotides next to them in the newly forming chain. The newly synthesized DNA is degraded with enzymes that split the radioactive phosphorus from the molecule to which it was originally attached. The result is that the radioactive phosphorus becomes part of the structure of the nucleotide nearest the nucleotide originally labelled. In this way it is possible to determine the frequency with which particular pairs of nucleotides are neighbours in polynucleotide chains. Such nearest neighbour frequencies vary markedly from one DNA source to another and are specific for each kind of DNA. Generally, the DNA produced by the DNA-synthesizing enzyme in the test tube cannot be shown to possess the biological activity of the template.

Nearest neighbour frequency analysis

When single-stranded circular DNA of the virus ϕX 174 is used as a template for DNA synthesis, however, a complementary (or $-$) strand is formed along the original (or $+$) strand. The free ends of the new (or $-$) strand can be joined by an enzyme to form a cyclic double-stranded molecule. If the cyclic double-stranded molecules are broken with another enzyme, random single breaks occur in a proportion of the molecules in either strand. The broken molecules can be eliminated leaving a number of intact synthetic cyclic ($-$) strands. When these are tested, they are found to be capable of infecting susceptible cells and of giving rise to a new generation of normal virus particles.

Characteristics of RNA. *Physical properties.* Most DNA molecules share a common genetic function and structural pattern. RNA molecules, however, perform several functions in the cell, and their properties vary correspondingly. Three types of RNA have been studied chemically: ribosomal RNA (rRNA), transfer (soluble or adaptor) RNA (tRNA), and viral RNA. Messenger RNA (mRNA), despite its biological importance, has not yet been the subject of extensive chemical investigation. The RNA's of some viruses (*e.g.*, reovirus, wound tumour virus) have

properties in common with DNA, including a double helical structure and a critical melting temperature. Other viral RNA's (*e.g.*, tobacco mosaic virus) and ribosomal RNA, however, have less sharp melting characteristics and lack a complete double helical structure; instead the polynucleotide chains are folded back on themselves to provide small regions of base pairing and, therefore, hydrogen bonding and some helical sections. The small tRNA molecules are thought to be folded similarly into a cloverleaf structure, with regions of hydrogen bonding between base pairs forming hairpin-like helical sections.

The molecular weights of RNA molecules vary according to their type and source. Ribosomal particles, for example, contain a large fraction of the RNA in a cell; these particles can be split into two unequal subunits. The main component of ribosomal RNA from the larger of the subunits has a molecular weight of about 1,000,000; the RNA from the smaller subunit has a molecular weight of about 800,000. Ribosomes also contain RNA's of lower molecular weight (about 60,000). Transfer RNA is the smallest type of RNA molecule; its molecular weight varies from about 25,000 to 30,000. The molecular weights of messenger RNA molecules vary over a wide range.

Structure and base sequence. Research involving RNA structure has centred around tRNA (the type of RNA to which amino acids are attached during protein biosynthesis), the smallest nucleic acids (70 to 80 nucleotide units). The different tRNA molecules (each of the approximately 20 amino acids has at least one specific tRNA) in the cell can be separated and purified. In addition to their small size, tRNA molecules are distinguished from other forms of nucleic acids by their relatively high content of minor nucleotides (*e.g.*, pseudouridine, methylated bases). The relative ease with which these minor nucleotides can be identified has been useful in sequence determinations. By using enzymes that catalyze different specific reactions for various lengths of time on one type of tRNA, fragments of the polynucleotide are produced and can be separated and analyzed to determine the sequence of nucleotides in the intact molecule.

Laboratory synthesis of RNA. It has been suggested that RNA is synthesized in the living cell by a process in which one of the two strands of a DNA molecule is used as a template (or pattern) for the formation of a complementary strand of RNA. For detailed discussion of the biosynthesis of RNA, see GENE. Enzymes found in microorganisms and in animal tissues can catalyze the synthesis of RNA in the test tube. This synthesis occurs provided that a mixture of the proper nucleoside triphosphates and a small quantity of DNA are present. The composition of the newly synthesized RNA corresponds exactly to that of the DNA template (with uracil replacing thymine), and nearest neighbour sequence analysis indicates a similar correspondence in base sequence. In certain cases, a hybrid double helical molecule, which may form from template DNA and the RNA product, indicates an exact correspondence in structure between primer and product. In animal and bacterial cells infected with RNA viruses, the RNA of the virus acts as a template for the synthesis of more RNA.

BIBLIOGRAPHY. E. CHARGAFF and J.N. DAVIDSON (eds.), *The Nucleic Acids: Chemistry and Biology*, 3 vol. (1955–60), a classical modern text, basic chemistry valid but biological sections out of date; E. CHARGAFF, *Essays on Nucleic Acids* (1963); D. COHEN, *The Biological Role of the Nucleic Acids* (1965), a short general monograph; J.N. DAVIDSON and W.E. COHN (eds.), *Progress in Nucleic Acid Research and Molecular Biology*, 11 vol. (1963–71), series of essays published irregularly; D.W. HUTCHINSON, *Nucleotides and Coenzymes* (1964), an elementary monograph; J.C. KENDREW, *The Thread of Life: An Introduction to Molecular Biology* (1966), a popular book based on a series of TV programs; A. KORNBERG, *Enzymatic Synthesis of DNA* (1962), three classical lectures; G. PARKER, W.A. REYNOLDS, and R. REYNOLDS, *DNA: The Key to Life* (1966), a programmed series of questions and answers; R.M.S. SMELLIE, *A Matter of Life: DNA* (1969), a popular paperback; T.L.V. ULBRICHT, *Introduction to Nucleic Acids and Related Natural Products* (1966), an elementary monograph.

College level textbooks on this subject include: J.N. DAVIDSON, *The Biochemistry of the Nucleic Acids*, 6th ed. (1969); E. HARBERS, G.F. DOMAGK, and W. MULLER, *Introduction to Nucleic Acids: Chemistry, Biochemistry and Functions* (1968; orig. pub. in German, 1964); V.M. INGRAM, *The Biosynthesis of Macromolecules* (1965); A.M. MICHELSON, *The Chemistry of Nucleosides and Nucleotides* (1963); and A.R. PEACOCKE and R.B. DRYSDALE, *The Molecular Basis of Heredity* (1965).

(J.N.D./R.Y.T.)

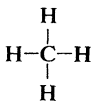
Nucleotides

Nucleotides are organic chemical compounds composed of nitrogen-containing units joined to sugar and phosphate units. They are of importance in biology as the structural groups composing nucleic acids, long chainlike molecules that may contain more than 1,000,000 nucleotides and that make up the fundamental genetic material responsible for storage and replication of hereditary information in living cells. Several nucleotides belong to a class of compounds called coenzymes, substances that act in conjunction with compounds called enzymes that catalyze (speed up) chemical reactions in biological systems; a few nucleotides are starting materials in biological systems from which compounds of other classes are synthesized.

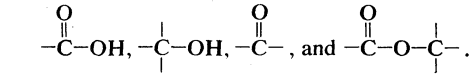
Discovery of the nucleotides

The chemistry of nucleotides has been studied since 1847, when a German chemist, Justus von Liebig, isolated inosinic acid from a meat extract, but nucleic acids were not obtained until more than 20 years later, and their relationship to nucleotides was made clear only after 1900. The first discovery of a nucleotide coenzyme occurred in 1904. The study of nucleotides has been greatly facilitated by the introduction of such analytical techniques as spectroscopy and chromatography about the middle of the 20th century.

General considerations. Understanding of nucleotides is based on knowledge of the structure of organic compounds, which are classified into families that have similar molecular structures and, as a result, comparable properties. All organic compounds have as their basic structural feature chains or rings of carbon atoms linked together by bonds, of a type called covalent, that result from sharing electrons between each pair of atoms. Each carbon atom has the capacity to form four such bonds with other carbon atoms or atoms of other elements. Empirical formulas for compounds are written with the symbol for an element representing a single atom and a subscript indicating the number of such atoms in the molecule; for example, methane has one carbon and four hydrogen atoms: CH_4. Structural formulas indicate the actual disposition of the atoms relative to one another in the molecule, each covalent bond being represented by a line between the relevant symbols:



Most chemical reactions of organic compounds involve only one or two of the numerous atoms and bonds, while the rest of the structure does not change. The unaffected part is called a radical (represented by the symbol R), a molecular fragment that is given a name derived from that of the compound formed by bonding a hydrogen atom to it; *e.g.*, the methyl radical, CH_3, is named from methane, CH_4. There are also tightly knit groups of atoms, called functional groups, that form part of an organic molecule without losing their characteristic properties in chemical reactions, whatever kind of molecule they are attached to. The most important of these are the carboxylic acid group, the hydroxyl group, the carbonyl group, and the ester group, for which formulas are written respectively:



When small organic molecules bond with one another to form chains, the repeating unit is called a monomer, or subunit, and the chain, or macromolecule, a polymer. Several kinds of monomers may also bond together in a re-

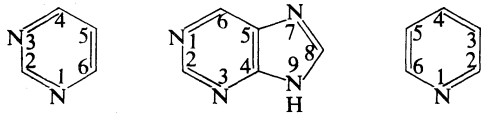
peating pattern to form polymers of varied classifications. Nucleic acids belong to a class of polymers, the subunits of which are nucleotides, themselves composed of three subunits.

The formation of nucleic acids in plant and animal cells is a process of joining a very large number of nucleotides end to end. The reaction is catalyzed by enzymes called polymerases.

Chemistry of the nucleotides and related compounds.

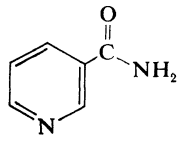
Nitrogen bases

Composition. The nitrogen-containing bases combined with sugar and phosphate units in the naturally occurring nucleotides are derivatives, and so named, of members of three families of heterocyclic compounds: the pyrimidines, purines, and pyridines. The derivatives have formulas in which one or more hydrogen atoms in a pyrimidine, purine, or pyridine structure has been replaced by simple radicals or functional groups, usually amino (NH_2), hydroxyl (OH), or methyl (CH_3). The most abundant of these nitrogen bases are the pyrimidines cytosine, thymine, and uracil; the purines adenine and guanine; and the pyridine nicotinamide. The structures of these compounds are represented by structural formulas in which the symbols for the elements are joined by single lines for single covalent bonds and two lines for double bonds. In a ring or cyclic structure the carbon atoms are at the corners of the ring but are not shown; noncarbon atoms, however, are. Hydrogen atoms attached to ring carbons also are not shown. The same rules are followed with multiple ring structures. Carbon atoms are shown in all chains attached to the rings. Ring positions are numbered for purposes of nomenclature and identification of bond positions. The parent ring systems of the nucleotides are shown as:

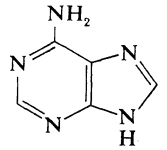


pyrimidine purine pyridine

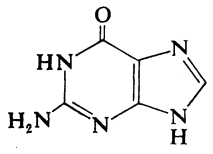
Major nitrogen bases formed from the nucleotides are:



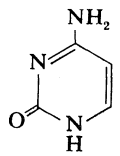
nicotinamide



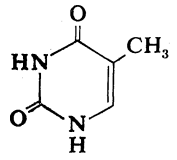
adenine



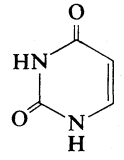
guanine



cytosine



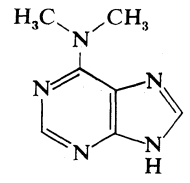
thymine



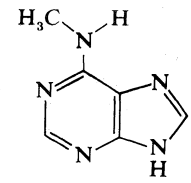
uracil

Certain nitrogen bases occur only in nucleotides present in a few specific nucleic acids; these rarer bases have the structures shown below.

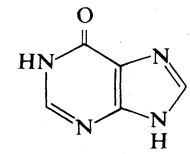
The purines and pyrimidines related to the nucleotides are crystalline, high-melting solids. They do not absorb light in the visible region of the spectrum (hence, are colourless), but their characteristic absorption of ultraviolet light is of great value for analytical determinations.



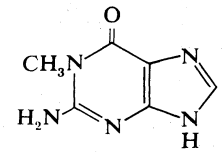
6-dimethylaminopurine



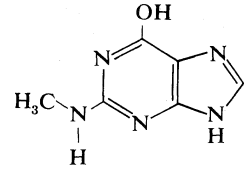
6-methylaminopurine



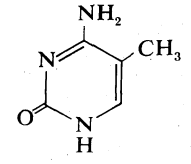
hypoxanthine



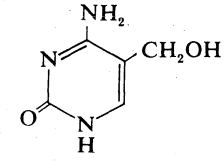
1-methylguanine



2-methylamino-6-hydroxypurine



5-methylcytosine



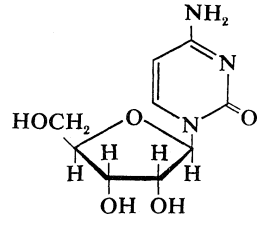
5-hydroxymethylcytosine

The purines and pyrimidines exhibit similar chemical behaviour because the properties of the pyrimidine ring are not much affected by the presence of the second ring of the purines.

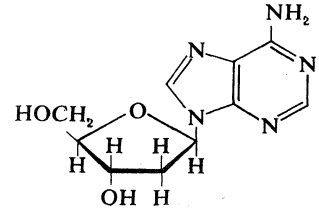
The nitrogen bases can all be prepared from simpler compounds in the laboratory. Pyrimidines are usually synthesized from urea and a second component that furnishes the three carbon atoms that occupy positions designated 4, 5, and 6 in the product. Purines are commonly made from pyrimidines that have amino groups at positions 4 and 5; these amino groups become the nitrogen atoms at positions 7 and 9 of the purine structure.

Several compounds analogous in structure to purine and pyrimidine bases have been synthesized, and their effects on biological systems have been studied. Small variations in the molecular structure can exert profound influences upon processes such as the formation of proteins or of deoxyribonucleic acids (DNA's), often because the modified compound cannot undergo some essential chemical reaction.

Relationship to nucleosides. Nucleosides are compounds in which one of the nitrogen bases is combined with a sugar unit, but the phosphate portion of the nucleotides is absent. The nucleosides are named as derivatives of the bases: for example, cytidine is composed of



cytidine
(a ribonucleoside)

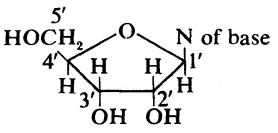


deoxyadenosine
(a deoxyribonucleoside)

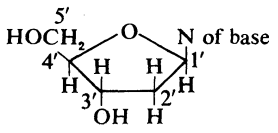
cytosine and the sugar ribose. Nucleosides can be decomposed into their constituent bases and sugars by hydrolysis; that is, a reaction with water in which a chemical bond is broken and the products are the parts of the original molecule—formerly connected by the bond that was hydrolyzed—now combined with the parts of a water molecule (a hydrogen atom and a hydroxyl group). The sugars that are obtained by the process of hydrolysis of nucleosides are usually the five-carbon compounds ribose or 2-deoxyribose.

Sugars

It has been found that the sugar is always present as a five-membered ring for which the structures are shown as:



ribose group

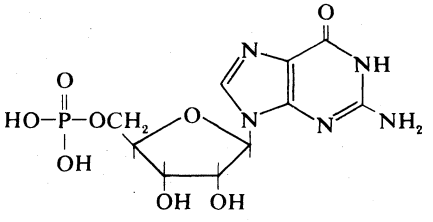


deoxyribose group

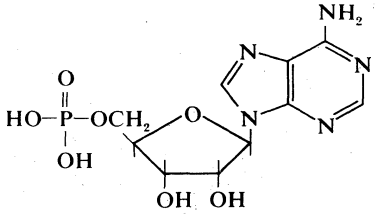
Laboratory preparation of nucleosides requires prior conversion of the sugar into a "protected" form that can combine with a nitrogen base only at position 1′ of the sugar ring. After the base–sugar bond has been formed, the protecting modifications are removed.

Structure of nucleotides. The constitution of nucleotides as three-part molecules was originally established by experiments in which the two chemical bonds joining the three units were broken by hydrolysis. The point at which the sugar is linked to the phosphate group may be determined by studies of the compounds in which these components remain bound together. It has been found that the location of this linkage may change during the chemical treatments used to isolate the nucleoside but that, in natural polynucleotides (nucleic acids), phosphate ester bonds extend from the 3′ position of one nucleotide to the 5′ position of the next.

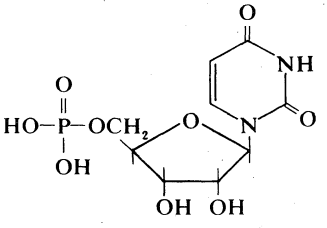
The structures of the nucleotides that make up most of the repeating units in ribonucleic acids (RNA's) are:



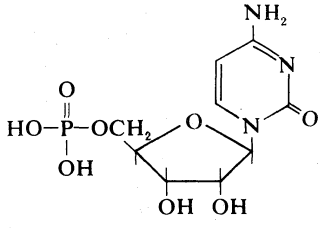
guanylic acid
(guanosine 5′-phosphate, GMP)



adenylic acid
(adenosine 5′-phosphate, AMP)

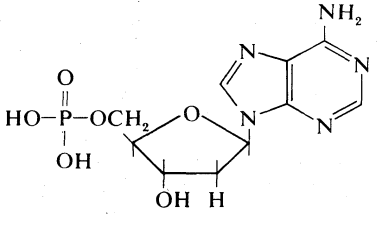


uridylic acid
(uridine 5′-phosphate, UMP)

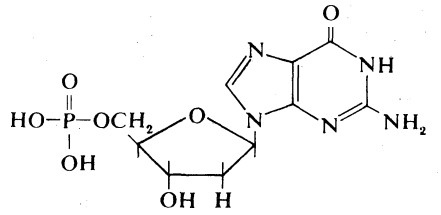


cytidylic acid
(cytidine 5′-phosphate, CMP)

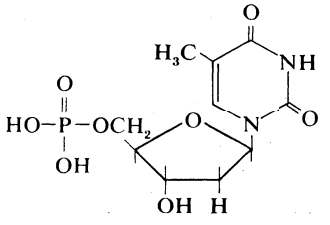
Those most common in DNA's are:



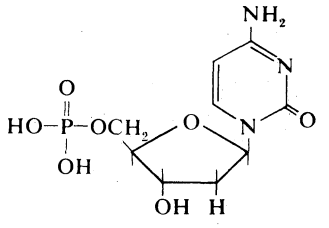
deoxyadenylic acid
(deoxyadenosine 5′-phosphate, dAMP)



deoxyguanylic acid
(deoxyguanosine 5′-phosphate, dGMP)



thymidylic acid
(thymidine 5′-phosphate, dTMP)



deoxycytidylic acid
(deoxycytidine 5′-phosphate, dCMP)

The nucleotides that serve as coenzymes or as the actual physiological building blocks in synthesis of nucleic acids often are esters (when an acid and an alcohol react they form water and an ester) not of phosphoric acid but of condensed relatives of it, pyrophosphoric or triphosphoric acids. They are named as derivatives of the nucleosides.

Laboratory conversion of a nucleoside to a nucleotide is most successful when both the nucleoside and the phosphate unit are employed in modified forms. The nucleoside must be altered so that only one of the two or three hydroxyl groups is available for reaction, while the compound used to introduce the phosphate unit must be capable of reacting with one hydroxyl group but no more. A variety of techniques has been developed for carrying out these difficult syntheses.

Oligonucleotides. An oligonucleotide is a compound in which a small number of nucleotides are joined together by the same kind of chemical linkages as those present in nucleic acids: phosphoric ester groups exist between 3′ and 5′ positions of sugar units. Oligonucleotides may be obtained either by partial hydrolysis of nucleic acids or by combining mononucleotides.

Decomposition of a nucleic acid into oligonucleotides is preferably effected by using enzymes because these substances hydrolyze specific sugar–phosphate bonds without inducing other reactions. Bovine pancreatic ribonuclease, for example, hydrolyzes some of the ester linkages in RNA's and does not act upon DNA's at all. Ribonucleases of differing specificities have been used in determination of the sequence of nucleotide units in RNA's. Such determinations in DNA's, however, are much more difficult because no suitable enzymes have been discovered.

Synthesis of oligonucleotides

Oligonucleotides can be synthesized by treating mononucleotides with reagents that induce the formation of phosphate diesters. Such reactions proceed satisfactorily with deoxyribonucleotides, which possess only one hydroxyl group that can react in this fashion; in ribonucleotides, however, there are two hydroxyl groups of similar reactivity and part of the product contains phosphate links between 2′ and 5′ positions. The desired result can be attained with ribonucleotides only when the 2′ hydroxyl group can be excluded from the reaction, as by prior conversion to an unreactive form from which it can be regenerated after formation of the 3′–5′ phosphate diester bond.

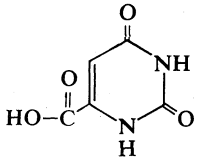
Oligonucleotides with specific base sequences often can be synthesized by enzyme action more easily than by chemical methods. Polymerase enzymes incorporate nucleotide 5′-triphosphates into a copy of a preformed oligonucleotide added to the reaction mixture.

Biological functions of nucleotides. Nucleotides are very widely distributed, apparently being present in all

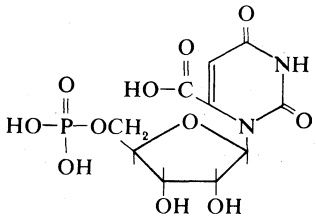
known life forms. Viruses contain DNA's or RNA's, although they are capable of replication only inside the cells of host organisms, utilizing nucleotides available there. Nucleotides have the same functions—those of coenzymes and of starting materials for nucleic acid synthesis—in all plants and animals, from the simplest to the most complex, and they are built up and broken down by processes that vary only in details throughout the same range.

Biological synthesis

In the biological synthesis of the pyrimidine nucleotides, ammonia, carbon dioxide, and the amino acid aspartic acid are utilized in the assembly of the pyrimidine ring of orotic acid, which is then attached to the ribose phosphate unit to produce orotidine 5′-phosphate, from which all the other pyrimidine nucleotides are formed. The structure is:

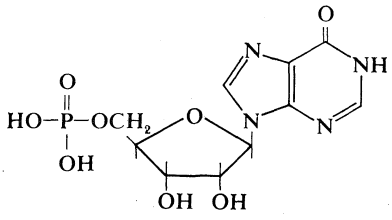


orotic acid



orotidine 5′-phosphate

The precursor (in the sequence of reactions in the organism) of the purine nucleotides is inosinic acid,

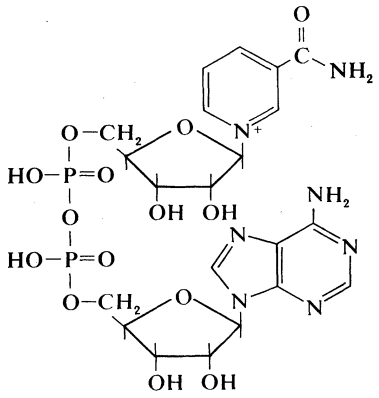


inosinic acid

which is built up from ribose phosphate by stepwise addition of the atoms of the five-membered ring, then those of the six-membered ring.

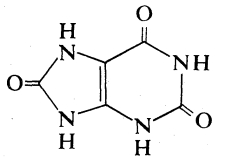
Transformations of orotidine 5′-phosphate or inosinic acid into other nucleotides involve three major classes of reactions: (1) The monophosphates are converted into diphosphates and triphosphates by the action of adenosine 5′-triphosphate, or ATP, which is generated during metabolism of carbohydrates. (2) Ribonucleotides are converted to deoxyribonucleotides by removal, through the action of an enzyme, of an oxygen atom from the ribose ring. (3) Groups such as amino, hydroxyl, and methyl are introduced into the pyrimidine and purine rings by replacement of hydrogen atoms or of other groups.

The important pyridine nucleotide, nicotinic acid ribonucleotide, is utilized in the synthesis of the coenzyme nicotinamide adenine dinucleotide (NAD+), in which the structure is:



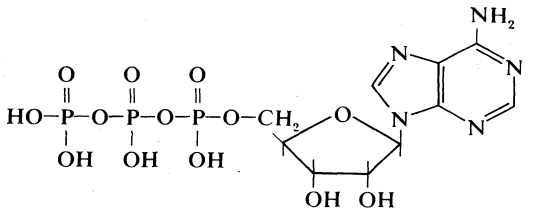
nicotinamide adenine dinucleotide, NAD+

Metabolic breakdown. Conversion of the purine and pyrimidine nucleotides to simpler compounds by reactions occurring in the living organism follows separate routes, after cleavage of the bond between the sugar and the nitrogen base. The pyrimidine bases are decomposed in the body by hydrogenation (addition of hydrogen) and hydrolysis reactions to ammonia, carbon dioxide, and the amino acids β-alanine or β-aminoisobutyric acid. The purines are converted into uric acid, which is excreted by man and other primates. In other species, uric acid is broken down into a series of simpler compounds, including allantoin, allantoic acid, urea, glyoxylic acid, ammonia, and carbon dioxide.



uric acid

Nucleotides as coenzymes. Several nucleotides perform essential roles as coenzymes, the function of which often is to drive physiological reactions more nearly to completion. Adenosine 5′-triphosphate, ATP, is one of the most important coenzymes, participating in syntheses of fats, carbohydrates, proteins, and nucleotides. Its structure is:



adenosine 5′-triphosphate (ATP)

Other nucleotide coenzymes are involved in interconversions among carbohydrates and in the synthesis of membranes and cell walls and of phospholipids, which are essential in the oxidation of organic substances to carbon dioxide.

NAD+ and flavin adenine dinucleotide (FAD) participate in oxidation–reduction reactions, in which hydrogen and oxygen atoms are exchanged between compounds. Coenzyme A transfers groups, such as acetyl (CH_3CO, derived from acetic acid), from one compound to another in the formation of fatty acids, isoprenoids, and steroids. Adenosine 3′, 5′-phosphate (cycloadenylic acid or cyclic AMP) activates the enzyme phosphorylase kinase, one of the substances controlling the interconversion of glucose and glycogen, a storage form of sugar. Cyclic AMP also regulates release of the hormone insulin from the pancreas.

Cyclic AMP

Nucleotides as starting materials. The utilization of nucleotides as starting materials for synthesis of other classes of compounds is exemplified by formation of the amino acid histidine from ATP and of the folic acid coenzymes from the compound GMP.

Formation of nucleic acids from nucleotides by polymerases requires the presence of a metal ion (usually magnesium), a preformed nucleic acid, and the four appropriate nucleoside triphosphates. The nucleic acid is a chain of nucleoside monophosphate units, the remaining two phosphate units of each triphosphate appearing as the inorganic pyrophosphate ion.

BIBLIOGRAPHY. E. CHARGAFF and J.N. DAVIDSON (eds.), *The Nucleic Acids*, 3 vol. (1955–60), a detailed discussion of the chemistry and biochemistry of nucleosides and nucleotides (best suited to the reader with some prior knowledge in the field); M. FLORKIN and E.H. STOTZ (eds.), *Comprehensive Biochemistry*, vol. 8, 14, and 15 (1963–66), good descriptions of the properties of nucleosides, nucleotides, and coenzymes (suitable for further reading by those with little prior knowledge in the field); D.W. HUTCHINSON, *Nucleotides*

and Coenzymes (1964), an elementary account of nucleosides, nucleotides, and coenzymes; A.M. MICHELSON, *The Chemistry of Nucleosides and Nucleotides* (1963), a more advanced, detailed description of nucleosides and nucleotides.

(J.V.K.)

Nucleus, Atomic

An atom, according to one definition, consists of a nucleus surrounded by electrons, the nucleus being a small and relatively heavy object at the centre, about which the electrons gyrate. In this model the atom can be compared to the solar system, the nucleus corresponding to the Sun and the electrons to the revolving planets, except that the solar system is flat with all the planetary orbits in almost the same plane, whereas the electrons in the atom orbit in all planes creating a sphere of effectiveness.

From the point of view of massive substance, the atom is almost all nucleus; the mass of the nucleus is about 4,000 times that of all the electrons together. The nucleus is electrically positively charged, the electron is negatively charged, and in a normal atom all of the electrons together have the same amount of charge as the nucleus so that, since the charges exactly balance, the whole atom is electrically neutral, with a charge of zero. The chemical properties of atoms are determined by the numbers of their electrons of which they may have more or less than normal. The variety of atoms found in nature, explained by the number of different kinds of nuclei, is the basis for the great variety of chemical substance.

Composition of the nucleus

A nucleus consists of protons and neutrons, called nucleons. The proton is positively charged, whereas the neutron has no charge. A neutron and proton have almost equal masses; hence the net charge or atomic number of a nucleus depends on the number of its protons, and the nuclear mass depends on the total number of its nucleons. The nucleons cling together to form a nucleus because of a strong, attractive force known as a nuclear force, one nucleon exerts on another close to it. There also exist opposing repulsive forces arising from the nuclear force (at extremely close range) and from the electrostatic force due to the positive charges on the protons. Nucleons also display much weaker magnetic fields, which can be attractive or repulsive.

Individual nucleons spin on their own axes and circulate vaguely about a common nuclear centre in orbits. These orbits have various energies and are ordered in shells, being arranged according to the laws of quantum mechanics. The shape of a nucleus as a whole is spherical or distorted into an oblate ellipsoid like a tangerine or a prolate ellipsoid like a lemon. It can vibrate, rotate, and change its shape, phenomena that can be identified because of the change in the total energy of the nucleus. Nuclei can change into one another through the process of radioactivity, and certain very heavy nuclei can split into two parts of almost equal size, known as a fission reaction.

Knowledge of atomic nuclei facilitates many practical applications of their properties and enriches other fields of study. One nucleus is said to be an isotope of another if it has the same nuclear charge, and thus the same chemical properties, but a different mass. A rare isotope introduced into one stage of a chemical or other process may be retrieved at another and thus serve as a tracer to indicate the path of the process. Radioactivity, which results from nuclear activity, can be used to treat cancer or, as a poison in the environment, can cause cancer. The energy released by causing interactions among nuclei is, on a weight basis, millions of times greater than that released in chemical reactions, and thus nuclear energy poses unprecedented vistas and threats to mankind. Laboratory studies of nuclei may provide explanations for the processes that release these same energies in stars, thus supporting theories of cosmic evolution.

HISTORICAL BACKGROUND

Since practically all of the mass of an atom is in its nucleus, the early chemical discoveries of the nature of atoms turned out to constitute knowledge about nuclei. In this sense some of the first information about nuclei came from the work of the English chemist John Dalton, who in 1803 proposed an atomic theory of atoms and compounds; of the Italian physicist Amedeo Avogadro, who in 1811 elucidated the number of atoms in a given volume of atoms and thus provided the first basis for determining what turned out to be nuclear masses; and of the Russian chemist Dmitry Ivanovich Mendeleyev, whose periodic table of the elements (1869) classified elements by atomic weights and like properties. After later recognition of the existence of nuclei, it could be inferred that nuclei are composed of "building blocks" of nearly equal mass, now recognized to be protons and neutrons.

Other paths that led more directly to the recognition that an atom has a central nucleus surrounded by electrons began in 1879 with the study of electrical discharges through low-pressure gases. Such discharges excited fluorescence; *i.e.*, produced the delayed emission of light from crystalline materials. While investigating this radiation, the French physicist Henri Becquerel in 1896 discovered radioactivity, a much more energetic and penetrating radiation emitted spontaneously from, in that case, uranium. Study of this new phenomenon in heavy elements was taken up most vigorously by two other French physicists, Pierre and Marie Curie, who in 1898 isolated radium and showed that it has a family of radioactive daughters with individual radiations that could be identified by their penetrations and the lifetimes over which they decay in intensity. Subsequent work traced other radioactive families and showed that different paths of radioactive evolution can lead to atoms with different radiations but the same chemical properties, suggesting isotopes of the same element. Such energetic rays, so different from the light rays ordinarily emitted by atoms, hinted that there must be some special part of the atom capable of emitting them. The radiations consist of alpha particles (later shown by the British physicist Sir Ernest Rutherford to be identical with helium nuclei); beta rays, or streams of electrons; and gamma rays, or electromagnetic radiations, which, like X-rays, resemble ordinary light but are more energetic and penetrating.

Discovery of radioactivity

In 1897 an English physicist J.J. Thomson discovered that what were called cathode rays in the Crookes tube were in reality negatively charged particles generated at the cathode, or negative terminal, coasting past the positive terminal, or anode. After a series of experiments, he concluded that these particles always have the same ratio of electric charge (designated e) to mass (m): e/m. These particles came to be called electrons. Years later, in 1910, Thomson discovered that the so-called canal rays —that is, the rays that appear behind the negative terminal from a discharge in neon gas—consist of ionized atoms (atoms lacking one or more electrons) with different masses, in this case of two different mass numbers, 20 and 22. He thus established the existence of isotopes, atoms of the same chemical element having the same chemical properties but different mass, and confirmed that the nonintegral mass numbers of some elements, in the case of neon 20.2, are due to mixtures of isotopes.

The nuclear age may be said to have started in 1911 when Rutherford definitely established the existence of an extremely small and massive particle within the atom—its nucleus. He did so by letting alpha particles from radium impinge on thin foils of various metals and observing how the particles are deflected by counting the scintillations they produced when they hit a fluorescent screen. From these observations he calculated that the massive cores had diameters about $1/10{,}000$ of the diameter of the whole atom. This, in turn, led a Danish physicist, Niels Bohr, to posit the planetary nature of the atom, with electrons circulating in orbits about a nucleus, and to invent the idea of quantization (according to which only certain discrete orbital energies are possible) of those orbits based on the fundamental quantum constant that Max Planck, the German physicist, had discovered in radiation in 1900 and Einstein had applied in 1905 to the photoelectric effect. New vistas of theoretical physics were opened, and in 1925 the physicists Erwin Schrödinger of Austria and Werner Karl Heisenberg of Germany discovered what is known as wave mechanics or

The Bohr atom

quantum mechanics, which constituted a drastic revision of the laws of classical mechanics as applicable to small systems and provided the basis of modern theories of atomic and nuclear structure in which orbits are more nebulous than in Bohr's theory.

The simplest atom is hydrogen, with only one electron, and its nucleus consists of an elementary particle called the proton, a particle with mass 1,836 times that of the electron and a positive electric charge equal to that of the electron but opposite in sign.

For a time it was thought that the nucleus must consist of protons and electrons, the particles then known, but there were difficulties. In 1932 the British physicist Sir James Chadwick discovered the neutron, an uncharged particle with almost the same mass as the proton (about a tenth of a percent more). The serious study of nuclei as assemblages of protons and neutrons then began.

GENERAL PROPERTIES OF NUCLEI

Mass and electric charge. Nuclear masses are derived by measuring the masses of atoms or molecules and making a correction for the mass of the electrons contained in them. These masses can be measured by means of deflections in combined electric and magnetic fields in an instrument known as a mass spectrograph. It is found that the masses of nuclei differ from one another very nearly by integral or whole multiples of a fundamental unit, the nuclear mass unit, that is very slightly less than the mass of a proton or neutron. Thus each kind of nucleus is assigned a mass number (A) and its mass is said to be A mass units. For example, the mass number of one of oxygen's isotopes is 16, which is the sum of the neutrons and protons in this isotope's nucleus; its mass is 16 mass units.

Atomic and mass numbers

The electric charges of various nuclei also differ from one another by integral multiples of a fundamental unit, in this case, the electron charge. Because atoms are electrically neutral, the positive nuclear charge (or number of protons) is exactly compensated for by the negative charge of the same number of electrons. Each kind of atom, or each element in the periodic table, is assigned what is called its atomic number Z, denoting both the number of electrons of the neutral atom and the electric charge on its nucleus in units of the proton charge. (For example, oxygen has eight protons and, therefore, is of atomic number 8.) Thus each kind of nucleus is characterized by a charge number, Z (for oxygen, 8), and a mass number, A (for oxygen, 16). Since each element also has a chemical name abbreviated into a chemical symbol, it is customary to designate a nucleus by its chemical symbol preceded by a superscript A and a subscript Z. For example, helium, as it occurs in nature, has a charge number $Z = 2$ and a mass number $A = 4$ and is denoted as helium-4 or ${}^{4}_{2}\mathrm{He}$; an oxygen nucleus is represented by ${}^{16}_{8}\mathrm{O}$.

Nuclei with all values of Z from 1 to 92 (except Z = 43 and 61) occur in the earth's crust, and others (up to $Z = 105$) have been made in the laboratory. For each value of Z (charge number), there may be several values of A (mass number), and nuclei with the same Z but different A are known as isotopes of one another, the word isotope implying equal atomic number. Almost all elements have more than one isotope, those with an odd Z usually having two and those with even Z frequently have more, up to a maximum of ten for tin (Z = 50). Altogether, natural elements have about 350 isotopes. Such is the great variety of nuclei that nature has provided.

Stable isotopes

A chemical element as it occurs in nature is in general a mixture of isotopes, and the relative abundance of the various isotopes in the mixture is very nearly constant, no matter where on earth the element may occur. In many elements, one isotope is considerably more abundant than the others, and the chemical atomic weight of the mixture is very nearly the atomic number of that isotope. In ordinary carbon compounds, for example, there are about 90 atoms of the common isotope carbon-12 for every atom of the rare isotope carbon-13, and the resulting chemical atomic weight of the mixture is 12.011. Because they arise from mixtures of isotopes, chemical atomic weights are not always close to integers.

Apart from the light isotope of hydrogen, ${}^{1}_{1}\mathrm{H}$, the mass numbers A of the fairly light nuclei are either equal to or very slightly larger than twice their atomic numbers Z. Beyond the light nuclei this ratio gradually increases until, for the heaviest natural nuclei, such as the heavy isotope of uranium, uranium-238, A is about 2½ times Z. This trend is displayed in Figure 1, in which $A-Z$ is plotted against Z for all the natural isotopes and $A-Z$ is shown to start out about equal to Z and to increase to about 1½ times Z.

From R.R. Roy and B.P. Nigam, *Nuclear Physics* (1967); John Wiley & Sons, Inc.

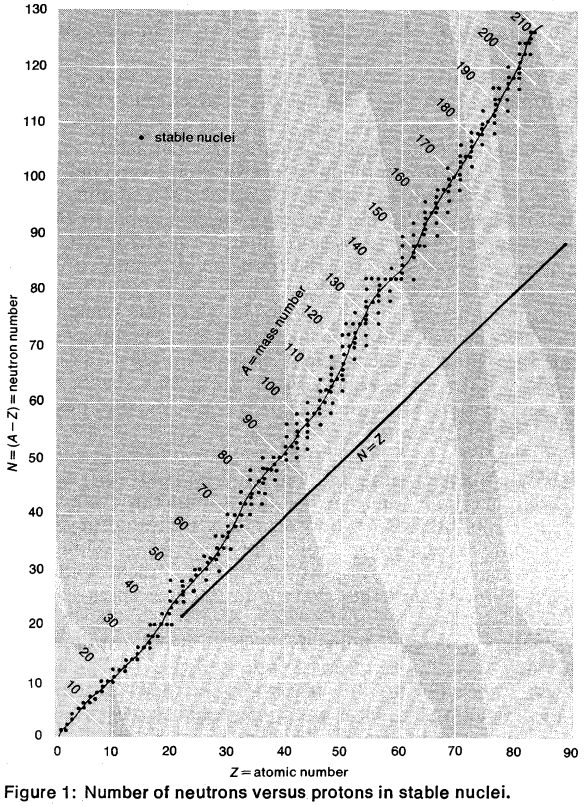


Figure 1: Number of neutrons versus protons in stable nuclei.

The size of nuclei. The radius of a typical atom is about 10^{-8} centimetre—$1/100{,}000{,}000$ of a centimetre. The radii of nuclei are a little less than 10^{-12} centimetre. Thus nuclear radii are roughly 10^{-4}, or $1/10{,}000$, of the radii of atoms. Atomic sizes vary periodically throughout the periodic table of elements, but heavy atoms are in general only slightly larger than moderately light elements. In sharp contrast to this, nuclear sizes increase very regularly with the mass of the nucleus. The volume of a nucleus is, in fact, directly proportional to its mass, so that it is possible to speak of "nuclear matter," the stuff that nuclei are in effect made of, as having a constant density. In this respect, a nucleus is indeed quite like a water droplet.

Nuclei, for the sake of defining an average radius, may be considered approximately spherical, the volume of a nucleus being proportional to the cube of its radius. What is observed by means of scattering of alpha particles (helium nuclei), or more recently of high-energy electrons, from nuclei, is that the nuclear mass is proportional to the radius cubed, and since volume is proportional to the radius cubed, the mass of a nucleus is directly proportional to its volume.

Nuclear angular momentum. One of the intriguing properties of an electron, proton, or neutron, tiny "elementary" particles that they are, is that each seems to be spinning like a top and has an angular momentum (a term describing the persistence of rotation of a body). In quantum mechanics, the unit of angular momentum is Planck's universal constant divided by two pi; *i.e.*, Planck's constant divided by the ratio of the circumference of a circle to its radius, or $h/2\pi$, in which h is Planck's constant (equal to 6.626×10^{-27} erg-sec). Elec-

tron, proton, and neutron spins are equal to one-half this unit of angular momentum and are referred to as spin-½ particles. A spin is said to have a certain orientation, meaning that it can spin in one direction or the other about its axis of rotation.

Nuclear spin

Many nuclei have angular momentum arising from both the spins and the orbital motions of its components. Because of the way it was discovered, this nuclear angular momentum is known as nuclear spin, though it is more complicated than the spinning of a top. Nuclear spin is called I; its value for a particular kind of nucleus is either an integer or half-integer (½, 1, 3/2, 2, etc.) and the magnitude of the spin is I times the unit of angular momentum.

Since electron spins and nuclear spins are both angular momenta arising from some sort of circulation of electrically charged matter, both behave as tiny magnets. The strength of a small magnet is called its magnetic moment. It is largely through the influence of their associated magnetic moments that electron spins and nuclear spins affect atomic spectra; *i.e.*, the energy of spectroscopic lines arising from electron transitions in the atom. When those electrons responsible for the total angular momentum of an atom dip in close to the nucleus, they create an average magnetic field at the nucleus in which the nucleus can assume one or more (in some cases, $2I + 1$, thus indicating the value of the nuclear spin I) different states of slightly different energies, visible in atomic spectroscopy as a hyperfine structure pattern (splitting of lines into further components).

Nuclei with even mass number and even charge number have no nuclear spin, and hence the spin is zero. The very few light nuclei that have a mass number even and a charge number odd (even isotopes of lithium, boron, and nitrogen) have integer 1 or 3 for their spins. Almost all nonvanishing nuclear spins are those of nuclei with odd mass number, and they have odd-integral spins varying all the way from ½ to 13/2.

Nuclear magnetic moments. The magnetic moment of a nucleus—*i.e.*, its strength as a magnet—is measured (with some difficulty) by observing its energy in an external magnetic field. This energy is equal to the projection of a vector, representing the magnetic moment, along the field multiplied by the strength of the magnetic field. A common way of portraying energy states of a molecule, atom, or nucleus is by means of a series of parallel horizontal lines, called energy levels. Thus the energy of a state is represented by the height of its level, and the energy difference of two states is proportional to the separation of their two energy levels. When there is a pattern of equally spaced energy levels formed by the quantization of a nuclear spin in an external magnetic field, what is observed is the energy jump between two adjacent levels.

Quantized magnetic moment

Ever since the late 1920s, measurements of nuclear magnetic moments have been made by measuring the energy spacings in hyperfine structure. The strong magnetic field at the nucleus, due to the electrons, takes the place of an external magnetic field, but this cannot be measured directly and can be determined theoretically only with limited accuracy; so magnetic moments for many nuclei have been determined by this method to an accuracy of only about 10 percent. The more accurate and more modern methods involve nuclear magnetic resonance (see MAGNETIC RESONANCE).

The accuracy of the nuclear resonance method is greater than that of the hyperfine method because the magnetic field being used is accurately known, since it is being supplied by an external magnet, and because the energy jump is induced by radio waves whose frequency can be precisely determined. The energy jumps between successive energy levels made by different orientations of the nuclear spin in the field are accordingly much smaller than those of hyperfine structure.

Nuclear magnetic resonance, the mechanism by which the radio waves cause the energy jump, can be described in a rather pictorial manner. The nuclear spin with its magnetic moment is acted on by the pull of an external magnetic field in the same way that a spinning top is acted on by the pull of gravity, and it precesses around the direction of the magnetic field, describing a cone just as does the axis of a top. In the nuclear case, the frequency of precession is characteristic of the nuclear spin and magnetic moment, and the angle of the cone and corresponding energy are quantized—that is, they can have only certain values depending on the permitted values of the nuclear spin projection. A child's swing also has a natural frequency, and if one pushes it repeatedly at just the right times in "resonance" with its frequency, the swing goes higher and higher. In the case of a top, if one pushes its axle in the direction it is precessing, say to the north when it is at its easternmost position, with just the right frequency, it gradually rises towards a vertical position as it precesses, absorbing energy. In the case of a nuclear spin, the magnetic field of the radio wave pushing on its magnetic moment similarly with the right frequency, in resonance with its precession, can cause it to change its cone angle with the field, jumping from one quantized angle to the next. Energy will be absorbed from the radio wave, indicating the magnetic moment involved.

Measured values of nuclear magnetic moments, sorted out according to values of the associated nuclear spin, are plotted in Figure 2 for nuclei with odd mass number and odd charge number and for nuclei with odd mass number and even charge number. The unit in which nuclear magnetic moments are measured is the magnetic moment or nuclear magneton, equal to the product of the electron charge and Planck's constant, divided by the product of four pi, the mass of the proton, and the velocity of light, or $eh/4\pi Mc$, in which M is the proton mass, e is electron charge, c is the velocity of light, and h is Planck's constant.

Shapes of nuclei and their electric quadrupole moments. A dumbbell with a positive electric charge on one end and a negative charge on the other is a simple illustration of an electric dipole moment. It is turned in

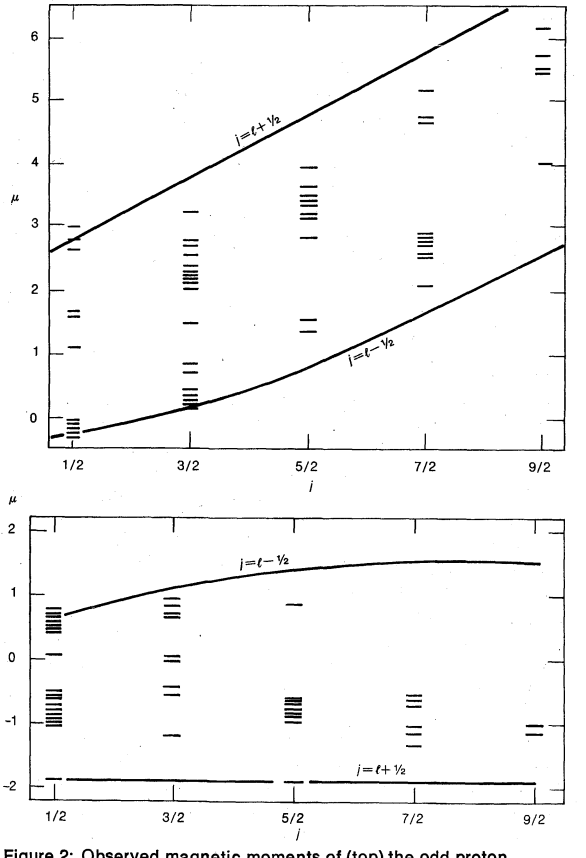


Figure 2: Observed magnetic moments of (top) the odd proton nuclei (having an odd number of protons and an even number of neutrons) and (bottom) the odd neutron nuclei. In each graph, the two lines give the calculated values for a single nucleon, with spin either parallel or antiparallel to the orbital angular momentum.

Example of quadrupole moment

one direction by a uniform electric field just as a compass needle is turned by a magnetic field. An electric quadrupole moment is illustrated by a dumbbell with equal positive charges on both ends. A uniform electric field pulls equally on both ends and does not tend to turn it. An inhomogeneous electric field, one whose strength is not uniform with direction, does pull harder on one end than the other, and thus the energy of an electric quadrupole moment in an inhomogeneous field depends on its orientation. Rotating it by 90° can take the quadrupole moment from a position of minimum energy to a position of maximum energy.

Some nuclei are spherical and have no electric quadrupole moment. Some are elongated like a football and have a positive electric quadrupole moment, like that of the positively charged dumbbell. Some are flattened like a doorknob and have a negative electric quadrupole moment.

An atom with electron orbits confined near one plane has an inhomogeneous electric field at the nucleus, and this makes the energy depend on the orientation of nuclear spin in a different way for the electric quadrupole moment than for the magnetic moment. The quadrupole moment thus modifies the pattern of the atom's hyperfine structure. These effects are small, and it is difficult to do more than estimate how inhomogeneous these fields are; so measurements of electric quadrupole moments deduced from this effect are not accurate.

When a beam of heavy ions (atoms that are charged due to the absence of one or more electrons) such as neon nuclei is allowed to impinge on target nuclei and then scatter, the inhomogeneity of the average field at the target nucleus can be calculated. Such transitory fields from passing ions twist on the quadrupole moment and make a target nucleus rotate, exciting quantized rotations. From the amount of excitation, the quadrupole moment can be determined. The quadrupole moment also helps the nucleus in the excited state radiate and jump back down to the nonrotating state. Measurement of the lifetime of the excited state is thus another way of measuring quadrupole moments.

NUCLEAR STRUCTURE AND FORCES

Nuclear components. Nuclei are made up of nucleons —protons and neutrons—having nearly equal masses. In terms of the mass unit, adjusted to make the mass of the carbon-12 atom exactly 12, and in terms of the electron mass m, their masses are shown in the Table. The mass of

Nucleon Masses

	atomic mass units	electron mass units
neutron mass	1.00867	1838.6
proton mass	1.00728	1836.1
(electron mass)	0.00055	1

the neutron is thus equal to the total mass of a proton plus about 2½ electrons. The neutron, being electrically neutral, has the total electric charge of a proton plus one electron. In respect to mass and charge, then, a neutron is equivalent to a proton and an electron with 1½ electron masses, or 1.5 m left over. According to the equivalence of mass and energy (energy is equal to the product of mass and the velocity of light squared), the mass left over is equivalent to the energy $1.5\ mc^2 = 0.78$ MeV. (The usual nuclear unit of energy is a million electron volts, 1 MeV, or the energy involved in transporting an electron through a potential difference of 1,000,000 volts, and $mc^2 = 0.511$ MeV, in which c is the velocity of light.)

Mass–energy equivalence

Because of this extra mass, a neutron is the unstable form of the nucleon, and when it is alone in free space it lives on the average only about 17 minutes before it becomes a proton and an electron flying energetically apart. At the same time there is emitted an elusive chargeless and nearly massless particle known as a neutrino. The maximum energy of the electron as it flies away is $1.5\ mc^2$, but it practically always has less energy because the neutrino carries away some. This process of changing into a proton and electron is known as beta decay. Neutrons in nuclei may or may not be stable. When neutrons do decay in a large number of similar nuclei, there is emitted a stream of electrons known as beta particles or beta rays—one of the three kinds of radioactivity of heavier nuclei.

Whether or not a neutron undergoes beta decay depends on its environment. If it is bound in a nucleus so tightly that the total energy of the nucleus would be increased by more than $1.5\ mc^2$ if the neutron should become a proton, then the decay is energetically unfavourable and does not take place. This is why neutrons, unstable in free space, can last forever as constituents of stable nuclei. Not only can a neutron emit an electron to become a proton if the energetic situation is favourable but the reverse process can also take place; a proton bound in a nucleus can emit a positive electron, or positron as it is called, and a neutrino to become a neutron bound in the nucleus.

Nucleons and statistics

A nucleon may be either a proton or a neutron and when bound in a nucleus may change from either form to the other depending on the energies involved in the transformation. Nucleons are the fundamental building blocks of nuclei. A nucleus with mass number (A) and charge number (Z) is composed of A nucleons, Z of which are protons and $A-Z$ or N of which are neutrons. When a neutron changes to a proton, emitting an electron, A remains constant and Z changes to $Z+1$. When the reverse process takes place with emission of a positron, A remains constant and Z decreases by one.

In several situations in which quantized aggregates of particles, such as nuclei, tend to settle in low quantized states (low energy levels), their behaviour depends markedly on whether they contain an even number or an odd number of spin-½ particles. One manifestation of their behaviour influences the statistical behaviour of condensing gases. Nuclei with an even number of particles are said to obey Bose–Einstein statistics; those with an odd number obey Fermi–Dirac statistics (see THERMODYNAMICS, PRINCIPLES OF), which keeps them from all moving in the same way or occupying the same low state of motion. By observations it is shown that the nucleus of nitrogen-14, or ^{14}N, obeys Bose–Einstein statistics and should have an even number of particles. Before the discovery of the neutron in 1932, one of the historic difficulties was the presumption that this nucleus must contain 14 protons and 7 electrons to make the positive charge come out equal to 7; but then the nucleus would have an odd number of particles making it inconsistent with Bose–Einstein statistics and demanding Fermi–Dirac statistics. Now one may consider that there are only nucleons present in a nucleus, and in the case of nitrogen-14, 7 protons and 7 neutrons, an even number of particles. One may think of each neutron as virtually a proton plus an electron and a neutrino, because in beta decay in a nucleus, an electron and neutrino are created as a neutron becomes a proton.

Mass defect and binding energy. Particles attracting one another tend to assemble into a state of lowest possible internal energy while giving off external energy, the way water attracted by gravity flows down through a mill to a level of lower energy while supplying energy to the mill. The energy of a nucleus is said to be zero when its nucleons are separated too far apart to interact and the internal energy of a normal nucleus is thus negative, the lowest or most negative energy attainable. The positive measure of this negative energy is called the binding energy, equal to the amount of external energy given off in the binding process. If two nuclei can be combined to make lower internal energy, external work can be done. According to the equivalence of mass and energy, not only the energy but also the mass of the bound nucleons is less than that of the separated nucleons, the missing mass being called the mass defect of the nucleus. Thus the mass of a nucleus is somewhat less than the sum of the masses of its parts. With the unit adjusted to make the mass of the carbon-12 nucleus, $^{12}_{6}C$, exactly equal to 12 nuclear mass units, the nucleons each have a mass greater than one mass unit by almost 1 percent, as already described, and so 12 of them have a total mass greater than

12 units by that percentage and mass defect of $^{12}_{6}C$ is somewhat less than 1 percent of its mass.

A heavy nucleus with many nucleons bound together naturally has more binding energy than a light nucleus with few nucleons. In fact, through most of the range of nuclear masses, the internal energy is almost, but not quite, directly proportional to the number of nucleons, so that the binding energy per nucleon is nearly constant. This is shown in Figure 3, in which the internal energy

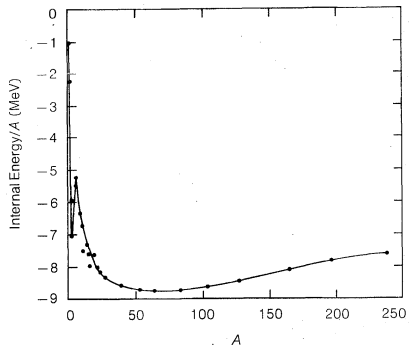


Figure 3: Energy per nucleon by which the nucleons are bound together to make nuclei of various mass numbers A.

per nucleon is plotted against the nucleon number, A, of the nucleus. Aside from minor fluctuations, the curve is seen to be quite flat for the intermediate-mass nuclei from $A = 100$ to 150 and it rises at both ends, quite gradually toward the heavy-nucleus end and more sharply for light nuclei in the region $A < 10$. For most nuclei the binding energy is about $A \times 8$ MeV, or 8 million electron volts per nucleon.

Nuclear energy levels: properties of nuclear states. Many nuclei are spherical, and nucleons in them have orbital motions and spins similar to those of electrons in atoms. The shapes of the orbits are different because there is no central attracting body at the centre. The nucleus has a more sharply defined surface or boundary, and the orbits may more nearly resemble those of a ball rolling or bouncing around the inside of a spherical enclosure, but the angular momentum of the nucleus is quantized just as in atoms. A nucleon also has spin like an electron, and each nucleon has an orbital angular momentum (l) and a spin angular momentum (s).

Nuclear shells

The shell configuration in nuclear physics is a useful concept for the orderly grouping of nucleons—protons and neutrons—that compose a nucleus. Unlike the atom, which has only one kind of particle, the electron, circling about a central field of force, the nucleus has two kinds of particles, the proton and neutron, moving about inside a force field, and therefore the grouping of particles is different in nuclear shells than in atomic shells. In nuclei of increasing atomic number and mass, the orbital and spin angular momenta are important to the way in which nuclear shells can be occupied. According to the Pauli exclusion principle, every particle in any one shell must have its own set of quantum numbers. The quantum number l, representing orbital angular momentum, is useful in assigning nucleons to shells. A shell may contain a maximum number of protons equal to $2(2l + 1)$. It is then called a closed shell of protons, and similarly for a shell of neutrons. The quantity $2(2l + 1)$ is equal to two for quantum number l equal to zero, six for $l = 1$, and so on. In the first shell the orbital angular momentum l is zero and the shell is closed for two protons and two neutrons, as shown by helium-4 (symbolized as $^{4}_{2}He$); and in the second shell it is one, and hence the shell becomes closed for either six protons or six neutrons, or doubly closed for protons and neutrons. The oxygen-16 nucleus, $^{16}_{8}O$, is extremely stable, having a helium-4 nucleus for an inner shell, and thus being doubly closed for both shells.

To illustrate how shells are filled as the number of nucleons increases, hydrogen-2, the deuteron, has one proton and one neutron in the first shell. A second neutron of opposite spin can be added, giving hydrogen-3, the triton. If hydrogen-4 were to exist, the third neutron could not be placed in the same shell as it would then be identical to one of the other neutrons and thus would have to occupy another shell. Helium-3, the mirror nucleus to hydrogen-3, has its first shell occupied by two protons of opposite spin and one neutron. Lithium-5 (which is unstable) contains the helium-4 shell for a core and requires a second shell for its third proton.

A nucleus ordinarily exists in its lowest energy, or ground, state but, under certain experimental conditions, it can be excited to have more energy in a higher state. The energies are discrete or individually distinct, being quantized usually in quite complicated ways. Some of the low states in nuclei having only one nucleon outside of closed shells are relatively simple. Oxygen-17, $^{17}_{8}O$, has one neutron in a third shell, outside of the first two closed shells of oxygen-16, in which the quantum number l can take two values, $l = 0$ and $l = 2$.

When the neutron spin s in a nuclear shell points in approximately the same direction as the orbital angular momentum of the neutron, the total angular momentum, given by the quantum number (j) of the neutron is: $j = l + \frac{1}{2}$. On the other hand, when s points approximately backward along l, $j = l - \frac{1}{2}$. Such a pair of states is known as a doublet with this value of l. The force that gives rise to the energy difference between the two states of the doublet is known as "spin–orbit coupling." If one could poise two gyroscopes at different angles on the point of a needle and if one could stretch a rubber band between the outer ends of their axles, the two gyroscopes would precess around a common direction. Spin–orbit coupling is roughly analogous to the rubber band and makes the spin and orbital angular momenta precess about the direction of the total angular momentum (j) in a similar fashion. The spin–orbit coupling for a neutron or a proton makes the state $j = l + \frac{1}{2}$, lower than the other member of the doublet, and indeed lower usually by several million electron volts (MeV).

Spin–orbit coupling

The energy states of oxygen-16 and oxygen-17, up to an excitation energy of 7 MeV, are shown in Figure 4. The ground state of oxygen-16 has nuclear angular momentum $I = 0$. This is an instance of the general rule that all

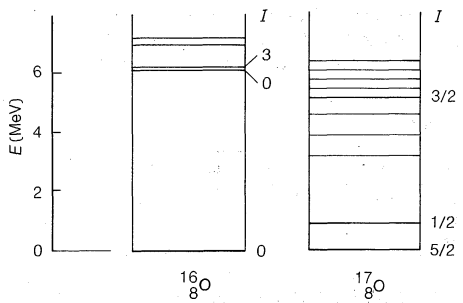


Figure 4: Energy levels of two isotopes of oxygen, contrasting a closed-shell nucleus with its neighbour.

nuclei with even proton number and even neutron number, or even–even nuclei, as they are called, have a nuclear angular momentum equal to zero. Thus the total angular momentum of the additional neutron in oxygen-17 is the total angular momentum of that nucleus. The ground state (zero energy) of $^{17}_{8}O$ is one with $l = 2$ and $I = j = \frac{5}{2}$ or $l + \frac{1}{2}$. The companion state of the doublet, with $l = 2$ and $I = \frac{3}{2}$, has an energy of 5.1 million electron volts (MeV). The other single-neutron state, with $l = 0$ and $I = \frac{1}{2}$ lies between them with the energy of 0.9 million electron volts. These three single-neutron states are indicated by heavy lines in the figure. They appear prominently in experiments known as stripping reactions in which a nucleus may acquire a nucleon that is stripped from a passing deuteron. But most nuclear states are more complicated than this and even the relatively simple nucleus oxygen-17 has many more complicated states. These are in some way related to the excited states of oxygen-16.

It is characteristic of closed-shell nuclei that the ground

state is particularly stable and farther separated from the first excited state than in other nuclei. Whereas the ground state is spherical, the excited states of oxygen-16 have a collective deformation similar to that encountered in the ground states of many heavier nuclei, deformations that involve lifting either two or four nucleons from the supposedly closed second ($l = 1$) shell into higher states with quantum number $l = 0$ or 2. In some of these states the deformation rotates, and in oxygen-17 these rotations combine with the angular momentum of the odd neutron to make a great variety of excited states. This is typical of the complexities found in excited states of nuclei.

Even the three simple single-nucleon states of oxygen-17 involve a little of this excitation from the closed second shell, as is manifest in the effective charge of a neutron needed to account for the magnitude of its electric quadrupole moment, even though a neutron has no charge. As the extra neutron swims almost freely through a sea of other nucleons, so to speak, it seems to drag some of the other nucleons along with it to a small extent. It retains the angular-momentum properties of a neutron, but in other ways the state differs slightly from that of a pure neutron and in this situation it is sometimes called a quasi-neutron, one kind of quasi-nucleon.

When there are two nucleons (or, more precisely, quasi-nucleons) outside of closed shells, their spin and orbital angular momenta can add up in two characteristic simple ways known as (LS) coupling and (jj) coupling or in some more complicated intermediate manner. (LS) coupling, common in atomic spectra, is encountered in nuclei only in the two lithium isotopes lithium-6 and lithium-7 and their immediate neighbours. It nevertheless provides an example of important symmetries in nuclear states.

Lithium-6 has two nucleons, a proton and a neutron, in the second shell in addition to the first closed-shell helium ($^4_2$He) core. Their two spins (designated $s = \frac{1}{2}$) can add up to make a total spin angular momentum of one or zero (designated $S = 1$ or 0). Similarly, the two vectors representing the two orbital angular momenta ($l = 1$) can add up to make a total orbital angular momentum of zero, one, or two ($L = 0$, 1, or 2), all in the usual angular momentum units. Hence the name (LS) coupling. When the total spin is one ($S = 1$), for one value of L these add up as vectors in such a way that they make three states with total nuclear angular momentum: L minus one, L itself, or L plus one ($I = L - 1$, L, or $L + 1$). They are known as a triplet for this value of the total orbital momentum (L), and if the spin–orbit coupling is weak, they are clustered together with only slightly different energies. With total spin equal to zero ($S = 0$), for one value of L there is only a single state, known as the singlet, with nuclear angular momentum equal to the total orbital angular momentum ($I = L$).

(LS) coupling

These two-nucleon states are described mathematically by wave functions, the nature of which cannot be described here (see MECHANICS, QUANTUM) other than to state that they are quantities whose value depends on where the two nucleons are and in which of two possible quantized directions the spins point (say up or down). If the positions of the two nucleons in space are interchanged—if the first nucleon takes the place of the second, and vice versa—the function may either remain unchanged or just change sign, say from positive to negative, and is accordingly called either symmetric or antisymmetric in space exchange. If the total angular momentum quantum number is even, it is symmetric, and if it is odd, it is antisymmetric in space exchange. The function is also either symmetric or antisymmetric in respect to interchanging which spin vector belongs to which nucleon. If the total spin is one it is symmetric and if it is zero it is antisymmetric in spin exchange.

A rule of the Pauli principle states that if the two nucleons are identical (either both neutrons or both protons) the function must be antisymmetric (that is, the function changes sign) on interchange with respect to both space and spin of the two nucleons.

This limitation does not apply when the two nucleons are a neutron and a proton, as in lithium-6, $^6_3$Li, but it does in the similar helium-6 nucleus, $^6_2$He, having two

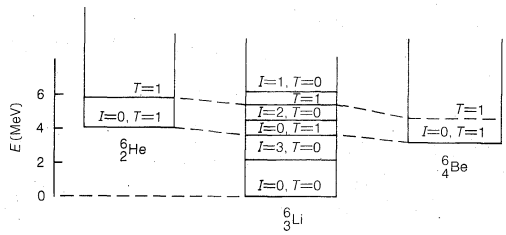


Figure 5: The energy levels of three nuclei with $A = 6$, relative to the ground state of ^{6}Li after subtraction of the calculated electrostatic energy. The isospin $T = 1$ levels exist in all three nuclei.

neutrons outside the closed shell. The low energy levels of these two nuclei are shown side by side in Figure 5 and it is seen that, because of this limitation, helium-6 has fewer states than lithium-6. In fact, instead of three triplets and three singlets it has only two singlets (the notation is as follows: $S = 0$ and $L = 0$ and 2) and one triplet with $S = 1$, $L = 1$, a total of five states in place of twelve. Nevertheless, they are the same states of motion as the corresponding states in lithium-6 and are observed to have very nearly the same energy separations. The other neighbour, beryllium-6, or $^6_4$Be, in which each of the two nucleons with orbital angular momenta equal to one are protons, is similar to helium-6 and is also depicted.

The variety of values of S, and the accompanying symmetries, arise from the existence of two spin vectors, each of which has a quantized up-or-down part (again in angular momentum units), designated as $+ \frac{1}{2}$ for up and $- \frac{1}{2}$ for down, and in addition, a small and indefinite sideways part. In the quantized system these values of $- \frac{1}{2}$ and $+ \frac{1}{2}$ add up in such a way either that $S = 0$, with the up-and-down part equal to zero, or that $S = 1$, with the up-and-down part, called M_S, either $M_S = 1$, 0, or -1. The values $M_S = 1$ or -1 correspond simply to having both spins up or both down, and interchanging them changes nothing, so these states are symmetric. When one nucleon spin is up and one down, they may combine their sideways parts in such a way as to make the total spins equal to one (which is symmetric) or equal to zero (which is antisymmetric).

The mathematical result of all this is that two quantities, each quantized to have two values, are combined into four other quantities, three of which are symmetric and one of which is antisymmetric. A similar situation is encountered in describing the dichotomy of the nucleon, which may be either a proton or a neutron. Quite separately from spin, its neutronness or protonness may be described mathematically as exactly analogous to spin up and spin down. In this analogy isobaric spin, or isospin for short, takes the place of real spin. Two nuclei or two nucleons are isobars of one another if they have the same mass number but different charge number. Isobaric spin up or down indicates which of its isobars a nucleon is, neutron or proton. When there are two nucleons, the two possible states of each combine in a way that is described as having total isospin T (analogous to the total spin, S) either one or zero, so that the combined state is symmetric for isospin equal to one and antisymmetric for isospin equal to zero. Completing the analogy, there are three isospin-equal-one states, either both neutron, both proton, or one of each, and one isospin-zero state with one neutron and one proton. This means that isospin-equal-one states can exist in the three isobars $^6_2$He, $^6_3$Li, and in $^6_4$Be, whereas the $T = 0$ state can exist only in $^6_3$Li. They form an isobaric triplet and an isobaric singlet, respectively.

Isobaric spin

A generalization of the Pauli principle says that a nuclear state must be antisymmetric in simultaneous exchange of any two nucleons in space, spin, and isospin. Since two changes of sign mean no change, this means that it must either be antisymmetric in all three or antisymmetric in one and symmetric in the other two. In lithium-6 there are isospin-one states that are antisymmetric in space and spin (which permits them to exist also in the other isobars) and isospin-zero states that are symmetric in space and spin. Thus in lithium-6 the isospin quantum number

of each state, isospin one or zero, indicates something about the symmetry of the state, and specifically whether or not it has such a symmetry in space and spin as to permit it to exist in the two isobaric nuclei, helium-6 and beryllium-6.

Symmetry may be summarized as requiring that no two nucleons in the same nucleus can behave in exactly the same way. If they are both protons or both neutrons and have the same spin direction, then the function whose magnitude describes where they may be in space is antisymmetric. This means that it changes sign when the position of the two particles is interchanged and becomes small when they are close together. Thus, they are seldom close together to take full advantage of the attractive force between them, and the energy of such a state is not as low as that of a state of the opposite symmetry, such as is permitted for a proton and a neutron.

The (LS) coupling scheme occurs only in a few of these light nuclei, but isospin has significance also for heavier nuclei, even those with many more neutrons than protons. In heavier nuclei, isobaric analogue states occur when a highly excited nucleus will have level spacings and other properties very similar to those of fairly low states in a neighbouring nucleus.

In most nuclei the (jj) coupling scheme prevails because spin–orbit coupling has a stronger influence on the orientation of the spin and orbital angular momentum vectors than have the attractive forces between the nucleons that hold them together in the nucleus. If there are two nucleons outside closed shells, for example, each nucleon has a spin (s) and orbital (l) that add up to its total nucleon angular momentum denoted by j (or $j = l + s$) and the nuclear momenta of the two nucleons (j_1 and j_2) add up to make the total nuclear angular momentum I for the nucleus. The two single-nucleon angular momenta are half-integral and add up to the nuclear angular momentum that may take all integral or whole values for the various states (from $j_1 - j_2$, if j_1 is the larger or they are equal, to $j_1 + j_2$, which makes $2j_2 + 1$ states altogether: for example, if $j_1 = 9/2$ and $j_2 = 5/2$, the sum is 7 and the difference is 2; the values of the states are 7, 6, 5, 4, 3, and 2, or 6 altogether as given by the expression $2j_2 + 1$).

There is an attractive force between two nucleons if they are close together but not if they are far apart on opposite sides of the nucleus. That is, the nucleon–nucleon interaction is short-ranged and attractive. Nucleon orbital angular momentum is something like that of a gyroscope: the mass distribution is near a plane and the orbital angular momentum vector is along an axis at right angles to the plane. Because the spin is usually considerably smaller than the orbital vector, which is the greater part of the total nucleon moment, the nucleon, as it circulates, is almost always near the plane through the centre of the nucleus and at right angles to the total nucleon angular momentum. When the two nucleon angular momenta are equal and they have exactly opposite directions so as to add up to make the total nuclear angular momentum equal to zero, the two nucleons are almost in the same plane and frequently close together to interact strongly. This makes the energy of the state in which $I = 0$ lower than that of all the others. Such nucleons are said to be paired, and the lowering of the energy is called pairing energy. It is responsible for the fact that all even–even nuclei have a total nuclear angular momentum equal to zero ($I = 0$).

Pairing energy

The (jj) coupling shell model. The average potential-energy well in which a nucleon moves in a typical nucleus is flat on the bottom and rises steeply at the sides. A parabola, symmetrical about a vertical axis, its vertex at the bottom, is a very simple function that has such a shape. In it the potential energy is proportional to the square of the distance from the vertex. A particle moving in a parabolic potential is known as a harmonic oscillator and has the special feature that quantized energy levels are equally spaced at intervals that differ by an amount equal to Planck's constant times the frequency with which a classical or nonquantized particle would vibrate in the potential well, namely, $h\nu$. A clock pendulum swinging through a small arc is an example of a one-dimensional harmonic oscillator, for its circular arc is almost the same as the part of a parabola for small angles. The potential energy of a nucleon in a spherical nucleus is approximated by a three-dimensional harmonic oscillator, with potential energy proportional to the square of the distance from the centre of the nucleus. The energy levels form a regularly spaced ladder, but each energy may correspond to several states. This ladder is shown in the left-hand column of Figure 6.

From D.R. Inglis, *Physics Today* (June 1969)

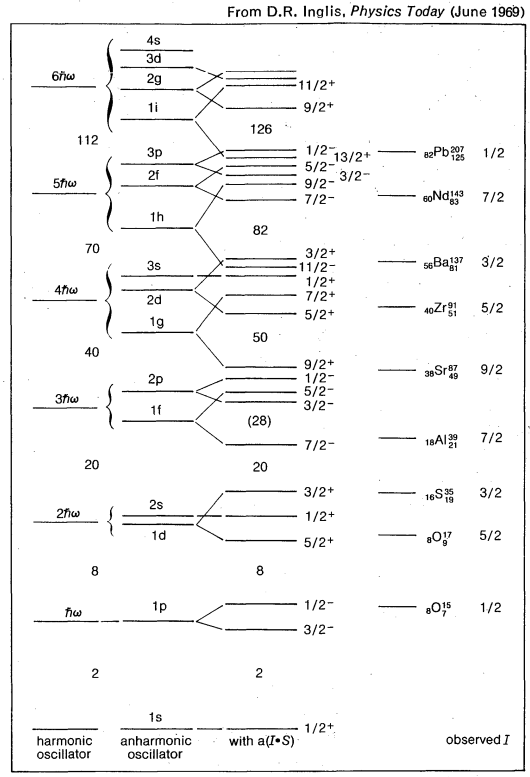


Figure 6: Energy level scheme for a single nucleon with strong spin–orbit coupling. In second column, letters s, p, d, f, g, etc. correspond to orbital angular momentum quantum numbers $l = 0, 1, 2, 3, 4, 5$, etc.

A better approximation would be a potential curve that is flatter on the bottom, rises more steeply, but then levels off at the energy of the separated nucleon. Such a potential tends to make the energy lower for states with higher orbital angular momentum (l) values that tend to concentrate near the edge. The rung of the ladder at two times Planck's constant times the frequency (written as $2h\nu$), for example, has two states, one with $l = 0$ and one with $l = 2$. With the flat-bottomed potential these are split apart, the latter state being lower, as shown in the second column for this and other cases.

Spin–orbit coupling modifies the energy further, splitting each level of orbital angular momentum (l) into two levels (written: $j = l + 1/2$ and $l - 1/2$). These splittings are distinctly larger for the larger l-values, as shown in the third column.

Each of the states in the third column can be filled with neutrons and again with the same number of protons. In the second column, the states are bunched in groups according to the oscillator levels of the first column, a large value being at the bottom of each group. In the top half of the figure, the spin–orbit splittings are so large that the higher total angular momentum (j) falls down into the next lower group, but still the levels fall into groups with large spaces between. The nuclei with levels just filled up to these spaces are especially stable, and a nucleon added to it has smaller than average binding energy. The number of neutrons, say, required to fill up through the last highest level below each of these spaces is known as a magic number, and these numbers, indicated in the spaces in the third column, are 2, 8, 20, 50, 82, and 126.

Magic numbers

Because of the pairing energy, which is especially

effective between like nucleons, an even number of protons, for example, has a total angular momentum zero and if there is just one neutron in addition to closed shells or one neutron short of closing the last shell, the total nuclear angular momentum is that due to that one particle or nucleon vacancy, called a hole. This is borne out perfectly by the observed ground-state angular momentum I, as shown in the last two columns of the figure. In many cases single-particle excited states corresponding to the nearby particle or hole levels in the diagram are observed. Such observations are testimony to the fact that the roundness of closed shells keeps these nuclei nearly spherical. For the ground states and simplest excited states of nuclei, such as these in the immediate vicinity of the magic-number closed shells, the (jj) coupling shell model works extremely well (with the exception of atomic masses equal to 6 and 7).

The general trend (but not the exact values) of the magnetic moments plotted in Figure 2 can be explained in terms of the (jj) coupling shell model with the magnetic moment ascribed to the single odd nucleon. The unit of these magnetic moments is the nuclear magneton, the magnetic moments of a proton circulating in a nuclear orbit with $l = 1$. It is equal to the product of the proton charge and Planck's constant divided by 4π times the product of the proton mass and the velocity of light, or $eh/4\pi Mc$, in which e is charge, h is Planck's constant, M is the mass, and c is the velocity of light. The spin of a proton is more richly endowed with magnetic moment than is the orbit, the proton spin magnetic moment being 2.79 nuclear magneton units, almost as great as that of an orbit with orbital angular momentum equal to three ($l = 3$), although the spin angular momentum is only one-half ($S = \frac{1}{2}$). When the spin points almost forward (along l to make $j = l + \frac{1}{2}$), the magnetic moment associated with spin angular momentum is large, as indicated by the upper solid line of Figure 2A. When the spin points almost backward (to make $j = l - \frac{1}{2}$), the spin magnetic moment tends to cancel out much of the orbital magnetic moment and the total is smaller, as indicated by the lower solid line in Figure 2A. In the case of a single neutron that carries no net charge around with it in its orbit, the orbit contributes no magnetic moment and the spin magnetic moment is negative, — 1.91 nuclear magnetons; so the total moment is positive when the spin points backward and negative when it is forward along l, as indicated by the upper and lower lines of Figure 2B. It is a remarkable sign of some simplicity in nuclei that the observed magnetic moments in Figure 2 fall into two groups, not on theoretical lines but between them in such a way that each group seems to belong to the $j = l \pm \frac{1}{2}$ of the nearest theoretical line. These data include both simple and more complex nuclei. For the simple nuclei with one nucleon or hole aside from closed shells, the discrepancy between theory and observation seems to verify that one is dealing with quasi-nucleons rather than with pure nucleons. For the other nuclei, the simple trend suggests that they are not as complex as one might expect and that the magnetic moment is contributed preponderantly by one quasi-nucleon.

The collectively deformed model. When two nucleons outside closed shells pair together and have their orbits almost in the same plane, they have little effect on the closed shells. When several pairs all attract each other and populate one region, they help each other exert a pull on the closed shells and deform them, partially exciting nucleons from the closed shells to higher states in the process. Such a cooperative or collective deformation of the nucleus may make it resemble either an oblate ellipsoid, flattened like a doorknob, or a prolate ellipsoid, elongated like a watermelon. In either case the nucleus, with a definite nonspherical shape, can rotate like a football tumbling end over end. A round nucleus does not rotate as a whole, just as the water in a bucket does not immediately rotate when the bucket is spun.

Rotational energy

The deformation of nuclei was first recognized by the ladders of rotational energies, similar to those already known in molecules, to which their quantized rotations give rise. The angular momentum of the rotation is an integer, usually expressed in angular momentum units, and the energy of rotation is proportional to one-half the product of the angular momentum and itself plus one, or $\frac{1}{2} I (I + 1)$. For even–even nuclei deformed in an ellipsoidal shape, this quantum number is an even integer 0, 2, 4, etc., which means that when these values are substituted into the foregoing expressions, the rotational energies are proportional to 0, 3, 10, 21, as shown by the characteristic pattern in Figure 7.

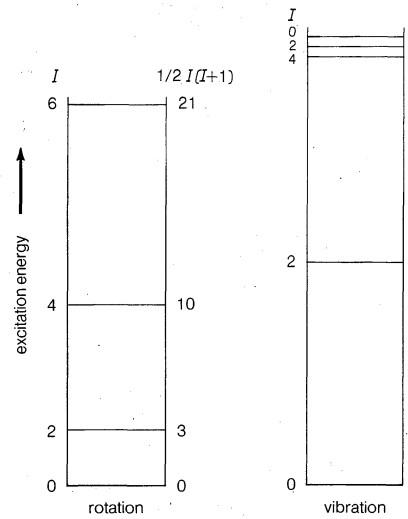


Figure 7: Patterns of the first few energy levels of collective excitation of an even–even nucleus.

The energy scale of the pattern is inversely proportional to the moment of inertia of the nucleus. Moment of inertia is a measure of the inertia involved in making a body rotate and is large for a flywheel, for example. If the internal structure of the nucleus were rigid it would have a relatively large moment of inertia and the spacing between the first two rotational states would be small, about ($\frac{1}{20}$) MeV. If, instead, the matter inside the nucleus were to behave like a liquid, it would move relatively little with the rotation, just as the water in an elliptical bucket does when the bucket is started rotating. This makes for a small moment of inertia, and spacing between the states is perhaps ten times as large as the rigid value. The observed situation in actual nuclei is about midway between these extremes, as may be explained in terms of the quantized orbits and the pairing energies of the nucleons that make up the nuclear fluid.

Nucleus analogous to a liquid droplet

A normally spherical nucleus is something like a liquid drop kept round by surface tension, and the vibrations of its surface are quantized, the second vibrational state having about twice the energy of the first. The nuclear surface tension is strong enough so that the first vibrational state is somewhat higher than the first rotational state of a typical deformed nucleus. Indeed, the heavy even–even nuclei with neutron number between 88 and 116, and again between 137 and 144, have exceptionally low first excited states and a ratio of second to first excitation energy near 3.33, indicating that these nuclei are permanently deformed and rotate, whereas those nearer the magic numbers 82 and 126, which do not have as many nucleons apart from closed shells to deform them, appear to be spherical, having the ratio nearer to two, characteristic of surface vibrations.

The simplest and most prevalent shape of a permanently deformed nucleus is ellipsoidal, with a circular cross section at right angles to an axis of symmetry. The nucleon orbits in such a nucleus are somewhat different from those in spherical nuclei. The deformation exerts a torque on the nucleon but none about the symmetry axis. The part of the angular momentum along this axis is quantized with a special quantum number (K), its different values corresponding to different energies. An even number of nucleons pair together to make this total quantum number equal to zero, and an odd mass nucleus has

several states with these quantum numbers that are half-integral because of the nucleon's one-half spin. With each of these, the nucleus can rotate collectively, and for each there is a band of states of increasing total angular momentum and energy. Two such bands observed in the fairly light aluminum-25 nucleus are shown in Figure 8.

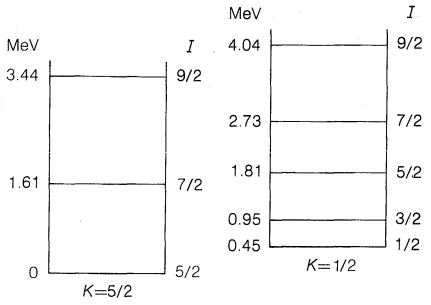


Figure 8: Two typical rotation bands in the odd-proton nucleus aluminum-25. The angular momentum *I* of lowest state of each band is due to internal nucleon motion, and in the states above it the nucleus is similar but rotates as a whole.

Deformation energies

The way the energy levels of an individual nucleon in a deformed nucleus vary with the deformation can be calculated approximately. The result is shown in Figure 9 for the lower levels. With a deformation (δ) equal to zero, the states are those of the spherical shell model and show gaps at the magic numbers 8, 20, 50, etc. As the deformation increases either in the prolate direction (positive deformation) or the oblate direction (negative deformation), the levels spread apart and eventually cross each other to make a complex pattern. The lines that slope upward with prolate deformation, for example, represent those states with angular momentum almost parallel to the nuclear axis, and thus a nucleon orbit in a plane almost perpendicular to the axis, in the direction that is squeezed by the deformation to raise the energy. The structure of intermediate and heavy deformed nuclei can be represented well by thinking of putting nucleons in

From D.R. Inglis, *Physics Today* (June 1969)

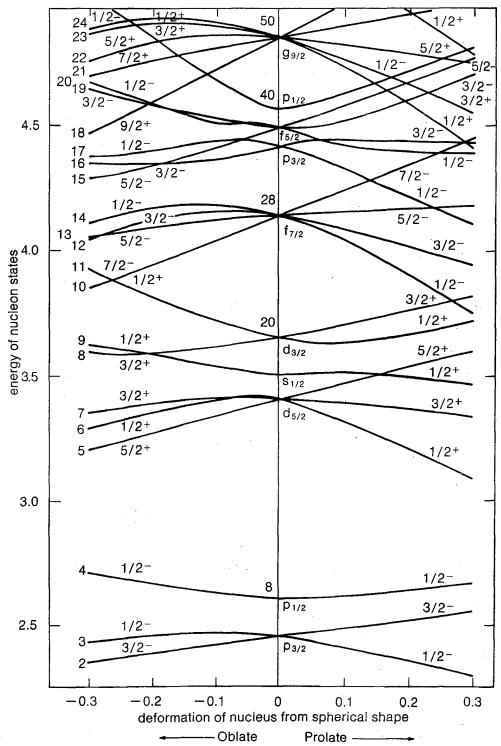


Figure 9: Energy levels of a single nucleon in a nucleus with an elliptical deformation from spherical shape. At the central line where the deformation vanishes, the levels converge to those of Figure 6.

these states, starting from the bottom, with two neutrons and two protons to a state as far as they go, with a given deformation. For an odd mass nucleus, the value of the quantum number K of the lone last nucleon is the total angular momentum of the nuclear ground state. Excited states of the nucleus are then formed by lifting nucleons from the topmost filled levels to nearby unfilled levels, accounting quite successfully for observed excited states.

Nuclear binding forces. The best understood part of the total force acting between nucleons is the electrostatic repulsion that one proton exerts on another. It is the same as the force between two electrons, giving rise to a potential energy inversely proportional to the separation or distance between the two protons. Since nuclear sizes are well-known and the potential varies slowly with separation, the contribution of repulsion to the energy of a nuclear state can be reliably calculated.

The energy levels of the three nuclei with atomic mass six have been plotted in Figure 5 after making allowance for the energy repulsion. It is remarkable that in this and similar situations the energies caused by all other interactions between the nucleons match up almost exactly between the neighbouring nuclei. From this, it may be concluded that the specific nuclear interactions, as they are called (those apart from electrostatic interaction), are practically the same between any two nucleons, be they two protons, two neutrons, or one of each.

The specific nuclear interactions that hold the nucleons together are short-ranged, strong, and attractive. The dependence on the separation may be represented approximately by a potential energy curve resembling a well with a radius or range of about 10^{-13} centimetre and a depth of roughly 50 million electron volts. That is, two nucleons interact quite strongly, but only when they are close together, their interaction being cut off much more abruptly than is the electrostatic interaction. This accounts for the abrupt change of particle density at the surface of a nucleus as compared with an atom.

The saturation property

Beyond these simple facts, nuclear forces are considered to be complicated in detail and no description of them is unique. A salient feature is the saturation property, which means, in effect, that one nucleon can interact with only about three other nucleons at a time. A phenomenological representation of this fact has been widely used to simplify calculations of nuclear structure. This formality describes the interactions as interchanging the identity of the two nucleons, interchanging mainly their positions but also to some extent their spins. It exploits the symmetry properties of the nuclear wave function to give the desired saturation, making a strong attraction when the function is symmetric in the interchange of the positions of the two nucleons and a weak repulsion when it is antisymmetric. (Because the Pauli principle strictly limits the degree to which it is symmetric, only a few neighbouring nucleons attract each other.) The more detailed and physical theories of nuclear interactions explain the important role of symmetrics. In its simplest form, the Pauli principle says that no two nucleons can behave in just the same way. They interact strongly only if they are similar (and symmetric) in space coordinates, but then they must differ in either spin or isospin. There are only two possible projections of spin, ½ and −½, and similarly two of isospin; so a group of four nucleons can differ from each other in either spin or isospin, and this is the number that can interact strongly with each other. A dramatic example is the exceptional stability of helium-4 and the fact that a fifth nucleon is not bound in helium-5.

The saturation property is responsible for the constant density of nuclear matter. Without it, each of the nucleons in a heavy nucleus would be attracted by all the other nucleons, and they would all crowd into a small space within about the range of nuclear forces of each other and thus would have an enormous binding energy, roughly proportional to the atomic mass number squared, rather than to the first power, in spite of the increased kinetic energy associated with this crowding.

Mesons in nucleons

More detailed theories of nuclear interactions are based on the fact that, in high-energy experiments, nucleons may be made to emit other particles known as mesons.

This has some similarity to the way they emit electrons in beta decay but involves much higher energies, 10^8 or 10^9 electron volts, and much stronger interactions between the nucleon and the emitted particle. Indeed, the building block of nuclei, the nucleon, may itself be considered to contain a whole host of mesons associated with it.

In all quantized systems, when particles are confined in a small space, the kinetic energy is high. When they are somehow permitted to spread out, the energy is lowered. When two atoms are brought together and their outer electrons can spread over both atoms, the energy is lowered and the atoms are bound in a molecule. Similarly, when two nucleons are close together, their mesons can spread over both and the energy is lowered, leading to an attractive interaction between the nucleons. When two atoms are pushed so close together that their closed shells overlap appreciably, the energy is raised and the atoms repel each other. Nuclear forces are rather generally considered similarly to include a very strong and very short range repulsion, though this is not certain. They may instead be markedly dependent on the relative momentum of the two interacting nucleons.

Particles have associated with them a length known as the de Broglie wavelength; that is, their shortest wavelength when they are moving with the speed of light. The wavelength lambda (λ) is equal to Planck's constant h divided by the mass m of the particle, 2π, and the velocity c of light, that is, $\lambda = h/2\pi mc$. For the pi meson with a mass 273 times that of an electron, this length is 1.4×10^{-13} centimetre, essentially equal to the range of the nuclear forces for which it is mainly responsible.

The pi meson has no spin, but the heavier rho meson and omega meson do have spin, and their participation in the interaction process leads to strong spin–orbit coupling. Any particle with spin, carried around in an orbital motion, has spin–orbit coupling because of the relativistic precession of the spin's coordinate system accelerated in the orbital motion. For the bare nucleon spin, however, this effect is much too weak and is apparently greatly enhanced by the spin-carrying mesons being accelerated between the two nucleons. The meson exchange also contributes to another part of the interaction known as a tensor interaction. This depends on the direction of the spins relative to the line between the two nucleons. If it is positive when the two spins point in the same direction at right angles to this line, it is negative when they are parallel to it.

Meson exchange

Such meson theories suggest the form but do not yet reliably give the magnitude of various terms of the nuclear interaction. Detailed calculations of nuclear structure thus assume a complicated form of interaction and, after determining the magnitude of the various terms by fitting some of the available experimental data, go on to calculate other results that may be compared with other experiments to test the consistency of the analysis.

NUCLEAR PHENOMENA AND REACTIONS

Radioactivity and the region of stable nuclei. Both alpha and beta radioactivity (see RADIOACTIVITY) help to determine which nuclei are present in nature. The region of stable nuclei is indicated in Figure 1, in which the number of neutrons is plotted against the number of protons. Light nuclei have about equal numbers of protons and neutrons, whereas the heavy nuclei have a considerable neutron excess, with about 1½ times as many neutrons as protons. This can be understood in terms of the specific nuclear interaction being the same between all nucleons, the positive potential energy of the electric repulsion between protons, and the freedom of nucleons through beta decay to be either neutrons or protons, whichever makes the energy lower. Without the electric energy, the protons and neutrons would be equal in number because neutrons and protons separately fill the various nucleon states, starting with the lowest. Equal numbers of protons and neutrons fill the states up to the same level, but to make the neutron number greater than the proton number, nucleons would have to be lifted to a higher level, increasing the total kinetic energy. The addition of the positive electric energy between protons, however, tends to make it energetically advantageous to change protons to neutrons and counterbalances some of the excess kinetic energy and makes it favourable to have a limited neutron excess. The electric effect is unimportant in light nuclei in which each proton is repelled by only a few others, but it is increasingly important in heavier nuclei, making the graph in Figure 1 curve upward until the excess of neutrons over protons is about one-half the proton number for the heaviest nuclei. The nuclei, sufficiently stable to be present in nature, end at proton number 92 and nuclear number 238, because in that vicinity the increasing electric repulsion makes nuclei too unstable mainly against alpha decay, and also against spontaneous fission (see NUCLEAR FISSION).

Spontaneous fission

The increase of electric energy toward the heavier nuclei is also responsible for the upward curvature of the graph of energy per nucleon in Figure 3. The fact that this internal energy per nucleon is lower for nucleon number 100 or 140 than for 240 means that a nucleus with nucleon number about 240 can separate into those two fragments with a decrease of the total internal energy, the excess energy going largely into the kinetic energy with which they fly apart, propelled by the electric repulsion between the fragments. This is the act of fission, made possible as demonstrated by the shape of the curve in Figure 3. In addition, one or more neutrons, called prompt neutrons, are released at the time of fission.

The two fission fragments have the neutron–proton ratio about the same as that in the original nucleus. Because of the curvature of the region of stable nuclei in Figure 1, this means that they have neutron excesses considerably greater than those of the stable nuclei of the same mass number. They get rid of this difference in neutron excess by a series of beta disintegrations, the first relatively energetic and short-lived, with a half-life of a few seconds, and the last with a half-life that may be measured in years. The original fission fragments thus decay to the mostly radioactive fission products, such as the ecologically troublesome strontium-90 with a half-life of 28.1 years, that either were dispersed in the atmosphere from nuclear weapons tests or that must be carefully disposed of from power reactors. Some, such as iodine-131, are useful as sources of radiation for medical therapy or diagnosis and for tracer and monitoring techniques in science and industry.

Part way down the chains of beta decay from some fission fragments a neutron is emitted. Since beta decay is inherently slower than neutron emission, there are a few delayed neutrons in addition to the prompt neutrons from the fragments preceding beta decay, a fact of importance for the stability of the chain reaction in a power reactor.

Nuclear reactions. The first nuclear reaction induced by an incident particle was observed by Rutherford in 1919 when he let natural alpha particles hit nitrogen nuclei and observed that protons were emitted. This was the first artificial transmutation of elements, changing nitrogen to oxygen.

First observed nuclear reaction

Nuclear reactions are powerful tools for learning about the properties of nuclear states, such as energies, angular momenta, the extent to which states may be single-particle states, and the way nucleons may be partially correlated into alpha-particle-like clusters at the surface of a nucleus. In laboratory investigations, nuclear reactions are induced either by neutrons from nuclear research reactors, nuclear bombs, or by charged particles, such as protons, deuterons (^{2}H), tritons (^{3}H), and helium nuclei (^{3}He and ^{4}He), accelerated to precisely known energies in electrostatic accelerators, linear accelerators, or cyclotrons. In addition to such light-particle reactions, there are heavy-ion reactions such as those induced by carbon nuclei, requiring higher energy acceleration of the heavier projectile to overcome the higher barrier of the electric repulsion.

Neutron bombardment for nuclear reactions has the advantage that it encounters no potential barrier at all, and neutrons with energies with a fraction of an electron volt or a little more can be used, their energies determined by time-of-flight methods, to pick out the energies of very narrow and closely clustered states of heavy nu-

clei at an excitation energy equivalent to the neutron's binding energy.

A greater variety of experiments is possible with charged particles, their energies being commonly varied over a wide range from under a million electron volts to ten times that or more. Precisely defined energies of nuclear states are seen in two ways. As the bombarding energy is varied slowly, resonances are observed; that is, a certain reaction takes place at just the right bombarding energy but not at nearby energies, making a peak if one plots a curve of the intensity. The bombarding energy is said to be in resonance with the energy of a state of the compound nucleus, the nucleus made when the projectile enters the target nucleus. With the bombarding energy fixed to cause the reaction, a second significant energy is that of the outgoing particle, which may be observed to have several precisely defined values corresponding to several states of the final nucleus composed of the target nucleus plus bombarding particle and minus the outgoing particle, the total energy being divided between kinetic energy of the particle and the excitation energy of the final nucleus. The highest energy product particles are those that leave the final nucleus in its ground state. They may have either more or less energy than had the incoming projectile, the reaction being called exoergic or endoergic, depending on whether the final nucleus and outgoing particle together have more or less binding energy than the target nucleus and projectile.

Resonance energy

The uncertainty principle in the quantum domain states not only that position and momentum together are uncertain but also that energy and time together are uncertain, or that a state cannot have a precisely defined energy unless it lasts a long time. In some reactions, the sharpness of the resonances indicates that the states of the product nucleus have relatively long lifetimes (though perhaps only microseconds). This means that the projectile does not retain its energy in the state so as to be able to come right out again, but rather that the excitation energy is divided up among many nucleons, producing what is known as a compound nucleus. Only after some time will the excitation energy be concentrated in the projectile particle, causing it to be re-emitted, or in the product particle of the reaction observed.

There is another class of reactions in which the projectile, such as a deuteron or triton, as it flies by the target nucleus is either stripped of one or more of its nucleons, which are deposited in the target to make the product nucleus, or picks up one or more nucleons from the target as it goes by. These "stripping" and "pick-up" reactions do not show sharp resonances, there being no compound nucleus involved; but there are sharply defined energies of the outgoing particles, corresponding to states of the product nucleus, when the bombarding energy is definite.

Stripping

A typical stripping reaction is the reaction in which an incident deuteron drops its neutron into the target nucleus and the remaining proton flies off at an angle from the incident direction. The deuteron has internal kinetic energy, and in it the proton and neutron in the deuteron have equal and opposite internal momentum at any instant as they vibrate back and forth past each other. When the neutron joins the target nucleus at the nuclear surface, its internal momentum from the deuteron helps determine its angular momentum in the product nucleus, and at the same time the equal and opposite internal momentum of the proton contributes to the angle at which the proton flies off. Thus, observation of the angular distribution of the protons is useful in determining the angular momentum of the nuclear state formed.

In still more refined observations of reactions, either the incident particle or the outgoing particle is polarized, meaning that its spin direction is determined. By such means not only the orbital angular momentum but also the total angular momentum of the neutron in the final state may be determined.

There are reactions with more than one outgoing particle. For example, in the reaction in which one proton, at incident energies of 200 to 400 million electron volts, knocks out another proton from the nucleus, the angular distributions of the two outgoing protons, observed simultaneously, indicate the momentum distribution of a proton before it was plucked from a low closed shell within the nucleus. (An electron can similarly be used to knock out a proton.)

Among the peculiar properties of heavy-ion reactions, such as a carbon-12 nucleus hitting another carbon-12 nucleus, is the apparent formation of temporary states associated with the contact of the two surfaces, perhaps resembling molecular vibrations, before complete amalgamation of the two heavy nuclei. This has some similarity to one aspect of the fission of uranium (see NUCLEAR FISSION). Fission is a neutron-induced reaction in which the target nucleus flies apart into two almost equal fragments. It may be described in terms of a potential barrier. The fission barrier, however, is such that quantized vibrations of the almost-separated fragments give rise to quasi-stationary states that manifest themselves in regularly spaced groups of sharp resonances.

Heavy-ion reactions

Interaction of nuclei with electromagnetic radiation. A straight radio antenna is a dipole radiator. The electrons that constitute an electric current rush back and forth, determining which end is positive and which negative; and the consequent electric and magnetic fields around it are propagated out into space, most strongly in the plane at right angles to the antenna. Electric quadrupole radiation may be produced by placing two antennas end to end and making their currents go always in opposite directions, so that both outer ends are positive at one instant, both inner ends at another. The radiations from the two cancel each other in some directions and add in others, so together they radiate most strongly in the direction midway between the antenna direction and the plane at right angles to it. In a plane that includes the antenna direction, this makes a four-lobed rosette radiation pattern.

Gamma rays are extremely high frequency electromagnetic radiation, and nuclei, as they make transitions from one state to another, can emit or absorb either dipole or quadrupole radiation in quantized amounts of energy called photons. As in the case of atoms emitting lower frequency photons as visible light (see ATOMIC STRUCTURE), the energy of a photon is determined by the energy difference of the quantized states involved, and the energy is equal to Planck's constant times the frequency, or $h\nu$.

Nuclear states have another property called parity, which is formally defined in terms of symmetry but may be described by saying that a state has even or odd parity according to whether the sum of the angular momentum quantum numbers of all the nucleons in it is even or odd. Electric dipole radiation is emitted when there is a change of parity between initial and final state, electric quadrupole radiation when there is no change in parity. In a nucleus that is essentially a closed shell and a single proton, called a one-proton state, strong dipole radiation is emitted when the orbital angular momentum quantum number l changes from zero to one, for example. Much weaker quadrupole radiation could be emitted if it were to change from zero to two. A proton in an orbit has a back-and-forth motion like that of the electrons in an antenna; but in a quantized system such as the nucleus, it radiates only when there is a change in this motion. The intensity of the electric dipole radiation emitted can be calculated when a proton changes from one orbit to the next. Even for nuclei with proton outside a closed shell the observed electric dipole radiation is considerably weaker than calculated, indicating that there are no pure one-nucleon states. When the orbital quantum number changes by two, electric quadrupole radiation is possible but is much weaker still. There is also magnetic dipole radiation corresponding to having the spin magnetic moment flip alternately up and down, as well as higher magnetic and electric multipoles.

Parity

When all the nucleons of a nucleus move collectively with a change of nuclear shape, however, the situation as far as radiation is concerned is quite different because the nucleons cooperate systematically in causing the radiation. A prolate nucleus may vibrate between a longer and shorter ellipsoidal shape, for example. The motion of the positive charge inward and outward at the two ends,

always moving in opposite directions, is analogous to the case of the two antennas, and like them it emits electric quadrupole radiation. Because many protons move together and because the intensity of radiation is proportional to the square of the strength of the electric field involved, the intensity of this radiation is in many cases considerably stronger than the electric dipole radiation from a single proton. This enhancement applies to radiations from changes of either vibration or rotation quantum numbers of deformed nuclei, a fact that facilitates the study of the properties of such states.

It already has been stated that when nuclei make transitions from one state to another they can emit high-energy quanta called gamma rays. The most accurate observations of gamma-ray energies are made with a bent-crystal spectrometer in studies analogous to those of atomic spectroscopy with a diffraction grating.

When a slow neutron is absorbed with a binding energy of about eight million electron volts, the nucleus is excited by the same amount, and this energy is emitted usually in a cascade of several gamma rays as transitions are made down a ladder of states. Similar cascades occur when deformed nuclei are excited to high rotational states by heavy-ion bombardment. Beta decay (the emission of electrons or positrons by an excited nucleus) produces a daughter nucleus, which itself gamma decays, thus giving rise to a whole series of beta–gamma cascades. In these various cascades, two successive radiations may usefully be investigated at the same time in a coincidence experiment. The frequency of the coincidence is dependent on the angle between the two rays, and any slight time delay recorded from the first to the second emission indicates the half-life of the intermediate state between them. This angular correlation gives information about the multipolarity of the radiations and the angular momenta and parities of the three states involved.

Beta decay cascade

Excitations caused by nuclear forces in nuclear reactions also are often followed by gamma emission, and studies of the angular correlation between the deflection angle and the gamma-ray pattern can be revealing. A striking example is observed in the inelastic scattering of alpha particles from such nuclei as magnesium-24, $^{24}_{12}Mg$ (nuclei with just enough nucleons to make up internal alpha particles), exciting the nucleus from the nonrotating ground state to the first excited state, which rotates with a nuclear spin equal to two. The angular distribution of the subsequent quadrupole gamma radiation has a rosette pattern, the orientation of which rotates backward as the deflection angle of the alpha-particle detector is increased.

Excitation of nuclear rotations and vibrations by the electric field of bombarding ions may also be looked upon as an interaction of the nucleus with electromagnetic radiations, an absorption process because the rapidly varying electric field of the passing ion may be analyzed as a superposition of radiations of various multipolarities and gives information supplementing that obtained from radiation.

BIBLIOGRAPHY. M.G. MAYER and J.H.D. JENSEN, *Elementary Theory of Nuclear Shell Structure* (1955); J.M. BLATT and V.F. WEISSKOPF, *Theoretical Nuclear Physics* (1952); S. DEBENEDETTI, *Nuclear Interactions* (1964); M.A. PRESTON, *Physics of the Nucleus* (1962); R.R. ROY and B.P. NIGAM, *Nuclear Physics: Theory and Experiment* (1967); F. AJZENBERG-SELOVE and T. LAURITSEN, "Energy Levels of Light Nuclei," *Nucl. Physics*, 11:1–340 (1959), 78:1–176 (1966), and A114:143–160 (1968); D.R. INGLIS, "Nuclear Models," *Physics Today*, 22:29–40 (1969). More advanced treatises: G.E. BROWN, *Unified Theory of Nuclear Models and Forces*, 2nd rev. ed. (1967); A. BOHR and B.R. MOTTELSON, *Nuclear Structure*, vol. 1, *Single-particle Motion* (1969).

(D.R.I.)

Number Games and Other Mathematical Recreations

Mathematical recreations comprise puzzles and games that vary from naive amusements to sophisticated problems, some of which have never been solved. They may involve arithmetic, algebra, geometry, theory of numbers, graph theory, topology, matrices, group theory, combinatorics (dealing with problems of arrangements or designs), set theory, symbolic logic, or probability theory. Any attempt to classify this colourful assortment of material is at best arbitrary. Included in this article are the history and the main types of number games and mathematical recreations and the principles on which they are based. Details, including descriptions of puzzles, games, and recreations mentioned in the article, will be found in the references listed in the bibliography.

At times it becomes difficult to tell where pastime ends and serious mathematics begins. An innocent puzzle requiring the traverse of a path may lead to technicalities of graph theory; a simple problem of counting parts of a geometric figure may involve combinatorial theory; dissecting a polygon may involve transformation geometry and group theory; logical inference problems may involve matrices. A problem regarded in medieval times as very difficult may prove to be quite simple when attacked by the mathematical methods of today.

Mathematical recreations have a universal appeal. The urge to solve a puzzle is manifested alike by young and old, by the unsophisticated as well as the sophisticated. An outstanding English mathematician, G.H. Hardy, observed that professional puzzle makers, aware of this propensity, exploit it diligently, knowing full well that the general public gets an intellectual kick out of such activities.

The relevant literature has become embarrassingly extensive, particularly since the beginning of the 20th century. Some of it is repetitious, but surprisingly enough, successive generations have found the older chestnuts to be quite delightful, whether dressed in new clothes or not. Much newly created material is continually being added. Over 300 books and monographs appeared between 1925 and 1970, exclusive of reprints and new editions. Even more voluminous is the literature in mathematical and educational journals and in periodicals devoted exclusively to recreational mathematics, including the former magazines *Sphinx* (Belgium, 1931–1939) and *Recreational Mathematics Magazine* (U.S., 1961 to 1964) and the current periodicals *Journal of Recreational Mathematics* (U.S.) and *Pythagoras* (The Netherlands, England, and U.S.).

History

EARLY HISTORY

Men have always taken delight in devising "problems" for the purpose of posing a challenge or providing intellectual pleasure. Thus, many mathematical recreations of early origin that have reappeared from time to time in new dress seem to have survived chiefly because they appeal to man's sense of curiosity or mystery. A few survived from the ancient Greeks and Romans: little was known about them during the Dark Ages, but a strong interest in such problems arose during the Middle Ages, stimulated partly by the invention of printing, partly by enthusiastic writers of arithmetic texts, and partly by the rivalry and disputations among early algebraists and scholars. Such activities were most prominent on the Continent, particularly in Italy and Germany. Notable contributors included Rabbi ben Ezra (1140), Fibonacci (Leonardo of Pisa; 1202), Robert Recorde (1542), and Geronimo Cardano (1545), among others.

Kinds of problems. The problems in general were two kinds: those involving the manipulation of objects, and those requiring computation. The first required little or no mathematical skill, merely general intelligence and ingenuity, as for example, so-called decanting and difficult crossings problems. A typical example of the former is how to measure out one quart of a liquid if only an eight-, a five-, and a three-quart measure are available. Difficult crossings problems are exemplified by the dilemma of three couples trying to cross a stream in a boat that will hold only two persons, with each husband too jealous to leave his wife in the company of either of the other men. Many variants of both types have appeared over the years.

Manipulation of objects and computation

Some examples. Problems involving computation also took on a variety of forms; some were as follows:

Finding a number. Think of a number, triple it, and take half the product; triple this and take half the result; then divide by 9. The quotient will be one-fourth the original number.

"God-Greet-You" problems. For example, in "God greet you, all you 30 companions," someone says: "If there were as many of us again and half as many more, then there would be 30 of us." How many were there?

The chess-board problem. How many grains of wheat are required in order to place one grain on the first square, 2 on the second, 4 on the third, and so on for the 64 squares?

The lion in the well. This is typical of many problems dealing with the time required to cover a certain distance at a constant rate while at the same time progress is hindered by a constant retrograde motion. "There is a lion in a well whose depth is 50 palms. He climbs $^1/_7$ of a palm daily and slips back $^1/_9$ of a palm. In how many days will he get out of the well?

Courier problems. These are typified by the movements of bodies at given rates in which some position of these bodies is given and the time required for them to arrive at some other specified position is demanded.

PIONEERS AND IMITATORS

The 17th century produced books devoted solely to recreational problems not only in mathematics but frequently in mechanics and natural philosophy as well. The first important contribution was that of the Frenchman Claude-Gaspar Bachet de Méziriac, one of the earliest pioneers in this field, who is remembered for two mathematical works: his *Diophanti*, the first edition of a Greek text on the theory of numbers (1621), and his *Problèmes plaisans et delectables qui se font par les nombres* (1612). The latter passed through five editions, the last as late as 1959; it was the forerunner of similar collections of recreations to follow. The emphasis was on arithmetic rather than geometric puzzles. Among the outstanding problems given by Bachet were questions involving number bases other than ten; card tricks; watch-dial puzzles depending on numbering schemes; the determination of the least number of weights which would enable one to weigh any integral number of pounds from one pound to 40, inclusive; and difficult crossings or ferry problems.

Books and collections

In 1624 a French Jesuit, Jean Leurechon, writing under the pen name of van Etten, published *Récréations mathématiques*. This volume struck the popular fancy, passing through at least 30 editions before 1700 despite the fact that it was based largely on the work of Bachet, from whom he took the simpler problems, disregarding the more significant portions. Yet it did contain some original work, and it served as a model for others, including Mydorge and Schwenter. The first English edition (1633) bore the title: *Mathematicall Recreations, or a Collection of Sundrie Problemes, extracted out of the Ancient and Moderne Philosophers, as Secrets in Nature, and Experiments in Arithmeticke, Geometrie, Cosmographie, Horologographie, Astronomie, Navigation, Musicke, Opticks, Architecture, Staticke, Machanicks, Chimestrie, Waterworkes; Fireworks, etc. Not vulgarly made manifest untill this Time. . . . Most of which were written first in Greeke and Latine, lately compiled in French, by HENRY VAN ETTEN Gent. And now delivered in the English Tongue with the Examinations, Corrections, and Augmentations* [*translated by William Oughtred*].

The rising tide of interest was exploited by French mathematicians Claude Mydorge, whose *Examen du livre des récréations mathématiques* was published in 1630, and Denis Henrion, whose *Les Récréations mathématiques avec l'examen de ses problèmes en arithmétique, géométrie, méchanique cosmographie, optique, catoptrique*, etc., based largely upon Mydorge's book, appeared in 1659. Leurechon's book, meanwhile, had found its way into Germany, to be imitated by Daniel Schwenter, a professor of Hebrew, Oriental languages, and mathematics, who assiduously compiled a comprehensive collection of recreational problems based on a translation of Leurechon's book, together with many other problems that he himself had previously collected. This work appeared posthumously in 1636 under the title *Deliciae Physico-mathematicae oder Mathematische und Philosophische Erquickstunden*. Immensely popular, Schwenter's book was enlarged by two supplementary editions in 1651–53. For some years thereafter Schwenter's enlarged edition was the most comprehensive treatise of its kind, although in 1641–42 the Italian Jesuit Mario Bettini had issued a two-volume work called *Apiaria Universae Philosophiae Mathematicae in Quibus Paradoxa et Nova Pleraque Machinamenta Exhibentur*, which was followed in 1660 by a third volume entitled *Recreationum Mathematicarum Apiaria Novissima Duodecim. . . .* And in 1665 one Johann Mohr in Schleswig published an imitation of Schwenter under the title of *Arithmetische Lustgarten*.

In England, somewhat belatedly, William Leybourn, a mathematics teacher, textbook writer, and surveyor, in 1694, published his *Pleasure with Profit: Consisting of Recreations of Divers Kinds, viz., Numerical, Geometrical, Mechanical, Statical, Astronomical, Horometrical, Cryptographical, Magnetical, Automatical, Chymical, and Historical*. The title page further states that the purpose of the book was to "recreate ingenious spirits and to induce them to make farther scrutiny into these sublime sciences, and to divert them from following such vices, to which Youth (in this Age) are so much inclined." Much of the volume is conventional textbook material, for most of Leybourn's published works grew out of his teaching.

18TH AND 19TH CENTURIES

The 18th century saw a continuation of this interest. Published in England were volumes by Edward Hatton, Thomas Gent, Samuel Clark, and William Hooper. In 1775 Charles Hutton published five volumes of extracts from the *Ladies' Diary* dealing with "entertaining mathematical and poetical parts." On the Continent there appeared several writers, including: Christian Pescheck, Abat Bonaventura, the Dutch writer Paul Halcken, and A.H. Guyot's four volumes of *Nouvelles Récréations physiques et mathématiques*, etc. (1769, 1786). But by far the outstanding work was that of Jacques Ozanam, the precursor of books to follow for the next 200 years. First published in four volumes in 1694, his *Récréations mathématique et physiques* went through many editions; based on the works of Bachet, Mydorge, Leurechon, and Schwenter, it was later revised and enlarged by Montucla, then translated into English by Charles Hutton (1803, 1814) and again revised by Edward Riddle (1840, 1844).

The work of Ozanam

The first half of the 19th century produced only a moderate number of lesser writers on mathematical recreations, but the second half of the 19th century witnessed a rising crescendo of interest, culminating in the outstanding contributions of Édouard Lucas, C.L. Dodgson (Lewis Carroll), and others at the turn of the century. Lucas' four-volume *Récréations mathématiques* (1882–94) became a classic. The mathematical recreations of Dodgson included *Symbolic Logic* and *The Game of Logic; Pillow Problems* and *A Tangled Tale*, 2 vol. (1885–95).

20TH CENTURY

Among the more colourful figures at the turn of the 20th century was an American, Sam Loyd, or rather two Loyds, father and son. Tremendously successful in making puzzles, the elder Loyd sold his weekly puzzle column to a national syndicate for years, and, in addition, created or adapted hundreds of mechanical puzzles fashioned of cardboard, wood, and metal that were also financially rewarding. When he died in 1934 at the age of 60, it was estimated that Loyd II had produced at least 10,000 puzzles.

In Germany, Hermann Schubert published *Zwölf Geduldspiele* in 1899 and the *Mathematische Mussestunden* (3rd ed., 3 vol.) in 1907–09. Between 1904 and 1920 Wilhelm Ahrens published several works, the most significant being his *Mathematische Unterhaltungen und Spiele* (2 vol.) with an extensive bibliography.

Among British contributors, Henry Dudeney, a contributor to the *Strand Magazine*, published several very pop-

ular collections of puzzles that have been reprinted from time to time (1917–67). Although the first edition of W.W. Rouse Ball's *Mathematical Recreations and Essays* appeared in 1892, it soon became a classic, largely because of its scholarly approach. After passing through ten editions it was revised by the Canadian professor H.S.M. Coxeter in 1938; it is still a standard reference.

Outstanding work was that of Maurice Kraitchik, editor of *Sphinx* and author of several well-known works published between 1900 and 1942.

Modern shift in emphasis

About the middle third of the 20th century, there was a gradual shift in emphasis on various topics. Up to that time interest had focussed largely on numerical curiosities; simple geometric puzzles; arithmetical story problems; paper folding and string figures; geometric dissections; manipulative puzzles; tricks with numbers and with cards; magic squares; those venerable diversions concerning angle trisection, duplication of the cube, squaring the circle, and the elusive fourth dimension. Interest began to swing toward more mathematically sophisticated topics: cryptograms; recreations involving modular arithmetic, numeration bases, and number theory; graphs and networks; lattices, group theory; topological curiosities; packing and covering; flexagons; manipulation of geometric shapes and forms; combinatorial problems; probability theory; inferential problems; logical paradoxes; fallacies of logic; paradoxes of the infinite.

Types of games and recreations

ARITHMETIC AND ALGEBRAIC RECREATIONS

Number patterns and curiosities. Some of the properties of the natural numbers when operated upon by the ordinary processes of arithmetic reveal rather remarkable patterns, affording pleasant pastimes. For example:

$1 \times 8 + 1 = 9$	$3 \times 37 = 111$	$(1)^2 = 1$
$12 \times 8 + 2 = 98$	$6 \times 37 = 222$	$(11)^2 = 121$
$123 \times 8 + 3 = 987$	$9 \times 37 = 333$	$(111)^2 = 12321$
$1234 \times 8 + 4 = 9876$	$12 \times 37 = 444$	$(1111)^2 = 1234321$
etc.	etc.	etc.

Another type of number pleasantry concerns multigrades; *i.e.*, particular identities between sets of numbers and their powers—*e.g.*,

$1^n + 6^n + 8^n = 2^n + 4^n + 9^n$ (for $n = 1$ or 2);
$36^2 + 37^2 + 38^2 + 39^2 + 40^2 = 41^2 + 42^2 + 43^2 + 44^2$.

An easy method of forming a multigrade is to start with a simple equality *e.g.*, $1 + 5 = 2 + 4$—then add, for example, 5 to each term: $6 + 10 = 7 + 9$. A second order multigrade is obtained by "switching sides" and combining, as shown below:

$$1^n + 5^n + 7^n + 9^n = 2^n + 4^n + 6^n + 10^n \quad (n = 1 \text{ or } 2).$$

On each side the sum of the first powers (S_1) $= 22$, and of the second powers (S_2) $= 156$.

Ten may be added to each term to derive a third-order multigrade:

$$11^n+15^n+17^n+19^n=12^n+14^n+16^n+20^n \quad (n=1 \text{ or } 2).$$

Switching sides and combining, as before:

$1^n + 5^n + 7^n + 9^n + 12^n + 14^n + 16^n + 20^n$
$= 2^n + 4^n + 6^n + 10^n + 11^n + 15^n + 17^n + 19^n$
In this example $S_1 = 84$, $S_2 = 1{,}152$, and $S_3 = 17{,}766$ ($n = 1$, 2, or 3).

This process can be continued indefinitely to build multigrades of successively higher orders. Similarly, all terms in a multigrade may be multiplied or divided by the same number without affecting the equality. Many variations are possible, palindromic multigrades that read the same backward and forward, for example, and prime-number multigrades.

Narcissistic numbers

Other number curiosities and oddities are to be found. Thus, narcissistic numbers are numbers that can be represented by some kind of mathematical manipulation of their digits. One such manipulation occurs when a whole number, or integer, is the sum of the nth powers of its digits; *e.g.*, $153 = 1^3 + 5^3 + 3^3$, which is called a perfect digital invariant. On the other hand, a recurring digital invariant is illustrated by:

$$55:5^3 + 5^3 = 250;$$
$$250:2^3 + 5^3 + 0^3 = 133;$$
$$133:1^3 + 3^3 + 3^3 = 55.$$

(From *Mathematics on Vacation*, Joseph Madachy; Charles Scribner's Sons.)

A variation of such digital invariants would be:

$$165{,}033 = 16^3 + 50^3 + 33^3.$$

Another curiosity is exemplified by a number that is equal to the nth power of the sum of its digits:

$$81 = (8 + 1)^2 = 9^2;$$
$$4913 = (4 + 9 + 1 + 3)^3 = 17^3.$$

An automorphic number is an integer whose square ends with the given integer, as $(25)^2 = 625$, and $(76)^2 = 5776$. Strobogrammatic numbers are numbers that read backwards after having been rotated through 180°; *e.g.*, 69, 96, 1001.

It is not improbable that such curiosities should have suggested intrinsic properties of numbers bordering on mysticism.

Digital problems. The problem of the four n's calls for the expression of as large a sequence of integers as possible, beginning with 1, representing each integer in turn by a given digit used exactly four times. The answer depends upon the rules of operation that are admitted. Two partial examples are shown.

Problem of the four n's

For four "1s":

$1 = 1 + \frac{1}{1} - 1$
$2 = 1 + 1 + 1 - 1$
$3 = 1 + 1 + \frac{1}{1}$
$4 = 1 + 1 + 1 + 1$
$5 = (1 + 1 + 1)! - 1$
etc.

For four "4s" it would be possible to have:

$1 = (4 \div 4) \cdot (\frac{4}{4})$
$2 = \frac{4}{4} + \frac{4}{4}$
$3 = \frac{4}{4} + 4/\sqrt{4}$
$4 = \sqrt{(4)(4)} \cdot (\frac{4}{4})$
$5 = \sqrt{4} + \sqrt{4} + \frac{4}{4}$
etc.

(In M. Bicknell & V. Hoggatt, "64 Ways to Write 64 Using Four 4's," *Recreational Mathematics Magazine*, No. 14, Jan.–Feb. 1964, p. 13.)

Obviously, many alternatives are possible; *e.g.*, $7 = 4 + \sqrt{4} + 4/4$ could also be expressed as $4!/4 + 4/4$, or as $44/4 - 4$. The factorial of a positive integer is the product of all the positive integers less than or equal to the given integer; *e.g.*, "factorial 4," or $4! = 4 \times 3 \times 2 \times 1$. If the use of factorial notation is not allowed, it is still possible to express the numbers from 1 to 22 inclusive with four "4s"; thus $22 = (4 + 4)/.4 + \sqrt{4}$. But if the rules are extended many additional combinations are possible.

A similar problem requires that the integers be expressed by using the first m positive integers, $m > 3$ ("m is greater than three"), allowing a finite number of operational symbols used in elementary algebra. For example, using the digits 1, 2, 3, and 4:

$2 = 4 - 3 + 2 - 1$
$4 = \sqrt{4} + 3 - 2 + 1$
$6 = \sqrt{4} + 3 + 2 - 1$
$8 = \sqrt{4} + 3 + 2 + 1$
$9 = 4 + 3 + (2 \cdot 1).$

Such problems have many variations; for example, over 100 ways of arranging the digits 1 to 9, *in order*, to give a value of 100 have been demonstrated.

All of these digital problems require considerable ingenuity, but, in reality, they involve little significant mathematics.

Cryptarithms. These are mathematical problems usually calling for addition, subtraction, multiplication, or division in a set of operations in which the digits have been replaced by letters of the alphabet or some other symbols. The term "crypt arithmetic" was introduced in 1931, when the following multiplication problem appeared in the former Belgian journal *Sphinx:*

$$\begin{array}{r} ABC \\ DE \\ \hline FEC \\ DEC \\ \hline HGBC \end{array}$$

An analysis of the puzzle suggested the general method of solving a relatively simple cryptarithm:

1. In the second partial product D $\times$ A = D, hence A = 1.
2. D $\times$ C and E $\times$ C both end in C; since for any two digits 1–9 the only multiple that will produce this result is 5 (zero if both digits are even, 5 if both are odd), C = 5.
3. D and E must be odd. Since both partial products have only three digits, neither D nor E can be 9. This leaves only 3 and 7. In the first partial product E $\times$ B is a number of two digits, while in the second partial product D $\times$ B is a number of only one digit. Thus E is larger than D, so E = 7 and D = 3.
4. Since D $\times$ B has only one digit, B must be 3 or less. The only two possibilities are 0 and 2. B cannot be zero because 7B is a two digit number. Thus B = 2.
5. By completing the multiplication, F = 8, G = 6, and H = 4.
6. Answer: 125 $\times$ 37 = 4,625.

(From *150 Puzzles in Cryptarithmetic* by Maxey Brooke; Dover Publications, Inc., New York, 1963. Reprinted through the permission of the publisher.)

Such puzzles had apparently appeared, on occasion, even earlier; in recent years they have become quite popular. Alphametics refers specifically to cryptarithms in which the combinations of letters make sense, as in one of the oldest and probably best known of all alphametics:

$$\begin{array}{r} SEND \\ +\ MORE \\ \hline MONEY \end{array}$$

Unless otherwise indicated, convention requires that the initial letters of an alphametic cannot represent zero, and that two or more letters may not represent the same digit. If these conventions are disregarded, the alphametic must be accompanied by an appropriate clue to that effect. Some cryptarithms are quite complex and elaborate, and have multiple solutions. Electronic computers have been used for the solution of such problems.

Paradoxes and fallacies. Mathematical paradoxes and fallacies have long intrigued mathematicians. A mathematical paradox may be described as a mathematical conclusion so unexpected that it is difficult to accept even though every step in the reasoning is known to be valid. A mathematical fallacy, on the other hand, is an instance of improper reasoning leading to an unexpected result that is patently false or absurd. The error in a fallacy generally violates some principle of logic or mathematics, often unwittingly. Such fallacies are quite puzzling to the tyro, who, unless he is aware of the principle involved, may well overlook the subtly concealed error. A sophism is a fallacy in which the error has been knowingly committed, for whatever purpose. If the error introduced into a calculation or a proof leads innocently to a *correct* result, the result is a "howler," often called an illegal operation or "making the right mistake."

Many paradoxes arise from the concepts of infinity and limiting processes. For example, the infinite series

$$1 + \tfrac{1}{2} + \tfrac{1}{4} + \tfrac{1}{8} + \ldots$$

appears at first glance to have a continually greater "sum" the more terms are included, although the sum actually can never be equal to 2, although it approaches nearer and nearer to 2 as more terms are included. On the other hand, the series

$$1 + \tfrac{1}{2} + \tfrac{1}{3} + \tfrac{1}{4} + \tfrac{1}{5} + \tfrac{1}{6} + \ldots$$

is divergent and has no limit, the "sum" becoming ever larger the more terms are taken. Another paradox is the fact that there are just as many even natural numbers as there are even and odd numbers all together, thus contradicting the notion that "the whole is greater than any of its parts." This seeming contradiction arises from the properties of collections containing an infinite number of objects. Since both are infinite, they are for both practical and mathematical purposes equal.

Zeno's paradoxes

The so-called paradoxes of Zeno (*c.* 450 BC) are, strictly speaking, sophisms. Thus, in the case of Achilles and the tortoise: Achilles, who could run ten yards per second, competed with a tortoise that ran five yards per second; the tortoise was to have a handicap of ten yards. Zeno claimed that Achilles could never catch the tortoise; his argument being based on the "fact" that whenever Achilles reached a point where the tortoise had just been, the tortoise would have moved ahead. Obviously, Zeno did not believe what he claimed; his interest lay in locating the error in his argument. The same observation is true of the three remaining paradoxes of Zeno, the *Dichotomy*, "motion is impossible"; the *Arrow*, "motionless even while in flight"; and the *Stadium*, or "a given time interval is equivalent to an interval twice as long." Beneath the sophistry of these contradictions lie subtle and elusive concepts of limits and infinity, only completely explained in the 19th century when the foundations of analysis became more rigorous and the theory of transfinite numbers had been formulated.

Common algebraic fallacies usually involve a violation of one or another of the following assumptions:

1. If $a = b$, then $a/k = b/k$, provided $k \neq 0$.
2. If $a > b$, then $ka > kb$, provided k is positive.
3. If a is nonnegative, then $\sqrt{a^2} = +a$.

Three examples of such violations follow:

A. Solve: $6x - 18 = 4x - 12$
Factoring: $3(2x - 6) = 2(2x - 6)$
Dividing by $(2x - 6)$: $3 = 2$

B. Since $+1/-1 = -1/+1$, then $\sqrt{+1}/\sqrt{-1} = \sqrt{-1}/\sqrt{+1}$, and so $(\sqrt{+1})(\sqrt{+1}) = (\sqrt{-1})(\sqrt{-1})$, hence $+1 = -1$

C. Given two positive numbers, a and b:

then,	also,
$a > -b$	$b > -a$
$b > -b$	$a > -a$
Multiplying: $ab > b^2$	$ab > a^2$
Or $a > b$	$b > a$

Thus a is both greater than b and less than b.

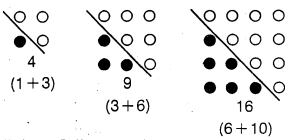
An example of an illegal operation or "lucky boner" is:

$$\tfrac{16}{64} = \tfrac{16}{64} = \tfrac{1}{4}.$$

Polygonal and figurate numbers. Among the many relationships of numbers that have fascinated man are those that suggest (or were derived from) the arrangement of points representing numbers into series of geometrical figures. Such numbers, known as figurate or polygonal numbers, appeared in 15th-century arithmetic books and were probably known to the Chinese as early as 2,000 years ago; but they were of especial interest to the ancient Greek mathematicians. To the Pythagoreans (*c.* 500 BC) numbers were of paramount significance; everything could be explained by numbers, and numbers were invested with specific characteristics and personalities. Among other properties of numbers, the Pythagoreans recognized that numbers had "shapes." Thus the triangular numbers, 1, 3, 6, 10, 15, 21, etc., were visualized as points or dots arranged in the shape of a triangle. Each triangular number is the sum of all successive numbers from 1 on up to some point.

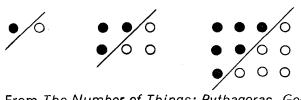
The "shapes" of numbers

Square numbers are the squares of natural numbers, such as 1, 4, 9, 16, 25, etc., and can be represented by squares. Inspection reveals that the sum of any two adjacent triangular numbers is always a square number.

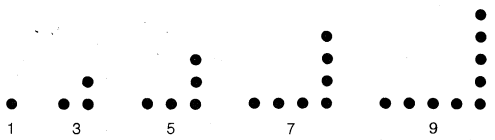


From *The Number of Things: Pythagoras, Geometry and Humming Strings* by Evans G. Valens, copyright © 1964 by Evans G. Valens; published by E.P. Dutton & Co., Inc., and used with their permission

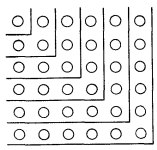
Oblong numbers are the sums of successive even numbers. It is notable that each oblong number is the product of two successive numbers; *e.g.*, $6 = 2 \cdot 3$; $12 = 3 \cdot 4$; $20 = 4 \cdot 5$; etc. The sum of any two equal triangular numbers forms an oblong number (see below).



The gnomons are all the odd numbers, represented by a right angle, or a carpenter's square (see below). Gnomons were useful to the Pythagoreans. They could build



up squares by adding gnomons to smaller squares and from such a figure could deduce many interrelationships: thus $1^2 + 3 = 2^2$, $2^2 + 5 = 3^2$, etc: or $1 + 3 + 5 = 3^2$, $1 + 3 + 5 + 7 = 4^2$, $1 + 3 + 5 + 7 + 9 = 5^2$, etc. (see below). Indeed, it is likely that Pythagoras first



realized the famous relationship $a^2 + b^2 = c^2$ by contemplating the properties of gnomons and square numbers, observing that any odd square can be added to some even square to form a third square. Thus

$$3^2 + 4^2 = 5^2, \text{ where } 3^2 = 4 + 5;$$
$$5^2 + 12^2 = 13^2, \text{ where } 5^2 = 12 + 13,$$

and in general, $a^2 + b^2 = c^2$, where $a^2 = b + c$. This is a special class of Pythagorean triples (see below).

Besides these, the Greeks also studied numbers having pentagonal, hexagonal, and other shapes. Many relationships can be shown to exist between geometric patterns and algebraic expressions.

Polygonal numbers constitute a subdivision of a class of numbers known as figurate numbers. Examples would be the arithmetic sequences

$$1, 2, 3, 4, \ldots r$$
and $$1, 3, 5, 7, \ldots (2r - 1).$$

When new series are formed from the sums of the terms of these series, the results are, respectively,

$$1, 3, 6, 10, \ldots$$
and $$1, 4, 9, 16, \ldots$$

These series are not arithmetic sequences but are seen to be the polygonal triangular and square numbers. Polygonal number series can also be added to form three-dimensional figurate numbers called pyramidal numbers.

The significance of polygonal and figurate numbers lies in their relation to the modern theory of numbers. Even the simple, elementary properties and relations of numbers often demand sophisticated mathematical tools. Thus it has been shown that every integer is either a triangular number, the sum of two triangular numbers, or the sum of three triangular numbers; *e.g.*, $8 = 1 + 1 + 6$; $42 = 6 + 36$; $43 = 15 + 28$; $44 = 6 + 10 + 28$.

Pythagorean triples. The study of Pythagorean triples as well as the general theorem of Pythagoras leads to many unexpected byways in mathematics. A Pythagorean triple is formed by the measures of the sides of an integral right triangle; *i.e.*, any set of three positive integers such that $a^2 + b^2 = c^2$. If a, b, and c are relatively prime—*i.e.*, if no two of them have a common factor—the set is a primitive P. triple.

A formula for generating all P. triples is

$$a = p^2 - q^2, b = 2pq, c = p^2 + q^2,$$

in which p and q are relatively prime, p and q are neither both even nor both odd, and $p > q$. By choosing p and q appropriately, for example, primitive P. triples such as the following are obtained:

p	q	(*a*) $p^2 - q^2$	(*b*) $2pq$	(*c*) $p^2 + q^2$
2	1	3	4	5
3	2	5	12	13

The only primitive triple that consists of consecutive integers is 3, 4, 5.

Characteristic properties of primitive triples

Certain characteristic properties are of interest:

1. Either a or b is divisible by 3.
2. Either a or b is divisible by 4.
3. Either a or b or c is divisible by 5.
4. The product of a, b and c is divisible by 60.
5. One of the following: a, b, $a + b$, $a - b$ is divisible by 7.

It is also true that if n is any integer, then $2n + 1$, $2n^2 + 2n$, and $2n^2 + 2n + 1$ form a P. triple for every value of n.

Some of the properties of P. triples were known to the ancient Greeks; *e.g.*, that the hypotenuse of a primitive triple is always an odd integer. It is now known that for an odd integer R to be the hypotenuse of a primitive triple, a necessary and sufficient condition is that every prime divisor of R be of the type $4k + 1$ (k is any positive integer $\neq 0$).

Perfect numbers and Mersenne numbers. Most numbers are either "abundant" or "deficient." In an abundant number, the sum of its proper divisors (*i.e.*, including 1, but excluding the number itself) is greater than the number; in a deficient number, the sum of its proper divisors is less than the number. A perfect number is an integer that equals the sum of its proper divisors. For example, 24 is abundant, its divisors giving a sum of 36; 32 is deficient, giving a sum of 31. The number 6 is a perfect number, since $1 + 2 + 3 = 6$; so is 28, since $1 + 2 + 4 + 7 + 14 = 28$. The next two perfect numbers are 496 and 8,128. The first four perfect numbers were known to the ancients. Indeed, Euclid suggested that any number of the form $2^{n-1}(2^n - 1)$ is a perfect number whenever $2^n - 1$ is prime, but it was not proved until the 18th century when the Swiss mathematician L. Euler proved that every even perfect number must be of the form $2^{n-1}(2^n - 1)$, where $2^n - 1$ is a prime.

Numbers of the form $2^n - 1$ are called Mersenne numbers after French mathematician Marin Mersenne; they may be prime (*i.e.*, having no factor except itself or 1) or composite (composed of two or more prime factors). A necessary though not sufficient condition that $2^n - 1$ be a prime is that n be a prime. Thus, all even perfect numbers have the form $2^{n-1}(2^n - 1)$ where both n and $2^n - 1$ are prime numbers. Until comparatively recently only 12 perfect numbers were known. In 1876, French mathematician Edouard Lucas found a way to test the primality of Mersenne numbers. By 1952 R.M. Robinson had applied Lucas' test and, by means of electronic digital computers, had found the Mersenne primes for $n = 521$, 607, 1,279, 2,203 and 2,281, thus adding five more perfect numbers to the list of 12 already known. By 1971 there were 23 known perfect numbers.

It is known that to every Mersenne prime there corresponds an even perfect number and vice versa. But two questions are still unanswered: Are there any odd perfect numbers? Are there infinitely many perfect numbers?

Many remarkable properties are revealed by perfect numbers. All perfect numbers, for example, are triangular. Also, the sum of the reciprocals of the divisors of a perfect number (including the reciprocal of the number itself) is always equal to 2. Thus

$$\text{for 6: } \tfrac{1}{1} + \tfrac{1}{2} + \tfrac{1}{3} + \tfrac{1}{6} = 2$$
$$\text{for 28: } \tfrac{1}{1} + \tfrac{1}{2} + \tfrac{1}{4} + \tfrac{1}{7} + \tfrac{1}{14} + \tfrac{1}{28} = 2$$

Fibonacci numbers. This sequence of numbers first became known in 1202 when Leonardo of Pisa, also called

Fibonacci, published *Liber abaci.* The sequence results from a recreational problem: "How many pairs of rabbits can be produced from a single pair in one year if it is assumed that every month each pair begets a new pair which from the second month becomes productive?" Thus

Month:	1	2	3	4	5	6	7	8	9	10	11	12
No. of pairs:	1	1	2	3	5	8	13	21	34	55	89	144

The second row represents the first 12 terms of the Fibonacci sequence, in which each term (except the first two) is found by adding the two terms immediately preceding; in general, $x_n = x_{n-1} + x_{n-2}$, a relation that was not recognized until about 1600.

Over the years, and especially in the middle decades of the 20th century, the properties of the Fibonacci numbers have been extensively studied, resulting in a very considerable literature. Their properties appear to be inexhaustible; for example, $x_{n+1} \cdot x_{n-1} = x_n^2 + (-1)^n$. Another formula for generating the Fibonacci numbers is attributed to Lucas:

$$x_n = \frac{1}{\sqrt{5}}\left\{\left(\frac{1+\sqrt{5}}{2}\right)^n - \left(\frac{1-\sqrt{5}}{2}\right)^n\right\}.$$

Lucas also created and discussed the Lucas series

$$1, 3, 4, 7, 11, 18, \ldots$$

The ratio $(\sqrt{5}+1):2 = 1.618\ldots$, designated as Φ, is known as the golden number; the ratio $(\sqrt{5}-1):2$, the reciprocal of Φ, is equal to $0.618\ldots$. Both these ratios are related to the roots of $x^2 - x - 1 = 0$, an equation derived from the Divine Proportion of the 15th century Italian mathematician Lucas Pacioli, namely, $a/b = b/(a+b)$, when $a < b$, by setting $x = b/a$. In short, dividing a segment into two parts in mean and extreme proportion, so that the smaller part is to the larger part as the larger is to the entire segment, yields the so-called Golden Section, an important concept in both ancient and modern artistic and architectural design. Thus, a rectangle the sides of which are in the approximate ratio of 3:5 ($\Phi^{-1} = 0.618\ldots$), or 8:5 ($\Phi = 1.618\ldots$), is presumed to have the most pleasing proportions, aesthetically speaking.

The Golden Section

$$\Phi = (\sqrt{5}+1)/2 \qquad \Phi^4 = (3\sqrt{5}+7)/2$$
$$\Phi^2 = (\sqrt{5}+3)/2 \qquad \Phi^5 = (5\sqrt{5}+11)/2$$
$$\Phi^3 = (2\sqrt{5}+4)/2 \qquad \Phi^6 = (8\sqrt{5}+18)/2$$

The above exhibits one of many interrelations between the Fibonacci and the Lucas numbers; the successive coefficients of the radical $\sqrt{5}$ are Fibonacci's 1, 1, 2, 3, 5, 8, while the successive second terms within the parentheses are Lucas' 1, 3, 4, 7, 11, 18.

If a golden rectangle ABCD is drawn and a square ABEF is removed, the remaining rectangle ECDF is also a golden rectangle. If this process is continued and circular arcs are drawn, the curve formed approximates the

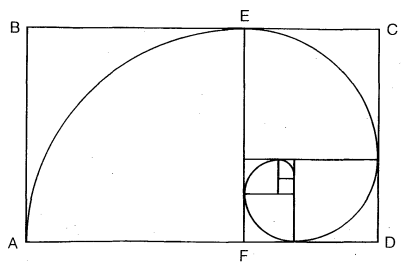


Figure 1: The golden rectangle.

logarithmic spiral, a form found in nature. The logarithmic spiral is the graph of the equation $r = k^\theta$, in polar coordinates, where $k = \Phi^{2/\pi}$. The Fibonacci numbers are also exemplified by the botanical phenomenon known as phyllotaxis. Thus, the arrangement of the whorls on a pinecone or pineapple, of petals on a sunflower, and of branches from some stems follows a sequence of Fibonacci numbers or the series of fractions.

$$\tfrac{1}{1}, \tfrac{1}{2}, \tfrac{2}{3}, \tfrac{3}{5}, \tfrac{5}{8}, \tfrac{8}{13}, \text{ etc.}$$

These relationships form the basis of the theory of dynamic symmetry that has been applied to the fine arts as well as to living forms. (For an excellent treatment of the Fibonacci numbers, see V.E. Hoggatt, *Fibonacci and Lucas Numbers*, 1969.)

Magic squares. A magic square is a square formed by numbers or letters in particular arrangements that were once thought to have magical properties. In arithmetical magic squares the numbers are arranged so that the sum of every column, of every row, and of the two main diagonals is the same.

A standard magic square of any given number contains the sequence of natural numbers from 1 to the square of that number. Thus the magic square of three contains the numbers 1 to 9 in three rows: 4, 9, 2; 3, 5, 7; and 8, 1, 6. The earliest magic square known, it can be traced back to about 2000 BC in China; it does not appear elsewhere until centuries later. For examples and discussion of magic squares composed of letters and words see BOARD AND TILE GAMES; for the mathematical theory of arithmetical magic squares, or Latin squares, see COMBINATORICS AND COMBINATORIAL GEOMETRY.

GEOMETRIC AND TOPOLOGICAL RECREATIONS

Optical illusions. The creation and analysis of optical illusions may involve mathematical concepts and geometric principles such as the areas of similar figures being proportional to the squares of their linear dimensions. Some involve physiological or psychological considerations, such as the fact that when making visual comparisons, relative lengths are normally more accurately perceived than relative areas.

For treatment of types of optical illusions and the sources of their illusory effects, including unorthodox use of perspective, distorted angles, deceptive shading, unusual juxtaposition, equivocal contours or contrasts, colour effects, chromatic aberration, and after images, see the article ILLUSIONS AND HALLUCINATIONS.

Geometric fallacies and paradoxes. Some of the more widely exhibited geometric fallacies include "proofs": (1) that every triangle is isosceles (*i.e.*, has two equal sides); (2) that every angle is a right angle; (3) that if ABCD is a quadrilateral in which $\overline{AB} = \overline{CD}$, then AD must be parallel to BC; and (4) that every point in the interior of a circle lies on the circle.

The explanations of fallacious proofs in geometry usually include one or another of the following: faulty construction; violation of a logical principle, such as assuming the truth of a converse, or confusing partial inverses or converses; misinterpretation of a definition, or failing to take note of "necessary and sufficient" conditions; too great dependence upon diagrams and intuition; being trapped by limiting processes and intuition—*e.g.*, in concluding that in Figure 2, as more and more "teeth" are

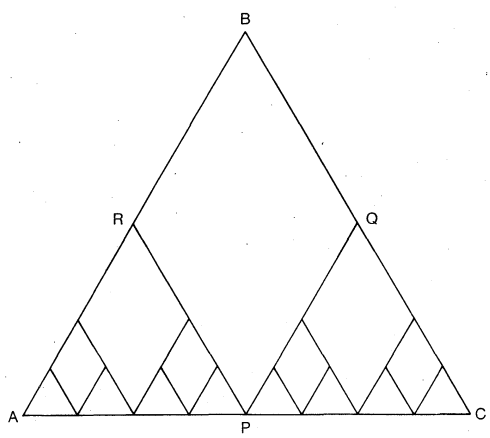


Figure 2: The constant zigzag (see text).

taken, the sawtooth zigzag from A to C approaches more and more nearly the length of AC, whereas actually, if P, Q, R are midpoints of the sides of an equilateral triangle ABC, it always equals $2 \times \overline{AC}$.

Some geometric curves, such as the so-called snowflake curve, a closed curve of infinite length but having a finite interior, furnish genuine paradoxes rather than fallacies. More unbelievable still (but true) is the space-filling curve; in seeming defiance that a curve is "one-dimensional" and thus cannot fill a given space, it can be shown that the curve suggested in various stages in Figure 3, when completed, will ultimately pass through *every* point in the square. In fact, by similar reasoning, the curve can be made to fill completely an entire cube.

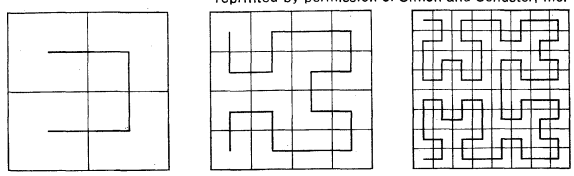


Figure 3: The space-filling curve (see text).

Tangrams; mazes. One of the oldest of Oriental amusements, a tangram is a set of geometric shapes, called *tans*, usually of wood or tile, that can be arranged to represent all sorts of objects, animals, people, and so on. The seven-piece set cut from a square is of Chinese origin; a Greco-Roman version, the loculus of Archimedes, or the stomachion, consisted of 14 pieces cut from a rectangle of which the length is twice its width.

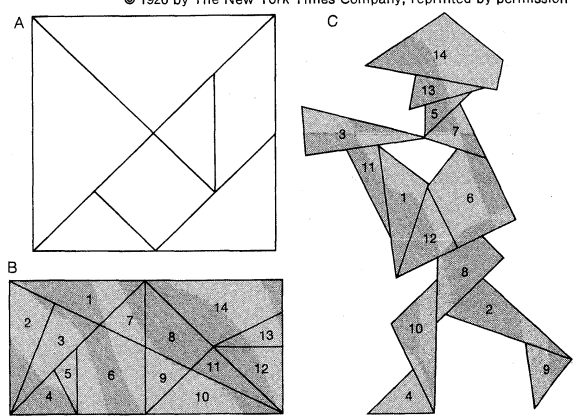


Figure 4: (A) Seven-piece tangram. (B) Fourteen-piece tangram. (C) Pied Piper made from B.

Whereas a tangram involves no mathematical principles, a maze presents a topological situation; the map of a maze is a topological invariant. A maze having only one entrance and one exit can be solved by placing one hand against either wall and keeping it there as it is traversed; the exit can always be reached in this manner, although not necessarily by the shortest path. If the goal is within the labyrinth, the "hand-on-wall" method will also succeed, provided that there is no closed circuit; *i.e.*, a route that admits of complete traverse back to the beginning (Figure 5).

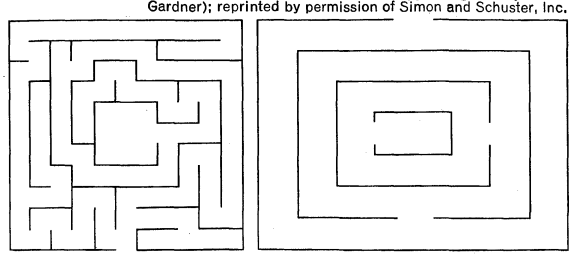


Figure 5: *Examples of mazes.*
(Left) "Simply connected" maze.
(Right) "Multiply connected" maze (see text).

If there are no closed circuits—*i.e.*, no detached walls—the maze is "simply connected"; otherwise the maze is "multiply connected." A classic general method of "threading a maze" is to designate a place where there is a choice of turning as a node; a path or node that has not yet been entered as a "new" path or node; and one that has already been entered as an "old" path or node.

The procedure is as follows:

1. Never traverse a path more than twice.
2. When arriving at a new node, select either path.
3. When arriving at an old node or at a dead end by a new path, return by the same path.
4. When arriving at an old node by an old path, select a new path, if possible; otherwise, an old path.

Ironically, while recreational interest in mazes has diminished, two areas of modern science have found them to be of value: psychology and communications technology. The former is concerned with the learning behaviour of men and animals, the latter with improved design of computers.

Geometric dissections. Geometric dissection problems involve the cutting of geometric figures into pieces that can be arranged to form other geometric figures; for example, cutting a rectangle into parts that can be put together in the form of a square and vice versa. Interest in this area of mathematical recreations began to manifest itself towards the close of the 18th century when Montucla called attention to this problem. As the subject became more popular, greater emphasis was given to the more general problem of dissecting a given polygon of any number of sides into parts that would form another polygon of equal area. Then, in the early 20th century, interest shifted to finding the *minimum* number of pieces required to change one figure into another.

Theory of equidecomposable figures

According to a comprehensive theory of equidecomposable figures outlined about 1960, two polygons are said to be equidecomposable if it is possible to dissect, or decompose, one of them into a finite number of pieces that can then be rearranged to form the second polygon. Obviously, the two polygons have equal areas.

According to the converse theorem, if two polygons have equal areas, they are equidecomposable.

In the method of complementation, congruent parts are added to two figures so as to make the two new figures congruent. It is known that equicomplementable figures have equal areas, and that if two polygons have equal areas, they are equicomplementable. As the theory advanced, the relation of equidecomposability to various motions such as translations, central symmetry, and, indeed, to groups of motions in general, was explored. Studies were also extended to the more difficult questions of dissecting polyhedra.

On the "practical" side, the execution of a dissection, such as converting the Greek cross or the tau cross into a square (Figure 6) requires the use of ingenious procedures such as the parallelogram slide, the quadrilateral slide, rational and step dissections, strip dissections, and tessellations, as described by H. Lindgren (see *Bibliography*).

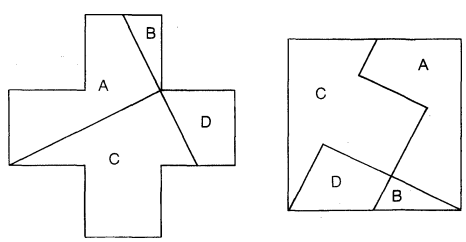


Figure 6: Greek cross converted by dissection into a square.

A quite different and distinctly modern type of dissection deserves brief mention, the so-called squaring the square, or squared rectangles. Thus, the problem of subdividing a square into smaller squares, no two of which are alike, long thought to be unsolvable, has been solved

by means of network theory. In this connection, a squared rectangle is a rectangle that can be dissected into a finite number of squares; if no two of these squares are equal, the squared rectangle is said to be perfect. The order of a squared rectangle is the number of constituent squares. It is known that there are no perfect rectangles of orders less than 9, and that there are exactly two perfect rectangles of order 9. The dissection of a rectangle or a square into unequal squares has been a comparatively recent development; that is, since about 1940.

From Martin Gardner, *The Second Scientific American Book of Mathematical Puzzles and Diversions* (copyright © 1961 by Martin Gardner); reprinted by permission of Simon and Schuster, Inc.

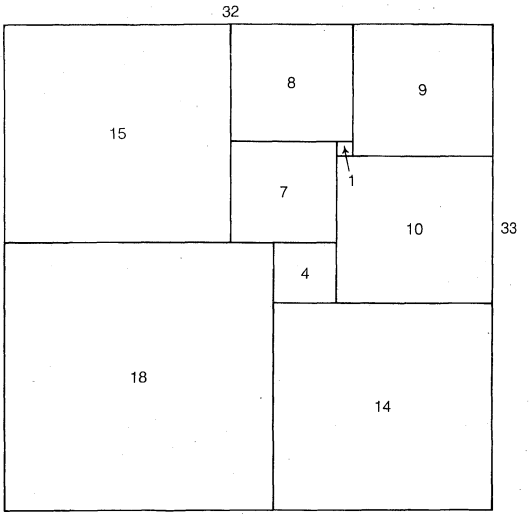


Figure 7: Squared rectangle (see text).

Graphs and networks. The word "graph" may refer to the familiar curves of analytic geometry and function theory, or it may refer to simple geometric figures consisting of points and lines connecting some of these points; the latter are sometimes called linear graphs, although there is little confusion within a given context. Such graphs have long been associated with puzzles.

If a finite number of points are connected by lines (Figure 8A), the resulting figure is a graph; the points, or corners, are called the vertices, and the lines are called the edges. If every pair of vertices is connected by an edge, the graph is called a complete graph (Figure 8B). A

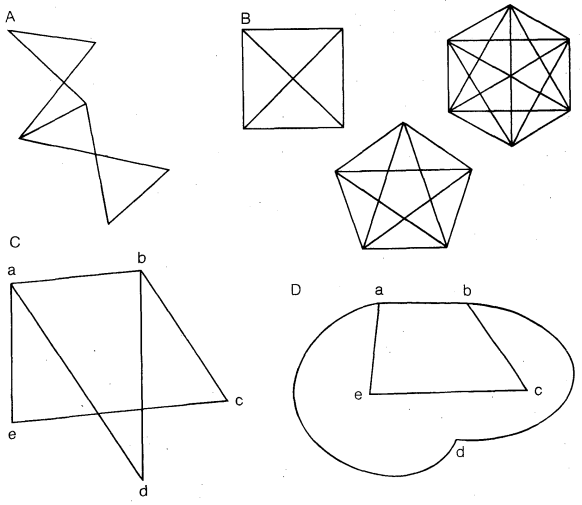


Figure 8: *Examples of linear graphs.*
(A) Graph. (B) Complete graphs. (C) Nonplanar graph.
(D) Nonplanar graph transformed into equivalent planar graph.

planar graph is one in which the edges have no intersection or common points except at the edges. (It should be noted that the edges of a graph need not be straight lines.) Thus a nonplanar graph can be transformed into an equivalent, or isomorphic, planar graph, as in Figures 8C and 8D. An interesting puzzle involves the problem of the three wells. Here (Figure 9) A, B, and C represent three neighbours' houses, and R, S, and T three wells. It is desired to have paths leading from each house to each well, allowing no path to cross any other path. The proof that the problem is impossible depends on the so-called Jordan Curve theorem that a continuous closed curve in a plane divides the plane into an interior and an exterior region in such a way that any continuous line connecting a point in the interior with a point in the exterior must intersect the curve. Planar graphs have proved useful in the design of electrical networks.

Problem of the three wells

From *Graphs and Their Uses*, by Oystein Ore. Copyright © 1963 by Yale University. Reprinted by permission of Random House, Inc.

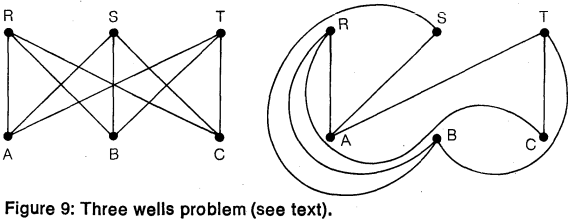


Figure 9: Three wells problem (see text).

A connected graph is one in which every vertex, or point (or, in the case of a solid, a corner), is connected to every other point by an arc; an arc denotes an unbroken succession of edges. A route that never passes over an edge more than once, although it may pass through a point any number of times, is called a path.

Modern graph theory (in the sense of linear graphs) had its inception with the work of Euler in connection with the Königsberg bridge problem, and was, for many years, associated with "unicursal curves"; *i.e.*, figures that can be drawn with one stroke of the pencil. The city of Königsberg (now Kaliningrad) embraces the banks and two islands of the river Pregel; seven bridges connect the islands with the mainland. The problem was: Could a person leave home, take a walk, and return, crossing each bridge just once? Euler showed why it is impossible.

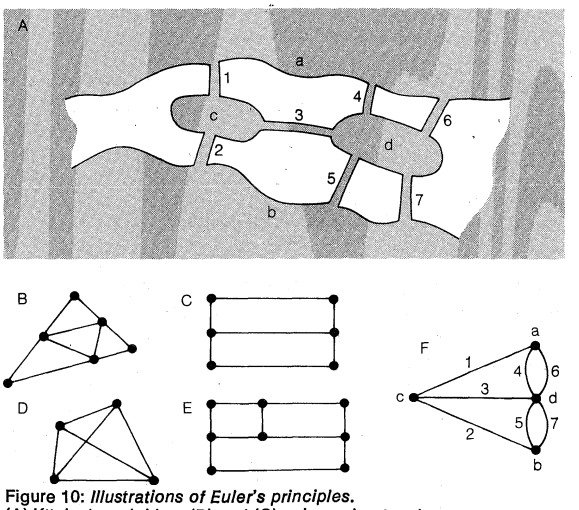


Figure 10: *Illustrations of Euler's principles.*
(A) Königsberg bridge. (B) and (C) unicursal networks.
(D) and (E) noncursal networks. (F) network corresponding to Königsberg bridge problem.

Briefly stated, Euler's principles (that apply for any closed network) are as follows:

1. The number of even points—*i.e.*, those in which an even number of edges meet—is of no significance.

2. The number of odd points is always even; this includes the case of a network with only even points.

3. If there are *no odd points*, the network is unicursal; one can start at any point and finish at the same point.

4. If there are *exactly two odd points*, the network is unicursal; one can start at either of the odd points and finish at the other odd point.

5. If there are *more than two odd points*, the network cannot be traced in one continuous path; if there

are $2n$ odd points and no more, it can be traced in n separate paths.

Thus Figure 10B and Figure 10C are unicursal; Figures 10D and 10E are not; Figure 10F shows a network corresponding to the K. bridge problem, in which the points represent the land areas and the edges the seven bridges.

Networks are related to a variety of recreational problems that involve combining or arranging points in a plane or in space. Among the earliest was a puzzle invented by an Irish mathematician, Sir William Rowan Hamilton (1859), which required finding a route along the edges of a regular dodecahedron that would pass once and only once through every point. In another version, the puzzle was made more convenient by replacing the dodecahedron by a graph isomorphic to the graph formed by the 30 edges of the dodecahedron (Figure 11). A Hamilton circuit is one that passes through each point exactly once but does not, in general, cover all the edges; actually, it covers only two edges at each vertex. The route shown in heavy lines is a Hamilton circuit.

The Hamilton circuit

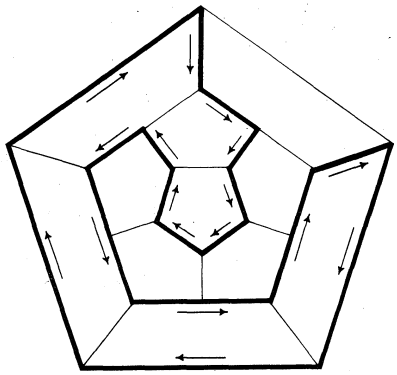


Figure 11: Hamilton circuit.
From Martin Gardner, *The Scientific American Book of Mathematical Puzzles and Diversions* (copyright © 1959 by Martin Gardner); reprinted by permission of Simon and Schuster, Inc.

Graph theory, being a branch of mathematics known as combinatorial topology, lends itself to a variety of problems involving combinatorics: for example, designing a network to connect a set of cities by railroads or by telephone lines; planning city streets or traffic patterns; matching jobs with applicants; arranging round-robin tournaments such that every team or individual meets every other team or individual.

Map colouring problems. Although geographers have long known that maps depicting subdivisions of areas can be coloured in such a way that any two subdivisions having a common boundary show different colours and no more than four distinct colours are used, the celebrated "four-colour map problem" bears little or no relation, historically, to cartography. A political map, in fact, is rather different. The mathematical question was originally framed in 1850 and publicized in 1878. Essentially, the problem is: How many colours are needed to colour any map so that no two regions sharing a common border (edge) will have the same colour? Are four colours both necessary and sufficient? In short, is it possible to construct a map for which five colours are necessary? No one knows. It seems likely that four colours are necessary and sufficient; this has never been proved. It is still a conjecture, not a theorem. That is not to say that the problem has been neglected; the literature is extensive.

The four-colour problem

Some of the salient results may be stated briefly. To begin with, the surface of a sphere is topologically equivalent to a plane. The following basic theorems have been established:

1. Any map on the plane can be coloured with two colours if, and only if, all of its points are even.

2. If the surface of a sphere is divided into regions, each of which has an even number of edges, and if three countries meet at each point, then the map can be coloured with exactly three colours.

3. If the surface of a sphere is divided into regions, each of which shares its boundaries with three neighbouring regions, then the map can be coloured with three or fewer colours.

4. A map can always be coloured in five or fewer colours.

5. Any map with fewer than 38 regions can be coloured with four or fewer colours.

But the general four-colour theorem still remains unproved.

Flexagons. A flexagon is a polygon constructed from a strip of paper or thin metal foil in such a way that the figure possesses the property of changing its faces when it is "flexed." First discussed in 1939, flexagons have become a fascinating mathematical recreation. One of the simplest flexagons is the trihexaflexagon, made by cutting a strip of suitable material and marking off 10 equilateral triangles. By folding appropriately several times and then gluing the last triangle onto the reverse side of the first triangle, the resulting model may be "flexed" so that one of the faces "disappears" and another face takes its place.

As mathematicians experimented with flexagons, they came up with hexaflexagons (19 triangles), tetraflexagons, hexahexaflexagons, tetrahexaflexagons, pentahexaflexagons and even decahexaflexagons, the latter having 10 faces and 82 variations.

MANIPULATIVE RECREATIONS

Puzzles involving configurations. One of the earliest puzzles and games that require arranging counters into some specified alignment or configuration was Lucas' Puzzle: three black counters and three white counters are placed in a row, leaving a blank space between them. The object is to move the black pieces from the left end to the right end and the white to the left end; black pieces can be moved only to the right and the white only to the left. A counter may move to an adjacent square or it may jump over one of the opposite colours. Variations of the puzzle use nine squares with four counters of one kind and four of another, and so on. For n counters of each kind the number of required moves is $n(n + 2)$.

Lucas' puzzle and Ticktacktoe

A similar puzzle uses eight numbered counters placed on nine positions. The aim is to shift the counters so that they will appear in reverse numerical order; only single moves and jumps are permitted.

Well known, but by no means as trivial, are games for two players, such as Ticktacktoe and its more sophisticated variations, one of which calls for each player to begin with three counters (3 black, 3 white); the first player places a counter in any cell, except the center cell, of a 3 $\times$ 3 diagram; the players then alternate until all the counters are down. If neither has won by getting three in a row, each, in turn, is permitted to move a counter to an adjacent square, moving only horizontally or vertically. Achieving three in a row constitutes a win. There are many variations. The game can be played on a 4 $\times$ 4 diagram, each player starting with four counters; sometimes diagonal moves are permitted. Another version is played on a 5 $\times$ 5 pattern.

Another interesting modification, popular in Europe, is variously known as Mill or Nine Men's Morris, played with counters on a board consisting of three concentric squares and eight transversals (for details see BOARD AND TILE GAMES).

Another game of this sort is played on a diamond-shaped board comprised of tessellated hexagons, usually 11 on each edge, where by "tessellated" we mean fitted together like tiles to cover the board completely. Two opposite edges of the diamond are designated "white"; the other two sides, "black." Each player has a supply of black or white counters. The players alternately place a piece on any vacant hexagon; the object of the game is for each player to complete an unbroken chain of his pieces between the sides designating his colour. Though the game does not end until one of the players has made a complete chain, it may meander across the board; it cannot end in a draw because the only way one player can block the other is by completing his own chain. The game was created by Piet Hein in 1942 in Denmark,

where it quickly became popular under the name of Polygon. It was invented independently in the United States in 1948 by John Nash, and a few years later one version was marketed under the name of Hex.

In addition to the aforementioned varieties of a class of games that can be loosely described as "three in a row" or "specified alignment" many others also exist, such as three- and four-dimensional Ticktacktoe and even a computer Ticktacktoe. The game strategy in Ticktacktoe is by no means simple; an excellent mathematical analysis is given in F. Schuh (see *Bibliography*).

Chessboard problems. Recreational problems posed with regard to the conventional chessboard are legion. Among the most widely discussed is the problem of how to place eight queens on a chessboard in such a way that none of the queens is attacking any other queen; the problem interested the great German mathematician C.F. Gauss (*c.* 1850). Another group of problems has to do with the knight's tour; in particular, to find a closed knight's tour that ends at the starting point, that does not enter any square more than once, but that passes through all the squares in one tour. Problems of the knight's tour are intimately connected with the construction of magic squares. Other chessboard problems are concerned with determining the relative values of the various chess pieces; finding the maximum number of pieces of any one type that can be put on a board so that no one piece can take any other; finding the minimum number of pieces of any one type that can be put on a board so as to command all cells; and how to place 16 queens on a board so that no three of them are in a straight line.

The Fifteen Puzzle. This is perhaps one of the best known of all puzzles. Invented by one Sam Loyd about 1878, it is also known as the Boss Puzzle, Jeu de Taquin, and Diablotin. It became popular all over Europe almost at once and literally created a rage. It consists essentially of a shallow square tray that holds exactly 15 small square counters numbered from 1 to 15, and one square blank space. With the 15 squares initially placed in random order and with the blank space in the lower right-hand corner, the puzzle is to rearrange them in numerical order by sliding only, with the blank space ending up back in the lower right-hand corner. It may overwhelm the reader to learn that there are more than 20,000,000,000,000 possible different positions that the pieces (including the blank space) can assume. But in 1879 two American mathematicians proved that only one-half of all possible initial arrangements, or about 10,000,000,000,000, admitted of a solution. The mathematical analysis is as follows. Basically, no matter what path it takes, as long as it ends its journey in the lower right-hand corner of the tray, any numeral must pass through an even number of boxes. In the normal position of the squares, regarded row by row from left to right, each number is larger than all the preceding numbers; *i.e.*, no number *precedes* any number smaller than itself. In any other than the normal arrangement, one or more

Solving the Fifteen Puzzle

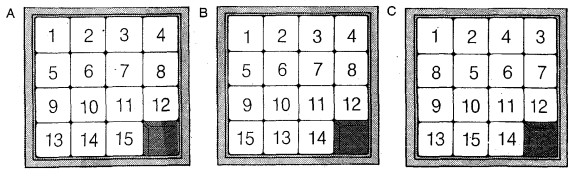


Figure 12: (A) Fifteen Puzzle with no inversions. (B) With two inversions. (C) With five inversions.

numbers will precede others smaller than themselves. Every such instance is called an inversion. For example, in the sequence 9, 5, 3, 4, the 9 precedes three numbers smaller than itself and the 5 precedes two numbers smaller than itself, making a total of five inversions. If the total number of *all* the inversions in a given arrangement is *even*, the puzzle can be solved by bringing the squares back to the normal arrangement; if the total number of inversions is *odd*, the puzzle cannot be solved. Thus, in Figure 12B there are two inversions, and the puzzle can be solved; in Figure 12C there are five inversions, and the puzzle has no solution. Theoretically, the puzzle can be extended to a tray of $m \times n$ spaces with $(mn - 1)$ numbered counters. In recent years, a "31" puzzle, as well as others, has been obtainable.

The Tower of Hanoi. This puzzle is believed to have been put out in 1883 by Lucas, under the name of M. Claus. Ever popular, made of wood or plastic, it still can be found in toy shops. It consists essentially of three pegs fastened to a stand and of eight circular disks, each having a hole in the centre. The disks, all of different radii, are initially placed on one of the pegs, with the largest disk on the bottom and, in order of decreasing radii, the smallest on top. The task is to transfer the individual disk from one peg to another so that no disk ever rests on one

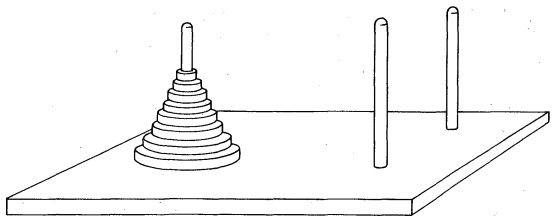


Figure 13: Tower of Hanoi.

smaller than itself, and, finally, to transfer the tower; *i.e.*, all the disks in their proper order, from their original peg to one of the other pegs. It can be shown that for a tower of n disks, there will be required $2^n - 1$ transfers of individual disks to shift the tower completely to another peg. Thus for 8 disks, the puzzle requires $2^8 - 1$, or 255 transfers. In a more imaginative version, the original "needle" (peg) was a tower with 64 disks, requiring $2^{64} - 1$, or 18,446,744,073,709,551,615 transfers; this is exactly the same number required to fill an 8×8 checkerboard with grains of wheat, 1 on the first square, 2 on the second, 4 on the next, then 8, 16, 32, etc. With no interruptions and no mistakes, it would take many thousands of millions of years to transfer the tower.

Polyominoes. The term designating this popular and relatively recent mathematical recreation was introduced in 1953. A polyomino is a simply connected set of equal-sized squares; *i.e.*, shapes consisting of congruent squares, each joined together with at least one other square along an edge. The simpler polyomino shapes are shown in Figure 14A. Somewhat more fascinating are the pentominoes, of which there are exactly 12 forms (Figure 14B). Asymmetrical pieces, which have "different shapes" when they are flipped over, are counted as one.

Manipulating squares and cubes

From Martin Gardner, *The Scientific American Book of Mathematical Puzzles and Diversions* (copyright © 1959 by Martin Gardner); reprinted by permission of Simon and Schuster, Inc.

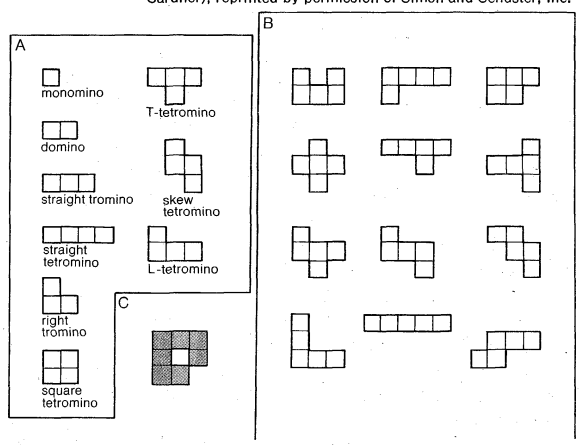


Figure 14: *Shapes made of squares.* (A) Monomino and simple polyominoes. (B) Pentominoes. (C) Heptomino with interior "hole."

The number of distinct polyominoes of any order is a function of the number of squares in each, but, as yet, no general formula has been found. It has been shown that there are 35 types of hexominoes, and 108 types of hep-

tominoes, if the dubious heptomino with an interior "hole" is included.

Recreations with polyominoes include a wide variety of problems in combinatorial geometry, such as forming desired shapes and specified designs; covering a chessboard with polyominoes in accordance with proscribed conditions; etc. Two illustrations may suffice.

The 35 hexominoes, having a total area of 210 squares, would seem to admit of arrangement into a rectangle 3×70, 5×42, 6×35, 7×30, 10×21, or 14×15; however, no such rectangle can be formed.

Another interesting problem is found in the question: Can the 12 pentominoes, together with one square tetromino, form an 8×8 checkerboard? A solution of the problem was shown around 1935. It is not known how many solutions there are, but it has been estimated to be at least 1,000. In 1958, by use of a computer, it was shown that there were 65 solutions in which the square tetromino is exactly in the centre of the checkerboard.

Soma Cubes. These were created by Piet Hein of Denmark, also known for his invention of the mathematical games known as Hex and Tac Tix. Hein stumbled upon the fact that all the irregular shapes that can be formed by combining no more than four congruent cubes joined at their faces can be put together to form a larger cube. There are exactly seven such shapes, as shown in Figure 15. No two shapes are alike, although the last two are mirror images of each other. The fact that these seven pieces (comprising 27 "unit" cubes) can be reassembled to form one large cube is indeed remarkable.

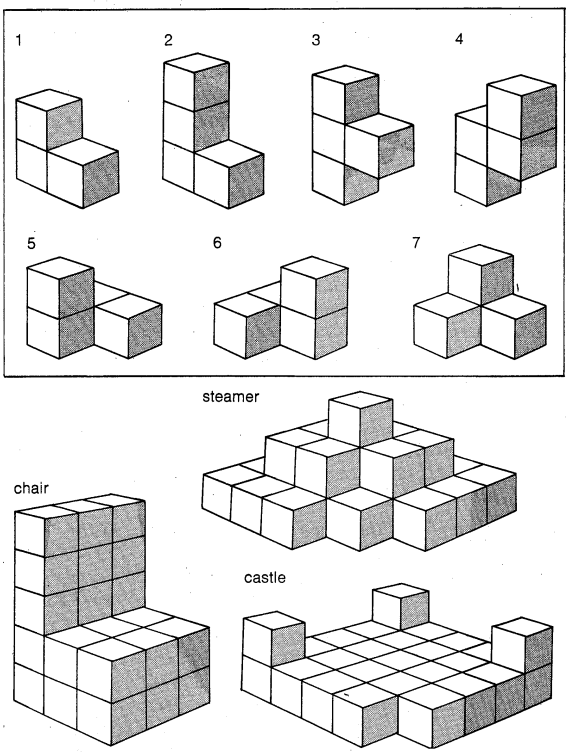


Figure 15: *Soma Cubes.*
(Top) The seven basic pieces. (Bottom) Examples of some of the shapes that can be built from Soma pieces.

Many interesting solid shapes can be formed from the seven Soma Cubes, shapes resembling, for example, a sofa, a chair, a castle, a tunnel, a pyramid, and so on. Even the reassembling of the seven basic pieces into a large cube admits of more than 230 essentially different solutions.

As a recreation, the Soma Cubes are fascinating. With experience, people often find that they can solve Soma problems mentally. Psychologists who have used them find that the ability to solve Soma problems is roughly correlated with general intelligence, although there are some strange anomalies at both ends of the distribution of intelligence. In any event, people playing with the cubes do not appear to want to stop; the variety of interesting structures possible seems endless.

Coloured squares and cubes. There is a wide variety of puzzles involving coloured square tiles and coloured cubes. In one, the object is to arrange the 24 three-colour patterns, including repetitions, that can be obtained by subdividing square tiles diagonally, using three different colours, into a 4×6 rectangle so that each pair of touching edges is the same colour and the entire border of the rectangle is the same colour.

More widely known perhaps is the 30 Coloured Cubes Puzzle. If six colours are used to paint the faces there result 2,226 different combinations. If from this total only those cubes that bear all six colours on their faces are selected, a set of 30 different cubes is obtained; two cubes are regarded as "different" if they cannot be placed side by side so that all corresponding faces match. Many fascinating puzzles arise from these coloured squares and cubes; many more could be devised. Some of them have appeared commercially at various times under different names, such as the Mayblox Puzzle, the Tantalizer, and the Katzenjammer. Stimulating discussions may be found in British combinatorial analysis authority Maj. P.A. MacMahon's original monograph *New Mathematical Pastimes* (1921); in F. Winter's *Spiel der 30 Bunten Würfel* (1934), and in T.H. O'Beirne's *Puzzles and Paradoxes* (1965), pp. 112–129.

A revival of interest in coloured-cube problems has been aroused by the appearance of a puzzle currently known as Instant Insanity consisting of four multicoloured unit cubes, each of which has its faces painted white, red, green, and blue, in a definite scheme. The puzzle is to assemble the cubes into a $1 \times 1 \times 4$ prism such that all four colours appear on each of the four long faces of the prism. Since each cube admits of 24 different orientations, there are 82,944 possible prismatic arrangements; of these only two are the required solutions.

Nim; Wythoff's Game; Tac Tix. A game so old that its origin is obscure, Nim lends itself nicely to mathematical analysis. In its generalized form, any number of objects (counters) are divided arbitrarily into several piles. Two people play alternately; each, in turn, selects any one of the piles and removes from it all the objects, or as many as he chooses, but at least one object. The player removing the last object wins. Every combination of the objects may be considered "safe" or "unsafe"; *i.e.*, if the position left by a player after his move assures a win for that player, the position is called safe. Every unsafe position can be made safe by an appropriate move, but every safe position is made unsafe by any move. To determine whether a position is safe or unsafe, the number of objects in each pile may be expressed in binary notation. If each column adds up to zero or an even number, the position is safe. For example, if at some stage of the game, the number of objects in each pile respectively is 4, 9, 15, in binary notation:

"Safe" and "unsafe" positions

$$\begin{array}{rr} 4 \rightarrow & 1\ 0\ 0 \\ 9 \rightarrow & 1\ 0\ 0\ 1 \\ 15 \rightarrow & 1\ 1\ 1\ 1 \\ \hline & 2\ 2\ 1\ 2 \end{array}$$

Since the second column from the right adds up to 1, an odd number, the given combination is unsafe. A skillful player will always move so that every unsafe position left to him is changed to a safe position.

A similar game is played with just two piles; in each draw the player may take objects from either pile or from both piles, but in the latter event he must take the same number from each pile. The player taking the last counter is the winner.

Games such as Nim make considerable demands upon the player's ability to "translate" decimal numbers into binary numbers and vice versa. Since digital computers operate on the binary system, however, it is possible to program a computer (or build a special machine) that will play a "perfect" game. This was done by E.U. Condon and an associate in 1940 in the invention of the

automatic Nimatron that was exhibited at the New York World's Fair. In still another version, developed by Piet Hein, the objects are arranged in a square formation.

Games of this sort seem to be widely played the world over. The game of Pebbles, also known as the game of Odds, is played by two people who start with an odd number of pebbles placed in a pile. Taking turns, each player draws one, or two, or three pebbles from the pile. When all the pebbles have been drawn, the player who has an odd number of them in his possession wins.

Predecessors of these games in which players distribute pebbles, seeds, or other counters into rows of holes under varying rules have been played for centuries in Africa and Asia. These games, known as Mancala games, are covered in the article BOARD AND TILE GAMES.

PROBLEMS OF LOGICAL INFERENCE

Logical puzzles. Under this heading are included challenging questions that, in general, do not involve numerical or geometrical considerations but call for deductive inferences based chiefly on logical relationships. Such puzzles are not to be confounded with riddles, which frequently rely upon deliberately misleading or ambiguous statements, a play on words, or some other device intended to "catch" the unwary. Logical puzzles do not admit of a standard procedure or generalized pattern for their solution and are usually solved by some trial and error method. This is not to say that the guessing is haphazard; on the contrary, the given facts (generally minimal) suggest several hypotheses, some of which may need to be rejected if found inconsistent, until, by substitution and elimination, the solution is finally reached. The use of various techniques of logic may sometimes prove helpful, but in the last analysis, success depends largely upon that elusive capacity called ingenuity. For convenience, logic problems are arbitrarily grouped in the following categories.

Solution of logical problems

The brakeman, the fireman, and the engineer. This puzzle, which dates back to at least the early 1920s, is used as a prototype for the first category of logic problems. In the intervening years it has become quite popular, the more recent versions being considerably more complicated. Three railway employees named Smith, Robinson, and Jones, are brakeman, fireman, and engineer, but not respectively, who live in the state of New York. Three businessmen, similarly named, also live in the state. The businessman Robinson and the brakeman live in Albany, the businessman Jones and the fireman live in Rochester, while the businessman Smith and the engineer live halfway between these two cities. The brakeman's namesake earns $3,500 a year; the engineer earns one-third as much as the businessman living nearest him. The railway man Smith beats the fireman at billiards. What is the engineer's name?

Overlapping groups. This crude designation suggests problems of the following sort. Among the members of a high-school language club, 31 were studying French; 26 were studying French and Spanish; 23, German; 36, Spanish; 11, French and German; 28, German and Spanish; two, French, German, and Spanish. How many members are there in the club, and how many are studying exactly two languages?

Truths and lies. Another favourite kind of logical inference puzzle concerns truths and lies. One variety is as follows: Some of the natives of a certain South Pacific island are pure-blooded and the rest are half-breeds, but they all look alike. The pure-blooded always tell the truth, but the half-breeds always lie. A visitor to the island, meeting three natives, asks them whether they are full-blooded or half-breeds. The first native mutters something inaudible. The second, pointing to the first, says, "He says that he is pure-blooded." The third, pointing to the second, says, "He lies." Knowing beforehand that only one of the natives is a half-breed, the visitor concludes what each of the three is.

In a slightly different type, four men, one of whom was known to have committed a certain crime, made the following statements when questioned by the police:

Archie: Dave did it.
Dave: Tony did it.
Gus: I didn't do it.
Tony: Dave lied when he said I did it.

If only one of these four statements is true, who was the guilty man? On the other hand, if only one of these four statements is false, who was the guilty man? (From 101 *Puzzles in Thought and Logic* by C.R. Wylie, Jr.; Dover Publications, Inc., New York, 1957. Reprinted through the permission of the publisher.)

Difficult crossings. These, as mentioned under *History* above, also are known as ferrying problems. Perhaps one of the oldest is that of the jealous husbands, one version of which was given about 1612 as follows: Three couples wish to cross a river, in a boat that will hold only two persons, in such a way as never to leave a woman in the company of a man unless her husband also is present. The original solution required 11 crossings. With two married couples, five passages are required. With four married couples, the problem has no solution. On the other hand, if the boat holds three persons, the problem can be solved for four couples. The use of graphs (networks) can facilitate the solution of such problems.

The problem of the jealous husbands

Many similar problems appeared in the writings of medieval mathematicians. Alcuin (*c.* 732–804), for example, gives the problem of a man and a wife and two children with a boat that can hold the man *or* his wife *or* the children. Alcuin also gives the problem of the man crossing a stream with a wolf, a goat, and a bundle of cabbages, the boat being so small that he can take only one of them at a time; he cannot leave the wolf alone with the goat nor the goat alone with the cabbages. A slightly different version is that of the man with a fox, a goose, and some corn who must cross a river and can take only one at a time. He cannot leave the goose with the corn to take the fox over and he cannot leave the fox alone with the goose. He can do it in six crossings, on two of which he will be empty-handed. Some wag has suggested that if the goose accompanies him each time—swimming beside the boat—only three crossings will be needed.

Logical paradoxes. Highly amusing and often tantalizing, logical paradoxes generally lead to searching discussions of the foundations of mathematics. As early as the 6th century BC, the Cretan prophet Epimenides allegedly observed that "All Cretans are liars," which, in effect, means that "All statements made by Cretans are false." Since Epimenides was a Cretan, the statement made by him is false, so that all statements made by Cretans are *not* false. Thus the initial statement is self-contradictory. A similar dilemma was given by an English mathematician, P.E.B. Jourdain, in 1913, when he proposed the card paradox. This was a card on one side of which was printed:

"The sentence on the other side of this card is TRUE."

On the other side of the card the sentence read:

"The sentence on the other side of this card is FALSE."

The barber paradox, offered by Bertrand Russell, was of the same sort: The only barber in the village declared that he shaved everyone in the village who did not shave himself. On the face of it, this is a perfectly innocent remark until it is asked "Who shaves the barber?" If he does not shave himself, then he is one of those in the village who does not shave himself and so is shaved by the barber, namely, himself. If he shaves himself, he is, of course, one of the people in the village who is not shaved by the barber. The self-contradiction lies in the fact that a statement is made about "all" the members of a certain class, when the statement or the object to which the statement refers is itself a member of the class. In short, the Russell paradox hinges on the distinction between those classes that are members of themselves and those that are not members of themselves. Russell attempted to resolve the paradox of the class of all classes by introducing the concept of a hierarchy of logical types but without much success. Indeed, the entire problem lies close to the philosophical foundations of mathematics.

The barber paradox

Miscellaneous problems. A considerable number of problem types do not lend themselves to ready classification. Some of them involve numerical considerations; all of them require some "inventive" use of logical faculty. Thus the classic problem of determining the least number of weights that would make possible the weighing of any number of pounds from one to 40 inclusive was first discussed by a mathematician who gave two solutions: (1) the series of weights 1, 2, 4, 8, 16, 32, and (2) the series 1, 3, 9, 27, depending on whether the weights may be placed only in one scalepan or in either scalepan. The problem was generalized in 1886, when it was shown that when any weight may be placed in either scalepan, and it is required (a) that no other weighings are possible and (b) that each weighing is to be possible in only one way, then there are eight possible sets of weights with which any number of pounds (one to 40 inclusive) may be accomplished. In seven of these solutions, some of the weights are equal (*e.g.*, 1, 3, 9, 9, 9, 9), but an earlier solution (1, 3, 9, 27) is one of the eight possible solutions, and is not only the one with the least number of weights but also the only one in which all the weights are different.

Weighing problems

Another type of recreation consists of detecting a false coin in a collection of otherwise identical coins by repeated weighings on a balance, with or without weights. The classic problem is a twelve-coin problem, which permits three weighings for 12 coins, where the false coin may be overweight or deficient in weight. The problem dates from about 1945; since then, various versions have become more sophisticated, and an extensive literature has been accumulated.

So-called decanting problems, or problems of partitioning liquids in vessels of specified capacity, made their appearance in 1484; they are still hardy perennials among mathematical recreations. In one version it is required to pour four quarts of wine from a full jug with a capacity of eight quarts, using only two empty jugs of five and three quarts capacity, respectively. Another is to partition 24 quarts into three equal parts, using only vessels with capacities of 5, 11, 13, and 24 quarts respectively. One solution of the first problem is as follows: Starting with the eight-quart jug full and the five-quart and three-quart jugs empty (8, 0, 0), the five-quart jug is filled from the eight-quart jug, leaving in it three quarts and leaving the three-quart jug still empty (3, 5, 0). Then the three-quart jug is filled from the five-quart jug giving 3, 2, 3; the three-quart jug is emptied into the eight-quart jug giving 6, 2, 0; and the two quarts in the five-quart jug are poured into the three-quart jug for 6, 0, 2. Next the five-quart jug is filled from the six quarts in the eight-quart jug giving 1, 5, 2; one quart from the five-quart jug is added to the two already in the three-quart jug giving 1, 4, 3; and, finally, the contents of the three-quart jug are added to the one quart in the eight-quart jug to solve the problem, 4, 4, 0.

Decanting problems

Other miscellaneous recreations involving imagination, visualization, or intuition more than mathematical ideas might include the jar of bacteria, the converging cyclists and the three smudged foreheads. In the first of these, a jar contains a single bacterium, which reproduces by simple division to produce two offspring every minute, at which rate it will fill the jar in exactly one hour. How full will the jar be after 59 minutes?

In the second instance, two cyclists are pedalling toward each other along a straight road at the rate of 12 miles an hour. When they are six miles apart, a dragonfly alights on one bicycle and forthwith flies toward the other, shuttling back and forth at the rate of 20 miles an hour until the two cyclists meet. How far did the dragonfly travel during that time?

The problem of the smudged faces is once more an instance of pure logical deduction. Three travellers were aboard a train that had just emerged from a tunnel, leaving a smudge of soot on the forehead of each. While they were laughing at each other, and before they could look into a mirror, a neighbouring passenger suggested that although no one of the three knew whether he himself was smudged, there was a way of finding out without using a mirror. He suggested: "Each of the three of you look at the other two; if you see at least one whose forehead is smudged, raise your hand." Each raised his hand at once. "Now," said the neighbour, "as soon as one of you knows for sure whether his own forehead is smudged or not, he should drop his hand, but not before." After a moment or two, one of the men dropped his hand with a smile of satisfaction, saying: "I know." How did that man know that his forehead was smudged?

A final example might be the paradox of the unexpected hanging, a remarkable puzzle which first became known by word of mouth in the early 1940s. One form of the paradox is the following: A prisoner has been sentenced on Saturday. The judge announces that "the hanging will take place at noon on one of the seven days of next week, but you will not know which day it is until you are told on the morning of the day of the hanging." The prisoner, on mulling this over, decided that the judge's sentence could not possibly be carried out. "For example," said he, "I can't be hanged next Saturday, the last day of the week, because on Friday afternoon I'd still be alive and I'd know for sure that I'd be hanged on Saturday. But I'd know this *before* I was told about it on Saturday morning, and this would contradict the judge's statement." In the same way, he argued, they could not hang him on Friday, or Thursday, or Wednesday, Tuesday, or Monday. "And they can't hang me tomorrow," thought the prisoner, "because I know it today!"

The paradox of the unexpected hanging

Careful analysis reveals that this argument is false, and that the decree can be carried out. The paradox is a subtle one. The crucial point is that a statement about a future event can be known to be a true prediction by one person but not known to be true by another person until *after* the event has taken place.

A rather different sort of a paradox is illustrated by the embarrassment of the executioner who is to do away with the prisoner. The condemned man is permitted to make one last statement before he is executed: if that statement is false, he is to be hanged; if it is true, he is to be beheaded. After a moment's thought, the prisoner makes a statement, whereupon the embarrassed executioner does nothing. What did the condemned man say?

BIBLIOGRAPHY. Further information may be found in the following references. They are grouped by topic and the type of material in each is indicated by its title.

General works: W.W.R. BALL and H.S.M. COXETER, *Mathematical Recreations and Essays* (1942, 1960); H.E. DUDENEY, *536 Puzzles and Curious Problems*, ed. by MARTIN GARDNER (1967); MARTIN GARDNER, *The Scientific American Book of Mathematical Puzzles and Diversions* (1959), *The Second Scientific American Book of Mathematical Puzzles and Diversions* (1961), *New Mathematical Diversions from Scientific American* (1966); J.A.H. HUNTER and J.S. MADACHY, *Mathematical Diversions* (1963); MAURICE KRAITCHIK, *Mathematical Recreations*, 2nd rev. ed. (1953); J.S. MADACHY, *Mathematics on Vacation* (1966); T.H. O'BEIRNE, *Puzzles and Paradoxes* (1965); HUBERT (CALIBAN) PHILLIPS, *Problem Omnibus*, 2 vol. (1962); FRED SCHUH, *The Master Book of Mathematical Recreations*, trans. by F. GOBEL, ed. by T.H. O'BEIRNE (1968).

Special topics: (Dissections): V.G. BOLTYANSKII, *Equivalent and Equidecomposable Figures* (1963); HARRY LINDGREN, *Geometric Dissections* (1964). (Fallacies): V.M. BRADIS, V.L. MINKOVSKII, and A.K. KHARCHEVA, *Lapses in Mathematical Reasoning* (1963); E.A. MAXWELL, *Fallacies in Mathematics* (1959). (Graphs): OYSTEIN ORE, *Graphs and Their Uses* (1963). (Logical inference): MAXEY BROOKE, *150 Puzzles in Crypt-Arithmetic* (1963); HUBERT (CALIBAN) PHILLIPS, *My Best Puzzles in Logic and Reasoning* (1961); G.J. SUMMERS, *Fifty Problems in Logical Deduction* (1969); C.R. WYLIE, *101 Puzzles in Thought and Logic* (1957). (Manipulative puzzles and games): MAXEY BROOKE, *Fun for the Money* (1963); SOLOMON GOLOMB, *Polyominoes* (1965); R.C. READ, *Tangrams: 330 Puzzles* (1965); T. SUNDARA ROW, *Geometric Exercises in Paper Folding* (1966); SIDNEY SACKSON, *A Gamut of Games* (1969); WALTER SHEPHERD, *Mazes and Labyrinths: A Book of Puzzles* (1961). (Polytopes): H.S.M. COXETER, *Regular Polytopes*, 2nd ed. (1963); H.M. CUNDY and A.P. ROLLETT, *Mathematical Models* (1967); L. FEJES TOTH, *Regular Figures* (1964). (Probability): WARREN WEAVER, *Lady Luck: The Theory of Probability* (1963).

(W.L.S.)

Number Theory

Number theory is a branch of mathematics concerned with the properties of integers, or whole numbers, such as 0, ± 1, ± 2, $\cdots$. These properties have been the object of fascination and investigation for thousands of years; interest in the natural numbers is as old as civilization itself. In this article some of the more elementary notions and examples are given first, the more sophisticated ideas following.

This article is divided into the following sections:

Elementary and algebraic number theory

ELEMENTARY THEORY OF NUMBERS

Divisibility and prime numbers. An integer a is said to be divisible by another integer b, not 0, if there is a third integer c such that $a = bc$. Thus 6 is divisible by 3 since the integer 2 exists for which $6 = 2 \cdot 3$. If this is the case, b is said to be a divisor or factor of a, and the fact is expressed by the notation $b|a$. If a and b are two integers, not both zero, the notation (a, b) is used for the greatest integer d that is a factor of both a and b. The integer d is known as the greatest common divisor of a and b. It was known to the 4th-century-BC Greek mathematician Euclid that integers x and y exist with the property that $ax + by = d$, d being the greatest common divisor of a and b.

Definition of prime

An integer p that is greater than 1, but has no positive divisors other than 1 and p, is said to be prime. Thus the first few prime numbers are 2, 3, 5, 7, 11, 13, 17, $\cdots$. An integer greater than one that is not prime is called composite. A fundamental theorem of arithmetic states that every integer greater than one may be written as a product of primes in a unique way (apart from rearrangements of the factors). Thus $666 = 2 \cdot 3 \cdot 3 \cdot 37$ but has no different expression as a product of primes except for the order of the factors. This result is a simple consequence of another of Euclid's theorems: if a prime p divides ab, then either p/a or p/b can itself be deduced from the theorem of Euclid mentioned above.

Euclid also knew that there are infinitely many prime numbers, and his proof is much quoted as an example of mathematical reasoning and an example of mathematical beauty revealed in elegance and simplicity. It runs as follows: If it be supposed, on the contrary, that there is only a finite number of primes and they are denoted by $p_1, p_2, \cdots, p_n$, then the number N that is a product of primes plus 1 (see Box, equation 1) is not divisible by any p_i (for which the subscript i is a number between 1 and n, possibly including 1 and n) and so must be divisible by a prime other than these (which may of course be N itself). This contradicts the hypothesis that there are no other primes.

The sieve of Eratosthenes. The first problem after the proof that infinitely many primes exist is that of finding some way of exhibiting all primes up to any given number n.

About 250 BC, a contemporary of the Greek mathematician Archimedes named Eratosthenes of Cyrene proposed the following procedure, which is now called the sieve of Eratosthenes. If it is required to find all primes less than 200, all the numbers up to 200 are written down (see 2). Every second integer after 2 is not a prime, every third after 3 is not, every fifth after 5 is not a prime, and so forth; these are struck out (see 3). The striking out of every 2nd, 3rd, and 5th has been illustrated. On the face of it, this seems like a laborious task, but in fact, this sieving is surprisingly efficient. It is efficient because it is only necessary to sieve by primes 2, 3, 5, 7, 11, 13, because 13 is the largest integer that is less than or equal to the square root of the number 200. The reason for this is that in general any integer n which is not a prime must have a prime factor $\leq \sqrt{n}$.

This simple principle has been refined, modified, and generalized into an extremely powerful tool in the theory of numbers.

(1) $N = p_1 p_2 \cdots p_r + 1$

(2) 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, $\cdots$, 197, 198, 199, 200

(3) 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, $\cdots$, 196, 197, 198, 199, 200

(4) $F_n = 2^{2^n} + 1$

(5) $\begin{cases} F_0 = 3, & F_1 = 5, & F_2 = 17, & F_3 = 257 \\ F_4 = 65{,}537 \end{cases}$

(6) $M_p = 2^p - 1$

(7) $\begin{cases} 6 = 1 + 2 + 3 \\ 28 = 1 + 2 + 4 + 7 + 14 \end{cases}$

(8) $2^{p-1}(2^p - 1)$

Fermat and Mersenne primes. The Fermat numbers F_n, for non-negative integer n, are defined by taking the 2^nth power of 2 and adding the number 1 (see 4). In 1640 a French mathematician, Pierre de Fermat, noted that the first five such numbers were all prime (see 5) and he was led to conjecture that F_n was prime for all n.

The 18th-century Swiss mathematician Leonhard Euler was able to show, however, that F_5 is composite. F_n is now known to be composite for all values of n from 5 to 16 and for several higher values. No further Fermat prime beyond F_4 has been discovered, and it is now thought that the total number of such primes is finite. The Fermat primes have an interesting connection with the classical problem of constructing a regular polygon by Euclidean methods (*i.e.*, by the use of straightedge and compass). The ancients knew that regular polygons of $2^k m$ sides could be constructed by such methods when $m = 3$ or 5. In 1796 the German mathematician Carl Friedrich Gauss proved—at the age of 18—that such constructions were possible if and only if m is a product of distinct Fermat primes.

The Mersenne numbers M_p are constructed from a prime p by taking the pth power of 2 and subtracting the number 1 (see 6). The French mathematician Marin Mersenne gave in 1644 a list of primes p for which he conjectured M_p was prime. Several mistakes have been found in this conjecture, though it was not until 1886 that the first error was discovered. A few very large primes of this form are known. In particular, in 1963, it was shown that $M_{11,213}$ is prime. Mersenne primes also have a connection with the ancient problem of perfect numbers. A number is said to be perfect if it is equal to the sum of its positive factors other than itself; examples are easily constructed (see 7). It was known to Euclid that if a number now called Mersenne is a prime then a certain power of two times this number is perfect (see 8). Euler showed that all even perfect numbers are of this type. Whether there exist any odd perfect numbers remains an unsolved problem.

Bernoulli numbers. The Swiss mathematician Jakob I Bernoulli introduced, in his *Ars conjectandi* (1713; "The Conjectural Arts"), an important sequence of numbers known as Bernoulli numbers. The nth Bernoulli number is denoted by the symbol B_n and the first few numbers

are simply expressed (see 9). The numbers of odd orders, other than B_1, are zero. The numbers of even order, other than B_0, alternate in sign.

One definition of Bernoulli numbers is a symbolic one (see 10), in which after expansion of the nth power of the sum $(B + 1)$, powers of B are degraded into subscripts. Thus B_3 is obtained by expanding a 4th power (see 11). In this way B_3 is obtained in terms of Bernoulli numbers of lower order. This is only one example of a host of recursion formulas for generating these numbers.

Historically the first definition of Bernoulli numbers was that given by Bernoulli himself. Bernoulli's approach was to re-express the sum of a common power of a finite number of successive integers as a linear combination of successive powers of a common integer. Terms in the re-expression provided the Bernoulli numbers as coefficients (see 12). Another definition of B_n, given in terms of an explicit formula, was due to Euler, who made a striking independent discovery of the numbers in 1748 (see 13).

The Bernoulli numbers have several applications in number theory—in particular in work on Fermat's last theorem. Perhaps the most famous property of B_n is expressed in the Von-Staudt-Clausen theorem (see 14) that expresses the $2n$th Bernoulli number as the difference between an integer and the sum of reciprocals of primes that exceed by 1 as divisor of $2n$. The 12th Bernoulli number provides an example (see 15).

Fermat's last theorem. Fermat's last theorem, an example of an unsolved theorem in number theory, concerns the equation (see 16) for which x, y, z, and n are nonzero integers. This is an example of a Diophantine equation (named after Diophantus of Alexandria, who lived about AD 250)—an equation the solutions of which are required to be integers.

The above equation can be solved when $n = 2$; for example, it is satisfied by the integers 3, 4, and 5 because $3^2 + 4^2 = 5^2$. Fermat's theorem states that if n is an integer greater than 2, there is no solution for this equation in nonzero integers. He wrote (*c.* 1637)—on the margin of his copy of the works of Diophantus—"I have discovered a truly remarkable proof which this margin is too small to contain." Although he proved it for $n = 4$, the general proof remains undiscovered. Leonhard Euler produced an incomplete proof for $n = 3$ in 1770, the missing steps being filled in by later mathematicians.

Attempts to prove Fermat's last theorem

To prove the theorem in general, it is sufficient to demonstrate the impossibility of a similar equation in which the common power is any odd prime greater than 3 (see 17). In 1823 Adrien-Marie Legendre of France showed that this was the case when the power is 5, and a countryman of his, Gabriel Lamé, in 1839 proved it for a power of 7. Many efforts were made to extend proofs to other powers, and in 1850 Ernst Eduard Kummer of Germany showed that the equation $x^p + y^p = z^p$ cannot be solved for positive integers x, y, and z if p is a prime that does not divide the numerators of $B_2, B_4, \cdots, B_{p-3}$, these being Bernoulli numbers. The criteria developed by Kummer have been used with a digital computer to show that Fermat's last theorem is true for all exponents less than 25,000.

The theorem has been attacked by many mathematicians and has acquired a unique reputation because of the great difficulties it presents. Some mathematicians think that Fermat's statement was incorrect, though it may be that the theorem will never be proved or disproved.

Residue classes and congruence. If m denotes a given positive integer, every integer a can be written uniquely in the form (see 18) in which q, an integer, denotes the quotient of a on division by m, and r, the remainder, takes one of the m values $0, 1, 2, \cdots, m - 1$. If two integers a and b have the same remainder on division by m, they are said to be congruent modulo m, written $a \equiv b \bmod m$. This is equivalent to the assertion $m|(a - b)$. The set of integers congruent to a given integer modulo m is called a residue class. Clearly there are m such residue classes.

Euler introduced the symbol $\varphi(m)$ to denote that number of positive integers a not exceeding m that are such that $(a, m) = 1$; if $(a, m) = 1$, a and m are said to be relatively prime. It is possible to prove that if $(m, n) = 1$, $\varphi(mn) = \varphi(m)\varphi(n)$; and that if p is a prime, $\varphi(p^r) = p^{r-1}(p - 1)$. These facts enable the calculation of $\varphi(m)$ for any given m. If $a \equiv b \bmod m$, $(a, m) = (b, m)$. In particular, an obvious meaning may now be attached to the term residue class modulo m, relatively prime to m. There are $\varphi(m)$ such residue classes.

Euler's theorem states that if $(a, m) = 1$, then $a^{\varphi(m)} \equiv 1 \pmod m$. This generalized a theorem of Fermat, who found that if p is prime and does not divide a then $a^{p-1} \equiv 1 \pmod p$.

$$(9)\quad \begin{cases} B_0 = 1, & B_1 = -\frac{1}{2}, & B_2 = \frac{1}{6} \\ B_3 = 0, & B_4 = -\frac{1}{30}, & B_5 = 0, \quad B_6 = \frac{1}{42} \\ B_7 = 0, & B_8 = -\frac{1}{30}, & B_9 = 0, \quad B_{10} = \frac{5}{66} \end{cases}$$

$$(10)\quad B_n = (B + 1)^n$$

$$(11)\quad (B + 1)^4 = B_4 + 4B_3 + 6B_2 + 4B_1 + B_0$$

$$(12)\quad \begin{cases} 1^k + 2^k + 3^k + \cdots + (N-1)^k = \\ \qquad \frac{1}{k+1}\left\{B_0 N^{k+1} + B_1\binom{k+1}{1}N^k + \cdots + \right. \\ \qquad \left. + B_k\binom{k+1}{k}N\right\} \\ \text{in which } \binom{n}{m} = \dfrac{n!}{m!(n-m)!} \end{cases}$$

$$(13)\quad B_{2n} = \frac{2(-1)^{n-1}(2n)!}{(2\pi)^{2n}}\left(1 + \frac{1}{2^{2n}} + \frac{1}{3^{2n}} + \frac{1}{4^{2n}} + \cdots\right)$$

$$(14)\quad B_{2n} = A_n - \frac{1}{p_1} - \frac{1}{p_2} - \cdots - \frac{1}{p_k}$$

$$(15)\quad B_{12} = -\frac{691}{2730} = 1 - \frac{1}{2} - \frac{1}{3} - \frac{1}{5} - \frac{1}{7} - \frac{1}{13}$$

$$(16)\quad x^n + y^n = z^n$$

$$(17)\quad x^l + y^l + z^l = 0$$

$$(18)\quad a = qm + r$$

$$(19)\quad \{1, a, a^2, \cdots, a^{\varphi(m)-1}\}$$

It may happen that for a given number m there exists a number a with the property that the set of certain powers of number a (see 19) contains precisely one number from each of the $\varphi(m)$ relatively prime residue classes modulo m. In such a case a is said to be a primitive root modulo m. Gauss showed that primitive roots exist when (and only when) m is 2, 4, p^k, or $2p^k$ for some odd prime p.

Three further theorems concerning congruences may be mentioned: (1) The Chinese remainder theorem, so called from its being known to the Chinese in the 1st century AD: If $m_1, \cdots, m_k$ are k positive integers, each pair of which are relatively prime, and if $r_1, r_2, \cdots, r_k$ are any k integers, then there exists an integer x satisfying the k congruences: $x \equiv r \pmod{m_1}, \cdots, x \equiv r_k \pmod{m_k}$. Moreover, any two such integers x are congruent modulo $m_1 \cdots m_k$. (2) Lagrange's theorem, named after the 18th-century French mathematician Joseph-Louis Lagrange: If p is a prime and $f(x)$ a polynomial of degree n with integer coefficients, then the congruence $f(x) \equiv 0 \bmod p$ is satisfied by at most n incongruent integers x. (3) Wilson's theorem, named after the 18th-century English mathematician John Wilson: If p is a prime, then $1 \cdot 2 \cdot 4 \cdots (p - 1) \equiv -1 \pmod p$.

Quadratic residues. If m is a positive integer, then an integer a that is relatively prime to m is said to be a quadratic residue of m if there exists an integer x satisfying $x^2 \equiv a \pmod{m}$. The most interesting case is when m is an odd prime p. In this situation $(p-1)/2$ of the positive integers not exceeding $p-1$ are quadratic residues of p and the remaining $(p-1)/2$ are not (they are known as quadratic non-residues). Numerous interesting results about quadratic residues are known; for example, -1 is a quadratic residue of an odd prime p if, and only if, $p \equiv 1 \bmod 4$; if $p \equiv 3 \bmod 4$, there are more quadratic residues in the first half of the interval from 1 to $p-1$ than in the second. The most striking and important of such theorems, however, is the law of quadratic reciprocity, which was first proved by Gauss, although its truth had been conjectured both by Euler and Legendre. It may be stated simply as follows: if p and q are different odd primes, then p is a quadratic residue of q if, and only if, q is a quadratic residue of p, unless both p and q are congruent to 3 modulo 4, in which case p is a quadratic residue of q if, and only if, q is a quadratic non-residue of p. A more succinct statement of this important law may be obtained by employing Legendre's symbol $(a|p)$, which is defined to equal 1 if a is a quadratic residue of p, -1 if a is not a quadratic residue, and 0 if $p|a$ (see 20).

$$(20)\qquad (p|q)(q|p) = (-1)^{\frac{1}{2}(p-1)\frac{1}{2}(q-1)}$$

$$(21)\qquad b_0 + \cfrac{a_1}{b_1 + \cfrac{a_2}{b_2 + \cfrac{a_3}{b_3 + \ddots}}}$$

$$(22)\qquad \frac{A_0}{B_0} = \frac{b_0}{1}, \quad \frac{A_1}{B_1} = \frac{b_0 b_1 + a_1}{b_1}, \quad \frac{A_2}{B_2} = \frac{b_0 b_1 b_2 + a_1 b_2 + a_2 b_0}{b_1 b_2 + a_2}, \quad \cdots$$

$$(23)\qquad \xi = [\xi] + \cfrac{1}{[\xi_1] + \cfrac{1}{[\xi_2] + \cdots}}$$

Continued fractions. A continued fraction is an expression of a form composed of a number plus a fraction in which the denominator of the fraction is again a number plus a fraction (see 21), which may or may not terminate. If it does terminate, it has an obvious meaning. If not, and the sequence (see 22), obtained by cutting off the continued fraction at successive stages, converges to a limit V, the continued fraction is said to converge, and to have the value V. (Plainly, for this to happen, B_n must be non-zero for sufficiently large n.)

Convergent fractions

The continued fraction is said to be simple if each $a_i = 1$ and each b_i $(i > 0)$ is a positive integer (b_0 is allowed to be an integer).

A simple continued fraction is always convergent. Every real number has an expansion as a simple continued fraction. To obtain the continued fraction expansion of a real number, symbolized by the Greek letter xi, ξ, it is necessary to write $[\xi] + \theta_1$, in which $[\xi]$ denotes the greatest integer not exceeding ξ, and $0 \leq \xi_1 < 1$. If $\theta_1 = 0$, the process terminates. Otherwise the proof continues with $\xi_1 = 1/\theta_1$ and then $\xi_1 = [\xi]_1 + \theta_2$ with $0 \leq \theta_2 < 1$. If $\theta_2 = 0$, the process terminates. If not, $\xi_2 = 1/\theta_2$ and the process goes on as before. In this way an expansion for ξ (see 23) is obtained. It may be shown that the expansion terminates if, and only if, ξ is rational, and otherwise it is periodic if, and only if, ξ is a quadratic surd. (Ed.)

ALGEBRAIC NUMBER THEORY

Whereas ordinary number theory is the study of the integers, $0, \pm 1, \pm 2, \cdots$, in algebraic number theory larger collections of numbers are studied. A number is algebraic if it is a root of a polynomial equation with integral coefficients; if the highest power of the variables has coefficient 1, the number is called an algebraic integer. For instance, $\sqrt{2}$ is an algebraic integer because it satisfies the equation $x^2 - 2 = 0$; so is $1 + \sqrt{3}$ in view of the equation $x^2 - 2x - 2 = 0$, which it satisfies; and the five roots of $x^5 - x + 2 = 0$ are all algebraic integers, although it is not possible to exhibit a simple formula for them.

Since about 1830, when the formal study of algebraic numbers began, evidence has multiplied that algebraic number theory is the best road to take in the search for a better understanding of ordinary numbers and a deeper insight into the mysteries they hold. In addition, it has been applied to nearly every other branch of mathematics.

Early history. For brevity certain standard symbols are used in this article, beginning with Z, the symbol for the set of ordinary integers $0, \pm 1, \pm 2, \cdots$. Such integers belong to this set and are said to lie in Z.

The simplest system of algebraic numbers going beyond this set is the set of all numbers of the form $a + ib$, in which a and b are in Z (*i.e.*, they are integers) and i is the square root of -1. These complex numbers are added and multiplied in the natural way, subject to the rule that $i^2 = -1$. For instance, $(2 + 3i) + (4 - 7i) = 6 - 4i$, and $(2 + 3i)(4 - 7i) = 8 - 14i + 12i - 21i^2 = 29 - 2i$. These numbers, first studied in depth in 1828 by Gauss, are called the Gaussian integers in his honour. The symbol $\mathfrak{G}$ is used for the set of Gaussian integers.

Gaussian integers

An illustration of the power of algebraic number theory lies in the proof that the equation $x^2 + 1 = y^3$ has in Z only the solution $x = 0$, $y = 1$. In other words, the only integers satisfying the equation are these values. This assertion was made by Euler in 1738 after he sketched a proof of a similar theorem. The first complete proof was given in 1875 by Pepin.

In the proof facts about $\mathfrak{G}$ are used to obtain information about Z, and the vital tool required for the proof is that in $\mathfrak{G}$, as in Z, numbers can be uniquely expressed as a product of primes. (Primes in $\mathfrak{G}$ may be defined in a similar way to primes in Z.) While this is true in $\mathfrak{G}$, it must be accompanied by a precaution concerning units (numbers that divide 1). In Z the only units are 1 and -1, and the ambiguity introduced is easily avoided by confining attention to positive numbers. On the other hand, in $\mathfrak{G}$, there are four units, ± 1 and $\pm i$, and it is necessary to accept the complication that primes of $\mathfrak{G}$ come in indistinguishable quadruplets. These primes are readily catalogued by observing the factors, lying in $\mathfrak{G}$, of primes in Z. Number 2, for example, becomes $i(1-i)^2$, with $1 - i$ a prime.

If n is an integer, $1, 2, 3, \cdots$, then primes of Z that can be expressed in the form $(4n - 1)$, such as $3, 7, 11, \cdots$, remain prime in $\mathfrak{G}$. Primes in Z of the form $(4n + 1)$, such as $5, 13, 17, \cdots$, have factors in $\mathfrak{G}$ that are two primes. Thus $5 = (2 + i)(2 - i)$, etc.

In this way a neat proof may be obtained of the theorem that any prime in Z of the form $4n + 1$ is uniquely expressible as a sum of two squares. This was stated by Fermat and proved by Euler.

The equation $x^2 + 1 = y^3$ has the form $(x + i)(x - i) = y^3$. It is first necessary to dispose of the possibility that $x + i$ and $x - i$ have a common factor. If this were the case, the common factor would divide their difference, $2i$, and it would follow that y is even and x is odd. This is quite impossible because $x^2 + 1$ would then be twice an odd number, whereas y^3 would be divisible by 8.

Unique factorization then yields the result that $x + i$ is a cube in $\mathfrak{G}$ (there is fortunately no difficulty with

units, because the equation $i = (-i)^3$ shows that all units in $\mathfrak{G}$ are cubes). If $x + i = (a + bi)^3$, then the equations $x = a^3 - 3ab^2$ and $1 = b(3a^2 - b^2)$ follow. Thus $b = \pm 1$, $1 = \pm(3a^2 - 1)$, $3a^2 = 0$ or 2, $a = 0$, $x = 0$. A similar method will show that $x^2 + 2 = y^3$ has in Z only the solutions $x = \pm 5$, $y = 3$; this is historically interesting as one of a number of statements made without proof by Fermat; it was not until 1875 that a complete proof was given, again by Pepin. For this problem unique factorization for numbers of the form $a + b\sqrt{-2}$ is required.

The first substantial study of algebraic number theory was made by Gauss in the memoir of 1828 mentioned above. In it he used $\mathfrak{G}$, the Gaussian integers, to break into new territory in the study of biquadratic (fourth power) residues in Z. For instance, it may be asked when the integer 2 is a biquadratic residue of a prime p, this meaning that there exists x with $x^4 - 2$ divisible by p. The question is not interesting if p has the form $4n - 1$, for then any quadratic residue is a biquadratic residue. For p of the form $4n + 1$, Gauss discovered the following criterion: 2 is a biquadratic residue of p if, and only if, p can be expressed in the form $a^2 + 64b^2$. As an illustration, the first prime having this form is $73 = 3^2 + 64(1^2)$, and indeed $25^4 - 2 = 390{,}623$ is divisible by 73. Again, here is a theorem stated wholly within Z; Gauss proved it by a skillful, detailed development of the properties of $\mathfrak{G}$.

Kummer's attempt to prove Fermat's last theorem

The next major advance in algebraic number theory was achieved by Kummer in his investigation of Fermat's last theorem. The problem is to prove the impossibility of solving $x^n + y^n = z^n$ in Z for $n \geq 3$. Kummer's method was to rewrite the equation in the form $(x + y) \cdot (x + uy) \cdot \cdots \cdot (x + u^{n-1}y) = z^n$, in which $u = e^{2\pi i/n} = \cos(2\pi/n) + i \sin(2\pi/n)$ is a primitive nth root of unity. For instance, when $n = 3$, $u = (-1 + \sqrt{-3})/2$, and the equation reads $(x + y)(x + uy)(x + u^2y) = z^3$. The problems that must now be overcome are precisely those encountered above in the study of the equation $x^2 + 1 = y^3$—units, relative primeness, and unique factorization. The difficulties concerning units and relative primeness were formidable, but Kummer mastered them completely. In his first work, however, he overlooked the need to prove unique factorization. The error was pointed out by the German mathematician Peter Gustav Lejeune Dirichlet. Because the first instance of non-uniqueness occurs at $n = 23$, it is not easy to exhibit the failure in connection with Fermat's last theorem. It is simpler to observe the phenomenon of non-uniqueness in an example in which the calculations are slight. The set of numbers $a + b\sqrt{-5}$ are considered in which a and b range over Z, the ordinary integers. There are two factorizations of 6 (see 24). It is easy to check that 2, 3, and $1 \pm \sqrt{-5}$ are all primes. Units play no role here because the only units are ± 1. Hence factorization is not unique in this case.

Despite the error that spoiled his proof, Kummer did not abandon the project. He had the novel, ingenious idea of restoring unique factorization by inventing additional "ideal numbers." By his new method Kummer made a great deal of progress on Fermat's last theorem, but did not arrive at a complete proof—nor did the numerous mathematicians who followed up his program. Computations based on Kummer's ideas, however, have shown $x^n + y^n = z^n$ to be impossible up to $n = 25{,}000$ (Richard M. Pollack and John L. Selfridge, 1964).

Foundations. In a systematic study of algebraic number theory, a basic role is played by the concept of a field (*i.e.*, a set of numbers in which the operations of addition, subtraction, multiplication, and division can be performed). The set $\mathfrak{Q}$ of rational numbers forms a field. On adjoining to $\mathfrak{Q}$ an algebraic number u, a typical algebraic number field, $\mathfrak{F} = \mathfrak{Q}(u)$, is obtained. The set $\mathfrak{R}$ of algebraic integers lying in $\mathfrak{F}$ is a collection admitting addition, subtraction, and multiplication (but not division); such a set is called a ring. In this ring uniqueness of factorization into primes may fail, as noted above in the case $\mathfrak{F} = \mathfrak{Q}(\sqrt{-5})$. Unique factorization may be restored by a method of the German mathematician Richard Dedekind by introducing ideals. An ideal in $\mathfrak{R}$

Rings

(24) $$6 = 2 \cdot 3 = (1 + \sqrt{-5})(1 - \sqrt{-5})$$

(25) $$6 = (2, 1 + \sqrt{-5})(2, 1 - \sqrt{-5}) \times (3, 1 + \sqrt{-5})(3, 1 - \sqrt{-5})$$

is a subset of $\mathfrak{R}$ that is closed under addition and subtraction and, furthermore, admits multiplication by any element of $\mathfrak{R}$. The most obvious example of an ideal is a principal ideal $x\mathfrak{R}$, consisting of all multiples of an element x of $\mathfrak{R}$. If every ideal is principal, then factorization is unique and the introduction of ideals is superfluous: this is, for instance, the case for Z, the integers, and $\mathfrak{G}$, the Gaussian integers. When factorization is not unique, so that there exist nonprincipal ideals, something valuable has been accomplished by the introduction of ideals, for Dedekind proved that every ideal is uniquely a product of prime ideals. (In the ring $\mathfrak{R}$ of algebraic integers in an algebraic number field, a prime ideal $\mathfrak{P}$ can be defined very simply as one admitting no factorization into ideals except the trivial factorization $\mathfrak{P} = \mathfrak{P}\mathfrak{R}$.) In the example $6 = 2 \cdot 3 = (1 + \sqrt{-5}) \cdot (1 - \sqrt{-5})$ of non-unique factorization given above, uniqueness is restored when the factors are decomposed further into prime ideals (see 25). Here the notation $(2, 1 + \sqrt{-5})$ denotes the ideal generated by 2 and $1 + \sqrt{-5}$; *i.e.*, the set of all numbers $2x + (1 + \sqrt{-5})y$ with x and y in the ring.

Class group and class number. It is possible, in a useful way, to measure the extent of the departure from unique factorization by an invariant called the class number. Crudely speaking, the larger the class number, the greater the failure of factorization to be unique.

The class number is essentially a measure of the number of nonprincipal ideals. To state this precisely, it is convenient to use fractional ideals—*i.e.*, ideals that may contain algebraic numbers that are not integers. These fractional ideals form a group under multiplication, with the principal ideals as a subgroup. The quotient group is called the class group. (For basic information on group theory see ALGEBRAIC STRUCTURES.) An important theorem of a Lithuanian-born mathematician, Hermann Minkowski, asserts that the class group is finite. The number of elements in the class group is the class number.

The class number is 1 precisely when unique factorization holds. Thus the class number of $\mathfrak{G}$ is 1, whereas the class number of $\mathfrak{Q}(\sqrt{-5})$ must exceed 1 (it can be shown to be 2). It is in general a difficult computation to find the class number of an algebraic number field, and its value oscillates unpredictably from one field to another. The case of quadratic field has been most thoroughly explored. (A quadratic algebraic number field $\mathfrak{Q}(\sqrt{m})$ is one obtained from the rational numbers $\mathfrak{Q}$ by adjoining a square root of a number m; m can be taken to be square-free [*i.e.*, to have no square factors other than 1], and the field is called real or imaginary according as m is positive or negative.) In his *Disquisitiones Arithmeticae* ("Discussions on numbers") Gauss listed nine imaginary quadratic fields with class number 1, given by $-m = 1, 2, 3, 7, 11, 19, 43, 67$, and 163. Gauss conjectured that these were all. There was no progress on this problem until 1934, when H. Heilbronn and E. Linfoot proved that there was at most one more. Computations by the U.S. mathematician D.H. Lehmer showed that a 10th field, if it existed, would have an enormous value for $-m$. Finally, in 1967 another U.S. mathematician, Harold M. Stark, proved that Gauss was right: the list of nine fields is complete. At about the same time, and independently, A. Baker developed methods that reduced the problem to a finite amount of computation. In retrospect it was recognized that in 1952 K. Heegner had made an attempt on the problem that could have been completed by techniques already known at that time.

Real and imaginary fields

This theorem of Heegner and Stark has an intriguing connection with an elementary question concerning primes. The question goes back to Euler, and the connection was

observed by G. Rabinovich in 1911. Euler had asked: when is it true for a prime p that $x^2 + x + p$ is prime for all the values 0, 1, 2, $\cdots$, $p - 2$? (It is fruitless to attempt to go higher than $x = p - 2$, for when x is set equal to $p - 1$, $x^2 + x + p$ takes the value p^2.) Euler gave the example 41, and there are five easily discovered smaller primes that work: 2, 3, 5, 11, and 17. If these six numbers are multiplied by 4 and 1 is subtracted, the numbers 7, 11, 19, 43, 67, and 163, the six largest values of $-m$ for fields of class number, are obtained. (The three values 1, 2, 3 are too small for there to be a connection with Euler's question.) It follows in fact from the theorem of Heegner and Stark that 41 is the last prime enjoying Euler's property.

For real quadratic fields (those of the form $\mathfrak{Q}(\sqrt{m})$ with m positive) numerical evidence indicates that the class number 1 occurs with great frequency. Gauss conjectured that there are an infinite number of real quadratic fields with class number 1, but this remains unsettled.

It is known that the class number of an algebraic number field can be arbitrarily large. One way of exhibiting large class numbers appears in Gauss's *Disquisitiones*. It is a consequence of a theorem of his on duplication of genera that the class number of the field $\mathfrak{Q}(\sqrt{m})$ is divisible by 2^{k-1} if the integer m has k or more distinct odd prime factors.

Dirichlet discovered a connection between the class numbers of certain quadratic fields and quartic fields. Given a square-free positive integer m, the quadratic fields $\mathfrak{Q}(\sqrt{m})$ and $\mathfrak{Q}(\sqrt{-m})$ can be formed, and in addition the field $\mathfrak{Q}(\sqrt{m}, \sqrt{-m})$ obtained by adjoining both $\sqrt{m}$ and $\sqrt{-m}$ to the rational numbers. Dirichlet's theorem asserts that the class number of $\mathfrak{Q}(\sqrt{m}, \sqrt{-m})$ is either the product of the class numbers of $\mathfrak{Q}(\sqrt{m})$ and $\mathfrak{Q}(\sqrt{-m})$ or else half that number. This result was the starting point for a considerable amount of subsequent research.

Units. In the ring $\mathfrak{R}$ of integers of an algebraic number field $\mathfrak{J}$, an element u of $\mathfrak{R}$ is called a unit if there exists an element v in $\mathfrak{R}$ with $uv = 1$. At least the numbers 1 and -1 are units; for some fields there are no other units. In a special category are those units u having the property that some power u^n is equal to 1; these are called roots of unity.

Dirichlet's unit theorem

In 1846 Dirichlet found a decisive theorem concerning the units of an arbitrary algebraic number field. This theorem gives a formula for the number of units, other than roots of unity, that are independent in a reasonable sense. $\mathfrak{J}$ is allowed to equal $\mathfrak{Q}(x)$ and is supposed to be n-dimensional over $\mathfrak{Q}$ so that the irreducible equation for x has degree n. Of the n roots of this equation, some are real and some imaginary, the imaginary ones coming in pairs; r_1 is written for the number of real roots and r_2 for the number of pairs of imaginary ones, so that $n = r_1 + 2r_2$. Dirichlet's formula for the number of independent units is $r_1 + r_2 = 1$.

Quadratic fields furnish simple illustrations for Dirichlet's unit theorem. If the field is imaginary, then $r_1 = 0$ and $r_2 = 1$, so that $r_1 + r_2 - 1 = 0$ and there are no units other than roots of unity. If the field is real, then $r_1 = 2$ and $r_2 = 0$, so that $r_1 + r_2 - 1 = 1$. There is then one basic unit called the fundamental unit, and all others are powers of it or their negatives. Stated in direct terms, this is the assertion about the equation named after the 17th-century English mathematician John Pell, $x^2 - Dy^2 = 1$, with D a non-square positive number. There is a solution in integers other than the trivial ones $x = \pm 1$, $y = 0$, and all solutions are obtainable in a simple way from a certain basic solution. As an illustration, if $D = 2$, the basic solution of $x^2 - 2y^2 = 1$ is given by $x = 3$, $y = 2$. If x^*, y^* is any solution, then $x^* + y^*\sqrt{2} = \pm(3 + 2\sqrt{2})^m$ for some m. Dirichlet's theorem can be viewed as a far-reaching generalization of the theory of the Pell equation.

The discriminant. If $\mathfrak{J}$ is an n-dimensional algebraic number field and $\mathfrak{R}$ its ring of integers, then there exists for $\mathfrak{R}$ an integral basis over the ring Z of ordinary integers. This is a set $u_1, \cdots, u_n$ of elements of $\mathfrak{R}$ with the property that every element of $\mathfrak{R}$ is uniquely a linear combination of the u's with coefficients in Z. For example, the numbers 1, i form an integral basis for the Gaussian integers. Now every member of $\mathfrak{J}$ has n conjugates that do not necessarily lie in $\mathfrak{J}$. When $u_{i1}, \cdots, u_{in}$ is written for the conjugates of u_i, d is the n by n determinant formed from the n^2 elements u_{ij}. It can be seen that $D = d^2$ lies in Z and that D does not depend on the choice of the integral basis; D is called the discriminant of $\mathfrak{J}$. The discriminant plays an important role in determining the properties of $\mathfrak{J}$.

A quadratic field $\mathfrak{Q}(\sqrt{m})$ serves as a simple illustration. If $m \equiv 2$ or 3 (mod 4), then an integral basis consists of 1 and $\sqrt{m}$, and $D = 4m$. If $m \equiv 1$ (mod 4), the elements 1 and $(1 + \sqrt{m})/2$ form an integral basis, and D turns out to equal m.

Some facts about the discriminant follow. Of these, the first is due to the Swiss mathematician Ludwig Stickelberger, the second to Dedekind, the third to the 20th-century Norwegian-born U.S. mathematician Oystein Ore, and the remaining three to the 19th–20th-century Lithuanian-born mathematician Herman Minkowski. (1) $D \equiv 0$ or 1 (mod 4); (2) the only primes in Z that acquire as factors multiple powers of prime ideals in $\mathfrak{R}$ (they are called the ramified primes) are those that divide D; (3) $2D$ is a multiple of the number of roots of unity in $\mathfrak{J}$; (4) D is never equal to 1 or -1; (5) there are only a finite number of fields with a given discriminant; (6) D becomes large with n; for example, there exists an absolute constant c such that $D > e^{cn}$.

The Zahlbericht and Hilbert's problems. At the invitation of the German Mathematical Society, the German mathematician David Hilbert prepared a report on algebraic number theory, published in 1897 and reprinted in his collected works. While initially planned as a joint project with Minkowski, he completed it alone. It was really a full length treatise, and in a masterful way it covered virtually all work in the field up to that date. Known as the "Zahlbericht" (number report), it became the "Bible" for all workers in the field. It fixed the notation, the names for the main concepts, and the style of proof used.

In 1900, at the International Mathematical Congress in Paris, Hilbert proposed 23 problems that he considered would be significant for research in the 20th century. The importance he attached to algebraic number theory is attested by three of the problems (the 9th, 11th, and 12th) being entirely in that field, and another (the 8th) partly so. These problems will be mentioned again below.

Hensel's p-adic numbers. As 20th-century work on algebraic number theory got under way, an important innovation was made by the German mathematician Kurt Hensel. Inspired partly by congruences to powers of primes that abound in number theory, and partly by an analogy with the power series that occur in algebraic function theory (see below), he conceived the idea of forming power series in a prime p. A typical p-adic number formed in this way can be written as a sum of powers of p with coefficients a_k (see 26) in which each coefficient a_i lies in the range $0 \leq a_i \leq p - 1$. If $a_{-1} = a_{-2} = a_{-3} = \cdots = 0$, the number is called a p-adic integer. This is just like a number written in the scale of p, except for the novelty that the digits are allowed to run on forever to the left. Addition and multiplication of p-adic numbers follow the usual rules of arithmetic. Such numbers could also be formed to the base 10, but there is then the serious disadvantage that divisors of zero occur—*i.e.*, there exist two non-zero 10-adic numbers the product of which is 0.

A simple illustration of the efficiency of using p-adic numbers is the statement that -1 has a square root in the 5-adic numbers. It is easily proved by an approximation procedure that is systematized once for all in a lemma (auxiliary proposition) named after Hensel. From this an infinite number of statements about ordinary integers may be deduced: for every n there exists an integer x such that $x^2 + 1$ is divisible by 5^n.

Out of Hensel's p-adic numbers there developed a major branch of modern algebra called valuation theory.

Quadratic forms and the Hasse principle. A quadratic form is an expression of the form $\Sigma\, a_{ij}\, x_i\, x_j$, the x's being variables. The earliest work on quadratic forms concerned the case in which the coefficients a_{ij} were ordinary integers. Eventually it was realized that it clarifies matters to study first the case of rational coefficients. This was the subject of a brilliant paper by Minkowski in 1890 in which he gave a complete classification of rational quadratic forms, which took advantage of virtually everything that had been done during the 19th century on integral quadratic forms. With this triumph fresh in his mind, Hilbert thought it timely in his 11th problem to ask for a classification of quadratic forms over an arbitrary algebraic number field.

The solution to Hilbert's 11th problem

The solution was given by the German mathematician Helmut Hasse in 1923, and the method was an impressive application of Hensel's p-adic numbers. Hasse first recast Minkowski's theorem in a new, attractive way. The field $\mathfrak{Q}$ of rational numbers can, for each prime p, be regarded as forming part of the field of p-adic numbers (say $\mathfrak{Q}_p$), and also part of the field $\mathfrak{R}$ of real numbers. The theorem states that two quadratic forms over $\mathfrak{Q}$ are equivalent (meaning that it is possible to pass from one to the other by a change of variable) if, and only if, they are equivalent over each $\mathfrak{Q}_p$ and over $\mathfrak{R}$. In this form the result is ready for a natural generalization to any algebraic number field, a generalization that Hasse succeeded in proving by a method, called the Hasse principle, that is regarded as a passage from local to global data. For the field of rational numbers, the local information is given one prime at a time, the field of real numbers being treated as a prime "at infinity." For a general algebraic number field, the primes are replaced by prime ideals, and there are usually several primes at infinity.

The Hasse principle has had other successes (for instance, in associative algebras and in algebraic groups), but for polynomial equations of degree higher than two it may fail. A simple example was given by the Norwegian Ernst Selmer in 1951: the equation $3x^3 + 4y^3 + 5z^3 = 0$ admits a nontrivial solution over the p-adic numbers for every p and also a solution over the real numbers, but it has no solution over the rational numbers.

Class field theory and reciprocity. A portion of algebraic number theory unites virtually everything that is known into a highly sophisticated structure. Two algebraic number fields $\mathfrak{F}$ and $\mathfrak{L}$ are studied, with $\mathfrak{L}$ an extension of $\mathfrak{F}$. Information is sought about how ideals behave in the passage from $\mathfrak{F}$ to $\mathfrak{L}$. The object is to classify the possibilities for $\mathfrak{L}$ in terms of readily computed objects in $\mathfrak{F}$. For a given $\mathfrak{F}$, there is a particularly important field $\mathfrak{L}$ that enjoys numerous special properties with respect to $\mathfrak{F}$: for instance, its dimension over $\mathfrak{F}$ is equal to the class number of $\mathfrak{F}$, and every ideal in $\mathfrak{F}$ becomes principal in $\mathfrak{L}$. The field $\mathfrak{L}$ is called the class field of $\mathfrak{F}$, the origin of the name class field theory.

There is a close relation between class field theory and reciprocity. In its original form, reciprocity referred to quadratic residues computed in the ring of ordinary integers and asserted the following: if p and q are odd primes, then p is a quadratic residue of q if, and only if, q is a quadratic residue of p, unless both p and q have the form $4n + 3$, in which case the behaviour is opposite. In his 9th problem, Hilbert asked for two advances to be made: squares were to be replaced by arbitrary nth powers, and the context was to be broadened to any algebraic number field.

Artin's reciprocity theorem

Many mathematicians (including Hilbert himself) made partial contributions until the climax in 1927, when Austrian-born Emil Artin discovered a definitive version of reciprocity and was able to prove it in full generality. It is typical of the increasing sophistication and abstraction of modern mathematics that in the theorem powers and reciprocity are no longer visible (indeed, it requires a fair amount of effort to extract a conventional reciprocity statement from Artin's theorem). Instead, Artin constructed two groups and a function from the first to the second; his theorem asserts that this function is an isomorphism (that is, it is one-to-one, onto, and preserves the product laws of the groups).

$$(26)\qquad \cdots + a_n p^n + \cdots + a_2 p^2 + a_1 p + a_0 + a_{-1} p^{-1} + + \cdots + a_{-m} p^{-m}$$

$$(27)\qquad \mathfrak{F} = \mathfrak{F}_0 \subset \mathfrak{F}_1 \subset \mathfrak{F}_2 \cdots$$

The class field tower. If $\mathfrak{F}$ be an algebraic number field and $\mathfrak{L}$ its class field, as described above, it is not necessarily the case that $\mathfrak{L}$ has class number 1; for, although every ideal in $\mathfrak{F}$ has become principal, some new ideals in $\mathfrak{L}$ may have arisen that are not principal. The process of passing to the class field can be iterated, resulting in the class field tower starting at $\mathfrak{F}$ and successively including other fields (see 27). It was an outstanding open question for many years whether, for any $\mathfrak{F}$, the class field tower had to end in a finite number of steps with a field having class number 1. Then, in 1964, the Russian mathematicians E.S. Golod and Igor Rostislavovich Shafarevich gave a counterexample. An explicit choice of a field $\mathfrak{F}$ having an infinite class field tower is $\mathfrak{Q}(\sqrt{-30030})$.

Abelian fields and the Jugendtraum. If u is a primitive nth root of 1, the algebraic number field $\mathfrak{F} = \mathfrak{Q}(u)$ is formed by adjoining u to the rationals; $\mathfrak{F}$ is called a cyclotomic field. It is known (by Galois theory: see the article ALGEBRAIC STRUCTURES) that $\mathfrak{F}$ is normal over $\mathfrak{Q}$ with a Galois group that is Abelian and has order $\varphi(n)$, in which φ is the Euler φ-function. If $\mathfrak{F}_0$ is any field between $\mathfrak{Q}$ and $\mathfrak{F}$, then $\mathfrak{F}_0$ likewise is Abelian over $\mathfrak{Q}$. In 1877 the German mathematician Leopold Kronecker proved the remarkable converse: any field Abelian over $\mathfrak{Q}$ is a subfield of a suitable cyclotomic field. In a letter to Dedekind dated March 15, 1880, Kronecker said his *liebster Jugendtraum* ("dearest dream of youth") had been to prove an analogous theorem concerning Abelian extensions of imaginary quadratic fields. This was achieved by the German mathematician Heinrich Weber, whose result showed that they are generated by certain values of elliptic functions (see ANALYSIS, COMPLEX). Hilbert's 12th problem called for extensions of this work to arbitrary algebraic number fields. Only scattered results have been obtained. Possibly further extensive development of class field theory is needed to obtain the requisite insight.

The analogy with algebraic functions. In studying algebraic functions in one variable the process begins with a field k, called the constant field, adjoins a variable x to get the field $\mathfrak{F} = k(x)$, and then forms a finite-dimensional extension of $\mathfrak{F}$ to get a field $\mathfrak{L}$. $\mathfrak{F}$ may be thought of as being analogous to the field $\mathfrak{Q}$ of rational numbers, and $\mathfrak{L}$ as being analogous to an algebraic number field. The ring $k[x]$ of polynomials in x is a subset of $\mathfrak{F}$; it is a principal ideal ring like the ring Z of integers. By taking appropriate integers inside $\mathfrak{L}$, a theory remarkably parallel to algebraic number theory is obtained. A great many theorems hold in both cases, and to a considerable extent a unified account covering both cases can be given. In particular, power series in x play a role quite similar to Hensel's p-adic numbers; in fact, they served to inspire him.

When the field k is finite, the analogy becomes closer still. All the major results of class field theory then hold, and there are perfect analogues of the unit theorem and the finiteness of the class number. Indeed, Artin and the U.S. mathematician George W. Whaples in 1945 set forth a simple set of axioms holding simultaneously for algebraic number fields and algebraic function fields over finite fields. The main axiom was the assertion of a neat product formula for valuations, asserting that certain infinite products are equal to 1. Arguing directly from their axioms, Artin and Whaples were able to give new unified proofs of the major theorems.

Analytic methods. In 1840 Dirichlet initiated the use of analysis on a large scale to prove results in number theory. Among other things he proved two remarkable theorems. The first was the existence of an infinite number of primes in every nontrivial arithmetic progression (an arithmetic progression is a sequence of the form a, a

Dirichlet's two analysis theorems

$+ d, a + 2d, \cdots$; it is nontrivial if a and d are relatively prime). The second was a class number formula. Since algebraic number theory had yet to be developed, Dirichlet's formula necessarily referred to binary quadratic forms. Translated into the language of algebraic number theory, his formula gave the class number of $\mathfrak{Q}(\sqrt{-p})$, with p a prime. It is simplest if p has the form $8n + 7$, and then is the sum from $r = 1$ to $r = (p - 1)/2$ of $(r|p)$, in which $(r|p)$ is the Legendre symbol, equal to 1 if r is a quadratic residue of p and -1 otherwise. For instance, if $p = 23$ the sum of $(1|p), \cdots, (11|p)$ is obtained. The numbers 1, 2, 3, 4, 6, 8, 9, are residues and 5, 7, 10, 11 are non-residues. Hence the class number of $\mathfrak{Q}(\sqrt{-23})$ is $7 - 4 = 3$.

The modernization of this work operates in an arbitrary algebraic number field $\mathfrak{F}$. The number, say $Z(t)$, of ideals with norm less than a given number t is counted; the norm of an ideal is an integer formed in analogy to the assignment of the integer $a^2 + b^2$ to the Gaussian integer $a + bi$. As t goes to infinity, $Z(t)/t$ can be proved to have a definite limit given by a formula built out of the class number, the discriminant, and the units of $\mathfrak{F}$. This result is then used to study the Riemann zeta function that can be formed relative to $\mathfrak{F}$. The result is that much information is obtained about the class number and the existence of prime ideals satisfying various restrictions, as called for in Hilbert's 8th problem. (I.K.)

Analytic number theory

THE SCOPE OF ANALYTIC NUMBER THEORY

Analytic number theory is a branch of number theory concerned with the interaction of analysis and number theory. Analysis can be used to prove certain properties of ordinary and algebraic integers and to establish quantitative results. Similarly, arithmetic properties can be used to analyze and shed light on analytic questions.

Among the earliest problems in number theory to which any sort of systematic study was brought to bear is the class of problems known as Diophantine equations, named after the Greek mathematician Diophantus of Alexandria who lived about AD 250. A Diophantine equation is an equation the solutions of which are required to be integers, the simplest example being the Diophantine equation involving first powers of x and y (see 28). The solution of these equations required the development of general properties of divisibility. The pursuit of Diophantine equations of higher degree led Gauss, in about 1820, to a study of a new set of integers, the Gaussian integers.

Diophantine equations

Properties of integers, whether ordinary or algebraic, may be of two types that are here characterized as qualitative, or descriptive, and quantitative. An example of the former is the question of whether, for a given integer n, the Diophantine equation that identifies n with a sum of second powers (see 29) has a solution in integers x_1, x_2, x_3, x_4. An example of the latter is the question of how many solutions this equation has.

In analytic number theory the problems investigated can be divided into three classes. The first class is that in which analysis is used to prove properties of integers, the statements of which can be formulated in terms of elementary concepts of mathematics. These are qualitative properties. One of the most noted examples is the Goldbach conjecture. This conjecture was made by the Prussian-born mathematician Christian Goldbach in 1742, and a related problem was partially solved by the Soviet mathematician Ivan Matveyevich Vinogradov in 1937. Goldbach conjectured that every number greater than or equal to 4 can be expressed as a sum of two primes. For example, the numbers 18 and 30 may be so expressed (see 30). While this conjecture had not been settled by the early 1970s, Vinogradov did prove in 1937 that every sufficiently large odd natural number can be written as the sum of three primes. It is notable that the statements of the conjecture of Goldbach and result of Vinogradov do not involve any analytic notions, yet to date the result of Vinogradov cannot be proved without deep and intricate analytic methods and techniques.

The second class of problem involves the use of analysis to establish quantitative results. Among the most important sequence of integers singled out for special study is the sequence of primes with which a function that indicates the number of primes less than or equal to x is associated (see 31).

Primes and $\pi(x)$ have been studied for many years, and one of the objectives was to find a simple formula for $\pi(x)$. This ambition was only partially realized when, in 1896, the French mathematician Jacques Hadamard and the Belgian mathematician Charles-Jean de la Vallée-Poussin independently proved that $\pi(x)$ is approximately $x/\log x$ in the precise sense that is expressed in terms of a limit (see 32).

The third class of problems makes use of arithmetic properties to analyze and shed light on analytic questions, the two interacting strongly.

Some problems dealt with under these three headings are given below. Qualitative problems include: (1) Various modifications of Goldbach's conjecture. (2) The 18th-century English mathematician Edward Waring's problem; *i.e.*, the solution of the Diophantine equation that identifies a sum of k powers with an integer n (see 33). (3) The existence of infinitely many primes in various sets, such as arithmetic progressions. (4) The existence of infinitely many pairs of primes differing by 2. (5) the existence of primes in various intervals—for example, the existence of a prime in the interval x to $x + x^{1/2}$; (6) the congruence properties of various functions, such as the partition function, coefficients of various modular functions; (7) the existence of solutions of Diophantine equations; (8) the properties of class numbers of algebraic number fields.

Quantitative problems include: (1) The prime factorization of an integer n (see 35). There are many functions associated with n such as the number of divisors of n (see 36). The behaviour of these functions and magnitude of their average values form part of the

(28) $ax + by = c$

(29) $n = x_1^2 + x_2^2 + x_3^2 + x_4^2$

(30) $18 = 13 + 5, \qquad 30 = 23 + 7$

(31) The function $\pi(x)$ associated with the sequence:
$2, 3, 5, 7, 11, 13, 17, 19, 23, \cdots$.
$\pi(x)$ = number of primes $\leqslant x$.
$\pi(10) = 4, \qquad \pi(20) = 8, \qquad \pi(35) = 10, \qquad \cdots$

(32) $$\lim_{x \to \infty} \frac{\pi(x) \log x}{x} = 1$$

(33) $x_1^k + \cdots + x_s^k = n$

(34) $$\sum_{n=0}^{\infty} p(n)x^n = \prod_{n=1}^{\infty} (1 - x^n)^{-1}$$

(35) $n = p_1^{\alpha_1} p_2^{\alpha_2} \cdots p_r^{\alpha_r}$

(36)
$d(n)$ = number of divisors of n
$\tau(n)$ = sum of the divisors of n
$\omega(n) = \alpha_1 + \alpha_2 + \cdots + \alpha_r$
$\Lambda(n) = \begin{cases} \log p \text{ if } r = 1 \\ 0 \text{ otherwise} \end{cases}$, the von Mangoldt function
$\mu(n) = \begin{cases} 0 \text{ if } \alpha_i \geqslant 2 \text{ for some } i \\ (-1)^r \text{ if } \alpha_i = 1 \text{ for all } i \end{cases}$, the Möbius function

(37) $x^m \cdot x^n = x^{m+n}$
$n^s \cdot m^s = (nm)^s$

object of this study. (2) The number of primes in various sequences such as arithmetic progressions, etc. (3) The interval between successive primes. (4) The magnitude of various exponential sums, including character sums. (5) Applications to the determination of the least primitive root mod p. (6) The magnitude of the class number of algebraic number fields. (7) Numbers of solutions of Diophantine equations; *e.g.*, number of solutions of $n = x_1^2 + x_2^2 + x_3^2 + x_4^2$, number of solutions in Waring's problem. (8) The orders of magnitude of coefficients of various power series (see 34). (9) The number of lattice points in various domains; for example, the number of lattice points in a circle. (10) Dirichlet and other densities.

Examples of the third class of problem include the following: (1) Kronecker's limit formulas; (2) Abelian functions; (3) complex multiplication; (4) modular functions over rational and algebraic number fields; (5) the German mathematician Erich Hecke's operators.

METHODOLOGY

History. Historically the use of analysis in number theory started with Euler in 1742 and is based on two simple observations, one concerned with integer powers of x and the second concerned with s powers of integers (see 37). In the first case the variable is x and in the second case the variable is s. Moreover, in the first case the product involves the sums $m + n$ of the integers m and n, and in the second case the product mn. Thus, it is to be expected that the first will be used in additive problems and the second in multiplicative problems. These can be illustrated with examples. For the first the equation that identifies k with squares of integers (see 38) is considered. Then, the series for squared powers of x (see 39) are considered. Multiplying these together leads to a relationship between series (see 40 and 41) in which the coefficient of x^k is the number of ways in which k can be written as a sum of two squares counting order (see 42). Because x^3 does not occur, then 3 cannot be written as a sum of two squares and so forth. If $r(k)$ is the number of solutions of the equation that identifies with k a sum of two squares of integers (see 43), then if $f(x)$ is an infinite sum of squared powers of x, the square of f can be computed in terms of a series with coefficients $r(k)$ (see 44) with the convention that $r(0) = 1$.

To illustrate the second case, the function $d(n) =$ number of divisors of n is considered: $d(n) = \Sigma 1$, the summation extended over all divisions of n. The series expressed in reciprocals of a power of positive integers (see 45) are used. Multiplying these two together leads to a square of such series (see 46).

The term $1/2^s$ occurs twice, as $1 \times 1/2^s$, $1/2^s \times 1$, likewise with $1/3^s$; but $1/4^s$ occurs three times as $1 \times 1/4^s$, $1/4^s \times 1$, $1/2^s \times 1/2^s$.

In general, the question is how many times $1/k^s$ occurs, or, in other words, what the coefficient of $1/k^s$ is. If $k = mn$, then the terms $1/m^s \times 1/n^s = 1/(mn)^s$ contribute 1 to the coefficient, and this happens for each decomposition; there are, however, as many decompositions as there are divisors of n. Hence, the coefficient of $1/k^s$ is $d(k)$ (see 47).

More generally, if any sequence of non-negative integers (see 48) is given, the function (in power series form with coefficients from the previous sequence; see 49) is called the generating function of the sequence. If a sequence of the form $b(k)$ is given (see 50) then the function that is formed from a series of reciprocals of s powers of integers, coefficients being selected from the given sequence (see 51), is called the Dirichlet series associated with the sequence.

The Dirichlet series

There are still other ways that a function can be associated with a given sequence, but attention is restricted for the time being to associations already given (see 49, 51). One association (49) is the more natural for additive problems, whereas another one (51) is more natural for multiplicative problems.

Classification and techniques. *Additive properties.* In the case of additive problems the general problem can be formulated as follows. It is required to write a natural number n as a sum of an element from a set A_1 plus an element from a set A_2 plus, etc. That is, it is required to determine whether an integer n can be written in the form of a sum of n_i (see 52) with $n_i \in A_i$ ($i = 1, 2, \cdots, r$), and the number $\nu(n)$ of such representations also is required. Evidently a representation exists if it can be shown that $\nu(n) \geq 1$.

As above, a sum of powers of x (see 53) is formed quite generally. Then it follows that a product of such sums is possible (see 54). Multiplying leads to another sum of powers of x (see 55). Setting n equal to a

(38) $$k = m^2 + n^2$$

(39) $$\begin{cases} \sum_{m=0}^{\infty} x^{m^2} = 1 + x + x^4 + x^9 + \cdots \\ \sum_{n=0}^{\infty} x^{n^2} = 1 + x + x^4 + x^9 + \cdots \end{cases}$$

(40) $$\sum_{m=0}^{\infty} x^{m^2} \sum_{n=0}^{\infty} x^{n^2} = \sum_{m,n} x^{m^2+n^2}$$

(41) $$(1 + x + x^4 + x^9 + \cdots)(1 + x + x^4 + x^9 + \cdots) = 1 + 2x + x^2 + 2x^4 + 2x^5 + x^8 + 2x^9 + 2x^{10} + \cdots$$

(42) $$\begin{cases} 1 = 1^2 + 0^2 = 0^2 + 1^2 \\ 2 = 1^2 + 1^2 \\ 4 = 4^2 + 0^2 = 0^2 + 4^2 \\ 5 = 1^2 + 4^2 = 4^2 + 1^2 \end{cases}$$

(43) $$k = m^2 + n^2$$

(44) $$f^2(x) = \sum_{m=0}^{\infty} x^{m^2} \sum_{n=0}^{\infty} x^{n^2} = \sum_{k=0}^{\infty} r(k) x^k$$

(45) $$\begin{cases} \zeta(s) = \sum_{n=1}^{\infty} \frac{1}{n^s} = 1 + \frac{1}{2^s} + \frac{1}{3^s} + \cdots + \frac{1}{n^s} + \cdots \\ \zeta(s) = \sum_{m=1}^{\infty} \frac{1}{m^s} = 1 + \frac{1}{2^s} + \frac{1}{3^s} + \cdots + \frac{1}{m^s} + \cdots \end{cases}$$

(46) $$\begin{cases} \zeta^2(s) = \left(1 + \frac{1}{2^s} + \frac{1}{3^s} + \cdots\right)\left(1 + \frac{1}{2^s} + \frac{1}{3^s} + \cdots\right) \\ \qquad = 1 + \frac{2}{2^s} + \frac{2}{3^s} + \frac{3}{4^s} + \cdots \end{cases}$$

(47) $$\zeta^2(s) = \sum_{k=1}^{\infty} \frac{d(k)}{k^s}$$

(48) $$A = \{a(0), a(1), a(2), \cdots, a(n), \cdots\}$$

(49) $$f(x) = \sum_{n=0}^{\infty} a(n) x^n$$

(50) $$B = \{b(1), b(2), \cdots, b(n), \cdots\}$$

(51) $$g(s) = \sum_{n=1}^{\infty} \frac{b(n)}{n^s}$$

(52) $$n = n_1 + n_2 + \cdots + n_r$$

(53) $$f_i(x) = \sum_{n_i \in A_i} x^{n_i}$$

(54) $$f_1(x) f_2(x) \cdots f_r(x) = \sum_{n_1 \in A_1} x^{n_1} \sum_{n_2 \in A_2} x^{n_2} \cdots \sum_{n_r \in A_r} x^{n_r}$$

sum of integers (see 56), it follows that x^n occurs each time there is a representation in the form given by (52); that is, a sum of integer powers of x each power multiplied by a coefficient $\nu(n)$ (see 57). Naturally, the determination of $\nu(n)$ is by no means resolved without detailed analysis of the functions $f_1(x), \cdots, f_r(x)$. The function $f(x) = f_1(x), \cdots, f_r(x)$ is the generating function for $\nu(n)$.

Generating functions

As another example, the generating function for the number of unrestricted partitions $p(n)$ of a natural number n can be found. A partition of n is a decomposition into an unrestricted sum of natural numbers m_i (see 58). For example, if $n = 5$, then all the cases can be listed (see 59); therefore, $p(5) = 7$.

To obtain the generating function, the partition must be made more systematic; r_1 is the number of 1's, r_2 the number of 2's, r_3 the number of 3's, $\cdots$, r_k the number of k's, occurring in a particular partition; so that n is a sum of products of integers times corresponding r_k (see 60). Thus a partition corresponds to a solution of an equation identifying n with a sum of integers, each integer in the sum belonging to a certain class (see 61). Therefore, the functions composed of sums of powers of x (see 62) are formed and the product is given by a general sum (see 63) with the convention that $p(0) = 1$. Noting that the expression for $f_k(x)$ is a geometric series, it follows that $f_k(x) = (1 - x^k)^{-1}$ and, therefore, that a product expression can be obtained (see 64). This is Euler's example. All questions of convergence have been ignored in this illustration.

Multiplicative properties. Here the problems dealt with are rooted in the multiplicative properties of integers. Let $a(1), a(2), \cdots, a(n), \cdots$ and $b(1), b(2), \cdots, b(m), \cdots$ be any two sequences. Guided again by the observation made earlier, the Dirichlet series is considered for each sequence (see 65). Then a product of sums can be reduced to a single sum (see 66).

Putting $mn = k$, the coefficient of k^{-s} is determined by a sum of terms $a(n)b(m)$, and $a(n)b(m)$ is part of the coefficient whenever $mn = k$. This is written succinctly as a sum over all integers nm whose product equals k (see 67) or equivalently as a sum (see 68), the sum being over all divisors of n.

As a further example, the Dirichlet series for the German mathematician Hans von Mangoldt's function may be found (see 69). If the f is set equal to a sum with typical term $\Lambda(n)$ divided by the s power of n, a moment's reflection shows that the sum of $\Lambda(d)$ over divisions d of n equals $\log n$ (see 70). Therefore $f(s) = -[\zeta'(s)/\zeta(s)]$, in which $\zeta'(s)$ is the derivative of $\zeta(s)$ (see 71).

SPECIFIC TOPICS IN NUMBER THEORY

The distribution of prime numbers. The earliest recorded fact about primes is Euclid's proof that the number of primes is infinite, occurring as proposition 20 in the 9th book of his *Elements*. The sieve of Eratosthenes, introduced about 250 BC, is a technique for exhibiting the primes below a certain number. The subject of primes and their distribution then lay dormant until 1640, when Fermat wrote that he was "almost convinced" that numbers of a form $2^n + 1$ were prime if n was a power of 2. Euler showed that this is false for $n = 2^5$ because $2^{32} + 1$ is divisible by 641.

The next contribution of importance was made by Euler, in a paper, "Variae observationes circa series infinitas," published in the Petersburg Academy, 1742. Euler connected the harmonic series (see 72), being not too much concerned with convergence, with the sum of the reciprocals of the primes (see 73). This ushered a new era in the theory of prime numbers, and a study of the function composed of a sum of s powers of positive integers (see 74) was initiated.

The next contribution was due to the French mathematician Adrien-Marie Legendre. In *Essai sur la théorie des nombres* (1798), he proposed that for large x, $\pi(x)$ had a specific form expressed in terms of the log x (see 75) approximately in which $\pi(x)$, as above, is the number of primes $\leq x$. Legendre was unable to give a proof,

$$(55)\quad f_1(x)f_2(x)\cdots f_r(x)=\sum_{\substack{n_1\cdots n_r\\ n_i\in A_i}} x^{n_1+n_2+\cdots+n_r}$$

$$(56)\quad n=n_1+n_2+\cdots+n_r$$

$$(57)\quad f(x)=f_1(x)f_2(x)\cdots f_r(x)=\sum_{n=0}^{\infty}\nu(n)x^n$$

$$(58)\quad n=m_1+m_2+m_3+\cdots+m_\ell$$

$$(59)\quad \begin{cases} 5=1+1+1+1+1\\ \quad=1+1+1+2\\ \quad=1+1+3\\ \quad=1+2+2\\ \quad=2+3\\ \quad=1+4\\ \quad=5 \end{cases}$$

$$(60)\quad n=1r_1+2r_2+\cdots+kr_k$$

$$(61)\quad \begin{cases} n=n_1+n_2+\cdots+n_k\\ n_1\in A_1=\{0,1,2,\ldots\}\\ n_2\in A_2=\{0,2,4,6,\ldots\}\\ \quad\cdot\\ \quad\cdot\\ \quad\cdot\\ n_k\in A_k=\{0,k,2k,\ldots\} \end{cases}$$

$$(62)\quad \begin{cases} f_1(x)=\sum_{n_1=0}^{\infty}x^{n_1}\\ f_2(x)=\sum_{n_2=0}^{\infty}x^{2n_2}\\ \quad\cdot\\ \quad\cdot\\ \quad\cdot\\ f_k(x)=\sum_{n_k=0}^{\infty}x^{kn_k} \end{cases}$$

$$(63)\quad \begin{cases} F(x)=f_1(x)f_2(x)\cdots f_k(x)\cdots\\ \quad=\sum_{n_1,n_2,\cdots,n_k,\cdots} x^{n_1+2n_2+\cdots+kn_k+\cdots}\\ \quad=\sum_{n=0}^{\infty}p(n)x^n \end{cases}$$

$$(64)\quad \sum_{n=0}^{\infty}p(n)x^n=\prod_{k=1}^{\infty}(1-x^k)^{-1}=F(x)$$

$$(65)\quad \begin{cases} f(s)=\sum_{n=1}^{\infty}\frac{a(n)}{n^s}\\ g(s)=\sum_{m=1}^{\infty}\frac{b(m)}{m^s} \end{cases}$$

$$(66)\quad f(s)g(s)=\sum_{n=1}^{\infty}\sum_{m=1}^{\infty}\frac{a(n)b(m)}{(nm)^s}=\sum_{k=1}^{\infty}\frac{c(k)}{k^s}$$

$$(67)\quad c(k)=\sum_{nm=k}a(n)b(m)$$

$$(68)\quad c(k)=\sum_{d|k}a(d)b\left(\frac{k}{d}\right)=\sum_{d|k}a\left(\frac{k}{d}\right)b(d)$$

and, as a matter of fact, later work shows that if for large x, $\pi(x)$ could be expressed in terms of A (see 76), then the best value of A is 1.

It was noted above that the primes appear to diminish in frequency, the number of primes in an interval of a given length getting steadily smaller as the beginning of the given interval is larger. In a letter to the German astronomer Johann Franz Encke, Gauss in 1792 suggested an explicit rate at which the frequency diminishes and proposed that the number of primes per unit interval beginning at x is $1/\log x$. Thus, he proposed a formula involving an integral of the reciprocal of a logarithmic function (see 77) approximately, the expression $\mathrm{li}(x)$ being an abbreviation for the integral.

The Russian mathematician Pafnuty Lvovich Chebyshev, in 1851, set out to prove that a certain limiting relation holds that involves the number of primes less than or equal to x (see 78). Although he did not succeed, he did at least establish the plausibility of this relation and, therefore, of those of Legendre and Gauss. He proved in particular that if the limit exists, then it must be 1.

Riemann's contribution

In 1859, the German mathematician Bernhard Riemann inaugurated a new era in the theory of distribution of primes. Returning to the function introduced by Euler, Riemann considered the function written as a sum of reciprocals of s powers the positive integers (see 79) with s a complex variable $s = \sigma + it$. This seemingly minor modification had a profound consequence. Using complex function theory, Riemann first derived a functional equation for $\zeta(s)$, expressed $\pi(x)$ in terms of $\zeta(s)$, wrote $\zeta(s)$ in terms of its zeros, and inferred the plausibility of the relation that the number of primes less than or equal to x is asymptotically the same as $\mathrm{li}(x)$ (see 80) in which the tilde sign is used and, in fact, gives a more precise statement. The $\sim$ sign is an abbreviation for the fact that the ratio tends to 1. The completion of this proof was carried out independently and simultaneously by Hadamard and de la Vallée-Poussin in 1896.

In the 20th century much of the work centring on the distribution of primes has been concerned with the finer properties of their distribution. Having established that $\pi(x)$ is approximately $\mathrm{li}(x)$, it is natural to determine how accurately $\mathrm{li}(x)$ in fact represents $\pi(x)$. This problem is intimately connected with the Riemann zeta function and its zeros. If the series form of the function is given with complex argument (see 81), then the series converges for $\sigma > 1$, but there is an analytic continuation into the whole complex plane, and $\zeta(s)$ is meromorphic with a simple pole at $s = 1$ with residue 1. Instead of working with $\pi(x)$, it is more natural analytically to work with a function defined in terms of the logarithm and written ψ (see 82). The connection with $\pi(x)$ is: the statement that $\pi(x)$ is asymptotically equal to x divided by the logarithm of x (see 83) is equivalent to the statement that $\psi(x) \sim x$.

The advantage of working with $\Lambda(n)$ lies in the Dirichlet series associated with $\Lambda(n)$ (see 84) as contrasted with the more complicated series expressed as a sum of reciprocals of s powers of p (see 85). The starting point for studying $\psi(x)$, and hence $\pi(x)$, is the relation that expresses the function ψ as a line integral that involves the zeta function (see 86).

Cauchy's theorem is then used to move the line of integration past the pole of the integrand, but here is encountered a great difficulty because the poles of the integrand involve the zeros of $\zeta(s)$, the function $\zeta(s)$ occurring in the denominator.

The Riemann hypothesis

In a now famous memoir of 1859, Riemann proposed the conjecture that all of the complex zeros $\beta + i\gamma$ of $\zeta(s)$ had $\beta = \frac{1}{2}$. The complex zeros are known to satisfy $0 < \beta < 1$. This is the so-called Riemann hypothesis, which remains unproved. What can be proved is that if the Riemann hypothesis is true, then there exists a constant A such that if $\Delta(x) = \psi(x) - x$, then the absolute value of Δ has a bound that depends on A and the logarithm of x (see 87) and a corresponding result for $\pi(x)$: If $\pi(x) - \mathrm{li}\, x = D(x)$, then the absolute value of $D(x)$ is bounded by a constant B times a function of x

(69) $$\Lambda(n) = \begin{cases} \log p & \text{if } n = p^k \\ 0 & \text{otherwise} \end{cases}$$

(70) $$\begin{cases} f(s) = \sum_{n=1}^{\infty} \dfrac{\Lambda(n)}{n^s} \\ \sum_{d \mid n} \Lambda(d) = \log n \end{cases}$$

(71) $$f(s) \cdot \zeta(s) = \sum_{n=1}^{\infty} \frac{\log n}{n^s} = -\zeta'(s)$$

(72) $$\sum_{n=1}^{\infty} \frac{1}{n}$$

(73) $$\sum_{p} \frac{1}{p}$$

(74) $$\zeta(s) = \sum_{n=1}^{\infty} \frac{1}{n^s}$$

(75) $$\pi(x) = \frac{x}{\log x - 1.08366}$$

(76) $$\pi(x) = \frac{x}{\log x - A}$$

(77) $$\pi(x) = \int_2^x \frac{dt}{\log t} = \mathrm{li}(x)$$

(78) $$\lim_{x \to \infty} \frac{\pi(x) \log x}{x} = 1$$

(79) $$\zeta(s) = \sum_{n=1}^{\infty} \frac{1}{n^s}$$

(80) $$\pi(x) \sim \mathrm{li}(x)$$

(81) $$\zeta(s) = \sum_{n=1}^{\infty} \frac{1}{n^s} \quad \text{with } s = \sigma + it$$

(82) $$\begin{cases} \psi(x) = \sum_{n \leq x} \Lambda(n) \\ \Lambda(n) = \begin{cases} \log p & \text{if } n = p^k \\ 0 & \text{otherwise} \end{cases} \end{cases}$$

(83) $$\pi(x) \sim \frac{x}{\log x}$$

(84) $$\sum_{n=1}^{\infty} \frac{\Lambda(n)}{n^s} = -\frac{\zeta'(s)}{\zeta(s)}$$

(85) $$\sum_{p} \frac{1}{p^s} = \sum_{n=1}^{\infty} \frac{\mu(n)}{n} \log \zeta(ns)$$

(86) $$\psi(x) = \frac{1}{2\pi i} \int_{a-i\infty}^{a+i\infty} \left(-\frac{\zeta'(s)}{\zeta(s)} \right) \frac{x^s}{s}\, ds \qquad a > 1$$

(87) $$|\Delta(x)| \leq A x^{\frac{1}{2}} \log^2 x$$

(88) $$|\pi(x) - \mathrm{li}\, x| \leq B x^{\frac{1}{2}} \log x$$

(see 88). By contrast the best result known to date is that there are constants c, A such that a bound for the absolute value of Δ as a function of x depends upon c, A, and a complicated function of x (see 89).

This result is due to A. Walfisz and is based on the work of Hermann Weyl, Godfrey Harold Hardy, John Edensor Littlewood, Ivan Matveyevich Vinogradov, and Nikolay Mikhaylovich Korobov.

In the opposite direction, Littlewood has proved that the error $\Delta(x)$ cannot be much different from that predicted by the Riemann hypothesis. More precisely, he has proved that for infinitely many values of x, there is for $c > 0$ a lower bound that depends on c as well as upon the logarithm of the logarithm of the logarithm of x (see 90) and for infinitely many values of x there is an upper bound for Δ of similar structure (see 91). Concerning the finer properties of the primes, there is the result of asymptotic nature concerning π as a function of x (see 92). Thus, if x is sufficiently large, there is a prime between x and $x + x^{5/8}$. Furthermore, it can be deduced that there is a constant A such that for every n an inequality holds for the absolute difference between adjacent primes (see 93). H.L. Montgomery, using sieving methods, has shown that $|p_{n+1} - p_n| \leq Ap_n^{3/5+\varepsilon}$, and M.N. Huxley has made a further refinement of the index to $7/12 + \varepsilon$. This is to be contrasted with the inference to be drawn if the Riemann hypothesis is true (see 94). It is an unsolved problem whether for x sufficiently large a difference expression involving π as a function of x is positive (see 95). Much effort has been devoted to the Riemann hypothesis, but by the 1970s the best known about a region in the strip $0 \leq \sigma \leq 1$ in which the function has no zeros is that σ is greater than or equal to 1 minus a function that involves A and the logarithm of t (see 96), a result due to Korobov and Vinogradov and used to prove the above quoted result on $\Delta(x)$.

(89) $|\Delta(x)| \leq Axe^{-c(\log x)^{3/5}(\log\log x)^{1/5}}$

(90) $\Delta(x) > cx^{\frac{1}{2}} \log\log\log x$

(91) $\Delta(x) < -cx^{\frac{1}{2}} \log\log\log x$

(92) $\pi(x+x^{\frac{5}{8}}) - \pi(x) \sim \dfrac{x^{\frac{5}{8}}}{\log x}$

(93) $|p_{n+1} - p_n| \leq Ap_n^{\frac{5}{8}+\varepsilon}$ for any $\varepsilon > 0$

(94) $|p_{n+1} - p_n| \leq Ap_n^{\frac{1}{2}} \log p_n$

(95) $\pi(x+x^{\frac{1}{2}}) - \pi(x) > 0$

(96) $\begin{cases} \zeta(s) \neq 0 \quad \text{for} \\ \sigma \geq 1 - \dfrac{A}{(\log t)^{\alpha}} \quad \text{with } t \geq 3, \qquad \alpha \geq 2/3 \end{cases}$

(97) $\begin{cases} \displaystyle\sum_{n=1}^{\infty} \frac{\mu(n)}{n^s} = \frac{1}{\zeta(s)} \\ \displaystyle\sum_{n=1}^{\infty} \frac{\varphi(n)}{n^s} = \frac{\zeta(s-1)}{\zeta(s)} \end{cases}$

(98) $\Phi(x) = \displaystyle\sum_{n \leq x} \varphi(n) = \frac{3}{\pi^2} x^2 + \Delta(x)$

(99) $|M(x)| \leq Axe^{-c\sqrt{\log x}}$

(100) $a, a+k, a+2k, \cdots$

(101) $\displaystyle\lim_{x \to \infty} \frac{\pi(a_1, k, x)}{\pi(a_2, k, x)} = 1$

(102) $\begin{cases} 1, 5, 9, 13, 17, 21, \cdots \\ 3, 7, 11, 15, 19, 23, \cdots \end{cases}$

(103) $\log \zeta(s) = \displaystyle\sum_p \frac{1}{p^s} + R(s)$

(104) $\displaystyle\sum_p \frac{1}{p}$

(105) $\varepsilon(n) = \begin{cases} 0 & \text{if } n \not\equiv a \pmod{k} \\ 1 & \text{if } n \equiv a \pmod{k} \end{cases}$

(106) $\displaystyle\sum_p \frac{\varepsilon(p)}{p^s} = \sum_{p \equiv a \pmod{k}} \frac{1}{p^s}$

(107) $\begin{cases} \chi(n) = \chi(m) \quad \text{if} \quad n \equiv m \pmod{k} \\ \chi(mn) = \chi(m)\chi(n) \\ \chi(n) = 0 \qquad \text{if } (n, k) > 1 \\ \chi(n) \text{ is not identically } 0 \end{cases}$

Other arithmetic functions. The analysis of other arithmetic functions such as those referred to above takes place by way of the Dirichlet series determined by them. For example, two summations are given, one of which involves the function μ and relates it to the ζ function, the other of which involves the function φ and relates it to the ζ function (see 97).

Typical results are Φ as a function of x expressed as a sum of φ as a function of n (see 98), in which $|\Delta(x)| \leq Ax \log x$. If $M(x) = \Sigma\mu(n)$, the summation for $n \leq x$, then the absolute value of M is bounded by a function expressed in terms of A and the exponential of a square root of the logarithm of x (see 99).

Primes in arithmetic progressions. If k and a be two relatively prime integers with $k \geq 2$ and the arithmetic progression, a sequence of integers beginning with a (see 100), is considered, then the questions raised by Legendre in 1803 were these:

Legendre's questions

(i) Are there infinitely many primes in this progression; *i.e.*, are there infinitely many $p \equiv a \pmod{k}$?

(ii) If so, and $\pi(a, k, x)$ be the number of such primes $\leq x$, then if $(a_1, k) = (a_2, k) = 1$, is it true that a limiting expression for a ratio of functions π equals 1 (see 101)? That is, are the numbers of primes in the different residue classes modulo k equidistributed?

The example gives the progressions $1 + 4n$, and $3 + 4n$—*i.e.*, two sequences, one beginning with 1 and one beginning with 3 (see 102). Are there as many primes in the first sequence as there are in the second in the sense of the limit of the ratio already given (see 101)? Dirichlet gave an affirmative answer to the first question in 1837 and laid the foundations for an affirmative answer to the second. In doing so, Dirichlet was guided by Euler's proof that $\pi(x) \to \infty$ and was led to the concept of a character. Euler had proved that the logarithm of the ζ-function can be expressed as a sum of reciprocals of s powers of primes plus a remainder (see 103) in which the first sum is over all primes and in which the term $R(s)$ remains bounded as $s \to 1$. It then followed that the sum of the reciprocals of the primes (see 104) diverged in view of the fact that $\zeta(s) \to \infty$ as $s \to 1$.

To duplicate the argument requires a function ε as a function of n being defined to take on values 1 or 0 on conditions that depend on a mod k (see 105) and then requires that the sum over all primes of the ratio of $\varepsilon(p)$ to p^s (see 106) be expressed in terms of functions the properties of which lend themselves readily to investigation. To produce such a function, Dirichlet postulates the existence of a function χ for fixed $k \geq 2$ with specific properties (see 107). The following facts are then established: (a) Such a function does indeed exist, and there are $\varphi(k)$ distinct such characters in which $\varphi(k) =$

the number of integers less than k and having no factors in common with k, called the Euler φ-function. (b) The values of $\chi(n)$ that are different from 0 are $\varphi(k)$-th roots of unity; *i.e.*, satisfy the equation $x^{\varphi(k)} = 1$. (c) There is a character χ_1 the values of which are as follows: $\chi_1(a) = 1$ if $(a, k) = 1$, $\chi_1(a) = 0$ if $(a, k) > 1$; χ_1 is called the principal character. Two further conditions may be expressed symbolically (see 108). The Dirichlet L-series is now defined by a sum over positive integers of ratios of the function χ to s powers of the integers for s a complex number (see 109).

These functions have many properties, some of which are displayed (see 110). The last relation is one of the fundamental and important properties of $L(s, \chi)$. Much attention has been devoted to a proof, but no simple proof is known.

With the help of the various properties enumerated, it is now possible to show that if a^* is chosen such that $aa^* \equiv 1 \pmod{k}$, then an equality exists between sums that relates the function χ, the function L, the function φ, and certain primes (see 111) in which $R(s)$ remains bounded as $s \to 1$. It follows that the limit as s approaches 1 of a restricted sum of reciprocals of s powers of primes is infinity (see 112). Thus the sum with $s = 1$ diverges, and, hence, there are infinitely many $p \equiv a \pmod{k}$. In fact, the above argument shows that the limit as s approaches 1 of a ratio of sums involving reciprocals of s powers of primes is equal to the reciprocal of φ at k (see 113). Thus the proportion of primes $p \equiv a \pmod{k}$ does not depend upon a and is $1/\varphi(k)$ in the sense of the limit of a ratio (see 113). In general, A is any set of natural numbers and B any subset. Then, the following limit (if it exists): the limit as s approaches 1 of a ratio of sums of reciprocals of s powers (see 114) is called the Dirichlet density of the set B with respect to the set A.

Dirichlet density

On the other hand, if $A(x)$ is defined as the number of elements of A that are less than or equal to x and $B(x)$ is similarly defined, then the limit (if it exists) of the ratio of B as a function of x to A as a function of x as x approaches infinity (see 115) is called the natural density of the set B with respect to the set A.

It is known that if the natural density exists, then the Dirichlet density exists. It is a fundamental problem to decide when the converse is true.

Distributions of other arithmetic functions in arithmetic progressions are handled in analogy with the distributions among the entire set of natural numbers, the L-functions usually replacing the zeta function.

Much attention has been devoted to the question of the uniformity of the distribution of primes in arithmetic progressions. De la Vallée-Poussin provided an expression for π in terms of a function φ and a function of x (see 116). The problem of a uniform error $\Delta(x)$, uniform in the sense that it does not depend on k or a, has played an important role in various questions. Some results are formulated in terms of the function ψ equal to a restricted sum of Λ as a function m (see 117).

As a consequence of Siegel's theorem on the class number, Walfisz proved that if $k \leq (\log x)^N$, there exists a constant c depending only on N such that the absolute value of Δ is less than or equal to a function involving constants A, c, and the logarithm of x (see 118) uniformly in k. Vinogradov used this result in his proof of the 3-prime version of the Goldbach conjecture.

Deeper results have been obtained by the mathematician E. Bombieri, which require sieving methods and estimates for the number of zeros of the L-functions. One result takes the following form. Let $\mathfrak{F}$ equal the maximum over a of Δ (see 119) and $\mathfrak{F}^*$ equal a certain maximum of $\mathfrak{F}$, the range considered being y less than or equal to x (see 120) and $\mathfrak{F}$ equal to a sum of $\mathfrak{F}^*$, the range of summation being k less than or equal to x (see 121). Then for any constant c_1 there exists a constant c_2 such that if $X = x^{1/2}(\log x)^{-c_2}$, the absolute value of $\mathfrak{F}$ is less than or equal to a function of x that includes constants A and c_1 also a functional expression X (see 122). The complexity of the result is in the nature of the subject because the primes are so irregularly distributed.

(108)
$$\begin{cases} (d)\ \displaystyle\sum_{a=1}^{k-1} \chi(a) = \begin{cases} \varphi(k) & \text{if } \chi = \chi_1 \\ 0 & \text{otherwise} \end{cases} \\ (e)\ \displaystyle\sum_{\chi} \chi(a) = \begin{cases} \varphi(k) & \text{if } a \equiv 1 \pmod{k} \\ 0 & \text{otherwise} \end{cases} \end{cases}$$

(109)
$$L(s, \chi) = \sum_{n=1}^{\infty} \frac{\chi(n)}{n^s} \qquad s = \sigma + it$$

(110)
$$\begin{cases} \text{(i) If } \chi = \chi_1 \text{ the principal character, then } L(s, \chi) \text{ differs from } \zeta(s) \text{ by a simple factor.} \\ \text{(ii) If } \chi \neq \chi_1, \text{ then } L(s, \chi) \text{ converges for } \sigma > 0. \\ \text{(iii) } L(s, \chi) = \displaystyle\prod_p \left(1 - \frac{\chi(p)}{p^s}\right)^{-1} \text{ its "Euler product."} \\ \text{(iv) If } \chi \neq \chi_1, \text{ then } L(1, \chi) \neq 0. \end{cases}$$

(111)
$$\sum_{\chi} \chi(a^*) \log L(s, \chi) = \frac{1}{\varphi(k)} \sum_{\substack{p \equiv a \\ \bmod k}} \frac{1}{p^s} + R(s)$$

(112)
$$\lim_{s \to 1} \sum_{p \equiv a \pmod{k}} \frac{1}{p^s} \to \infty$$

(113)
$$\lim_{s \to 1} \frac{\displaystyle\sum_{p \equiv a \pmod{k}} \frac{1}{p^s}}{\displaystyle\sum_p \frac{1}{p^s}} = \frac{1}{\varphi(k)}$$

(114)
$$\lim_{s \to 1} \frac{\displaystyle\sum_{n \in B} \frac{1}{b^s}}{\displaystyle\sum_{a \in A} \frac{1}{a^s}}$$

(115)
$$\lim_{x \to \infty} \frac{B(x)}{A(x)}$$

(116)
$$\begin{cases} \pi(a, k, x) = \dfrac{1}{\varphi(k)} \dfrac{x}{\log x} + \Delta(x) \\ \text{in which} \\ |\Delta(x)| \leq Axe^{-c\sqrt{\log x}} \end{cases}$$

(117)
$$\psi(a, k, x) = \sum_{\substack{n \leq x \\ n \equiv a \pmod{k}}} \Lambda(n)$$

(118)
$$|\Delta(x)| \leq Axe^{-c\sqrt{\log x}}$$

(119)
$$\mathfrak{F}(x, k) = \max_a \Delta(x) \qquad (a, k) = 1$$

(120)
$$\mathfrak{F}^*(x, k) = \max_{y \leq x} \mathfrak{F}(y, k)$$

(121)
$$\mathfrak{F}(x) = \sum_{k \leq x} \mathfrak{F}^*(x, k)$$

(122)
$$|\mathfrak{F}(X)| \leq Ax(\log x)^{-c_1}, \qquad X = x^{\frac{1}{2}} (\log x)^{-c_2}$$

Other applications of L-functions. There is a close connection between class numbers (see above *Algebraic number theory*) and the L-functions of Dirichlet. In fact, let $\mathfrak{J} = \mathfrak{H}(\sqrt{d})$ be a quadratic extension of the rational field $\mathfrak{H}$, with class number h. Consider the case when $d < 0$ and $d \equiv 1 \pmod 4$ [similar results hold when $d \equiv 2, 3 \pmod 4$]. Then Dirichlet proved that the product πh divided by the square root of d is equal to L (see

123) in which $\chi(n) = (d/n)$ is the Kronecker symbol (an extension of the Legendre symbol) and is a character mod $|d|$. The series for $L(1, \chi)$ may be expressed as a finite sum. In the more general case, there is a constant, symbolized by the Greek letter alpha, α, depending on the field K, such that αh equals the residue of ζ at s equals 1 (see 124), the function $\zeta_K(s)$ being the Riemann zeta function defined by ζ_k at s equals a summation over the class $\mathfrak{A}$ of reciprocals expressed in terms of N (see 125) summed over all integral ideals $\mathfrak{A}$. In certain cases $\zeta_K(s)$ can be written as a product of Dirichlet L-functions and its residue computed in terms of L-functions. Siegel proved that for a quadratic field of discriminant d a ratio of logarithms involving α and h converges to 0 (see 126) as $|d| \to \infty$. This was generalized by the German-born U.S. mathematician Richard Brauer in 1947.

The Circle method. The case of the partition function $p(n)$ was previously considered and its generating function determined. It was found that a power series with coefficients $p(n)$ equals a product over positive integers of the function of x, which in turn equals F (see 127).

In 1917, Hardy and the Indian mathematician Srinivasa Ramanujan obtained an asymptotic formula for $p(n)$ and laid the foundations for a method that was further developed by Hardy and Littlewood, which method has since been called the Circle method. This method has been modified and refined by Vinogradov and has been widely used to derive results on additive problems in number theory. The method will be described briefly as it applies to $p(n)$. Using Cauchy's theorem p as a function of n equals an integral over C of a function involving F (see 128), C being a simple closed contour inside the unit circle D, containing the origin. The function $F(x)$ is analytic inside D, but on the boundary of D, $F(x)$ has a dense set of poles. Thus it is not possible to integrate past the poles but only to come close to these. A circle of radius $N < 1$ is chosen and cut in a prescribed manner into arcs, the Farey fractions being used to determine the arcs and the number of arcs depending on N and tending to infinity as N tends to 1. Moreover, the circle approaches the unit circle as $N \to 1$.

The function $F(x)$ is now approximated along each of the arcs. The success of the method depends upon the precision with which $F(x)$ may be approximated. In this special case $F(x)$ satisfies a functional equation, and the approximation is so effective that a rapidly convergent series for $p(n)$ can be derived, as was shown by the German-born U.S. mathematician Hans Rademacher in 1937.

Vinogradov's modification

Vinogradov's modification is as follows: If, for example, the Goldbach problem for three primes be considered, then $r(n)$ may be the number of solutions of n equal to a sum of three primes (see 129). By use of an idea previously stated (see above *Methodology*), it is readily seen that a power series with coefficients r equals a third power of a sum extended over primes (see 130) in which the summation on the right is over primes.

Vinogradov replaces the infinite series with a finite sum and, because x is then no longer required to be in absolute value < 1, replaces x by $\exp(2\pi it)$, and integrates along the unit circle, which is then transformed into the unit interval or indeed any interval of length 1. If the function f equals a restricted sum of exponentials involving p and t (see 131), r as a function of n is an integral from zero to one of the third power of f times an exponential (see 132). The interval is now dissected into subintervals by analogy with the decomposition into Farey arcs. The intervals are in two classes, basic intervals and supplementary intervals. The basic intervals contribute the principal term. If the basic intervals are denoted by I and the supplementary intervals by S, r as a function of n is a sum of two integrals, one extended over I and one extended over S, each with integrands of the type: third power of f times an exponential (see 133). After suitable approximation on I and S, respectively, the conclusion is reached that r as a function of n is a sum of two terms C and λ all multiplied by a function involving the third power of the logarithm of n (see 134) when $C(n)$ is positive if n is odd and $\lambda(n) \to 0$ as $n \to \infty$. The first term comes from I and the second from S.

$$\frac{\pi h}{\sqrt{d}} = L(1, \chi) \tag{123}$$

$$\alpha h = \text{residue of } \zeta_K(s) \quad \text{at } s = 1 \tag{124}$$

$$\zeta_k(s) = \sum_{\mathfrak{A}} \frac{1}{N(\mathfrak{A})^s} \tag{125}$$

$$\frac{\log \alpha h}{\log \sqrt{|\alpha|}} \to 0 \tag{126}$$

$$\sum_{n=0}^{\infty} p(n)x^n = \prod_{k=1}^{\infty} (1 - x^k)^{-1} = F(x) \tag{127}$$

$$p(n) = \frac{1}{2\pi i} \int_C F(x) x^{-n-1}\, dx \tag{128}$$

$$n = p_1 + p_2 + p_3 \tag{129}$$

$$\sum_{n=0}^{\infty} r(n)x^n = \left(\sum_p x^p\right)^3 \tag{130}$$

$$f(t) = \sum_{p \leq N} e^{2\pi i p t} \tag{131}$$

$$r(n) = \int_0^1 f^3(t) e^{-\pi i n t}\, dt \tag{132}$$

$$r(n) = \int_I f^3(t) e^{-2\pi i n t}\, dt + \int_S f^3(t) e^{-2\pi i n t}\, dt = I_1 + I_2 \tag{133}$$

$$r(n) = (C(n) + \lambda(n)) \frac{n^2}{\log^3 n} \tag{134}$$

$$n = x_1^k + \cdots + x_s^k \tag{135}$$

$$\begin{cases} \text{If } s \geq [10k^2 \log k], \text{ then} \\ r(n) = C(n)\, \gamma(s, k) + \lambda(n) n^{s/(k-1)} \end{cases} \tag{136}$$

It follows that if n is sufficiently large then $r(n) > 0$. The same method has been used to obtain a variety of results. For example, if $r(n)$ is the number of solutions of an equation giving n as a sum of s terms of the type x_1 raised to the kth power (see 135) if s is greater than or equal to a specific function of k involving a logarithm, then r as a function of n is a sum of two terms, one a product of C times γ, one a product of λ and a certain power of n (see 136), in which $C(n)\gamma(s, k) > 0$ and $\lambda(n) \to 0$ as $n \to \infty$. This was proved by Vinogradov.

In particular, the English mathematician Harold Davenport has proved that if $k = 4$, then $r(n) > 0$ if and only if $s \geq 16$. Yury Vladimirovich Linnik (Soviet) and George Leo Watson (English) have proved that for $k = 3$, $r(n) > 0$ if $s \geq 7$.

Although the Goldbach conjecture for even numbers has not been proved, it can be shown using the circle method that almost all even numbers are sums of two primes. This means that if $E(N)$ is the number of even integers, m, that are not the sum of two primes and $m \leq N$, then the limit, as N approaches infinity, of the ratio $E(N)/N$ is zero.

Exponential sums. It is significant that both in the distribution of primes and in additive problems, as well as in numerous other questions, the basic problems frequently reduce to the problem of estimating the size of the exponential sum E as a function of x. This is a sum over all values of n less than or equal to N of an exponential involving f (see 137), in which $f(y)$ is some function of y. In the case of the Goldbach problem, f as a function of y takes on one of two values 0 or 1, depend-

ing on whether y is not or is a prime (see 138). In the case of the zeta function $\zeta(s)$, $(s = \sigma + it)$, the estimates of $\zeta(s)$ can be reduced to a function $E(x)$ in which f as a function of y is expressed in terms of the logarithm of y (see 139). It is to be noted that because each of the terms in the sum has absolute value 1, then trivially the absolute value of E is less than or equal to N (see 140). The basic problem is to show that under suitable conditions on $f(y)$ and on x, the limit as N approaches infinity of the ratio of the absolute value of E to N is 0 (see 141), and the rate at which this takes place is an important part of the problem. A typical result due to Korobov and Vinogradov is the following: If a number of conditions are satisfied that involve r, w, t, M, M', and θ, then the absolute value of a sum of exponentials of the logarithm of m and w is less than or equal to a certain power of M (see 142). The complexity of the statement of the result is fairly typical and it is noted once again that this intricacy is in the nature of the subject.

Elliptic, theta, and modular functions. In addition to analysis of the partition function, Euler considered a function that makes $\theta(x)$ equal a product over all positive integers of the difference between 1 and integer powers of x (see 143) and proved the so-called pentagonal number theorem—*i.e.*, θ as a function of x can be found by calculation to be 1 plus an infinite sum of positive and negative terms each being powers of x (see 144)—the name stems from the relation of the exponents to the pentagon.

More important, however, is the fact that this function was the first example of a theta function. From this result and other related ones, interesting identities and relations involving partitions are proved. Theta functions appeared in the works of Bernoulli (1713), in Fourier's *Théorie analytique de la chaleur* (1822; *The Analytical Theory of Heat*, 1878), but were only systematically introduced and studied by the German mathematician Karl Gustav Jacob Jacobi in *Fundamenta nova Theoriae Functionum Ellipticarum* (1829; "New Foundations of the Theory of Elliptic Functions"). There are four types of theta functions; one is defined by θ as a function of z and q is equal to 1 plus twice the sum over all positive integers of a product involving a cosine function (see 145) with $q = \exp(\pi i\tau)$, $\operatorname{im} \tau > 0$, which for $z = \pi$ takes the form θ as a function of π and q is equal to 1 plus twice the infinite sum over all positive integers of q raised to the squared power of an integer (see 146), a function useful for studying sums of squares.

Using the theory of θ-functions and their transformations, Jacobi proved several results on the numbers of representations of integers by sums of squares. In particular, he proved that the number $r_4(n)$ of representations of n as a sum of four squares is (i) 8 times the sum of the odd divisors of n when n is odd and (ii) 24 times the sum of the odd divisors of n when n is even.

Jacobi also proved the identity an infinite product extending over all positive integers is equal to an infinite sum extending over both negative and positive integers, the product expressed in three factors that involve x, z, and an integer, the sum involving two factors expressed in terms of z, x, and an integer (see 147) from which a number of results could be derived.

The introduction of modular forms and functions began a new era in the interplay between number theory and analysis.

Modular transformations

If $f(z)$ be a function of a complex variable z that is meromorphic in the upper half plane, then if a, b, c, d are integers with $ad - bc = 1$, the transformation w is a rational function of z with coefficients a, b, c, and d (see 148); this is called a modular transformation. The set of modular transformations forms a group generated by the particular transformations w_1 and w_2; these are each simple functions of z (see 149), and a modular transformation takes the upper half plane into itself. If f, as a function of z, is a factor raised to the $2k$ power times f, as a function of a rational function of z, the rational function involving coefficients a, b, c, and d that satisfy a simple condition (see 150), then $f(z)$ is called a modular form of weight (or dimension) $2k$. Examples of such modular forms are the Eisenstein series, G_k as a function of z, a restricted sum over two integer valued variables m and n of a simple function of two variables (see 151), in which m, n range over all integers except the pair (0, 0). The form Δ, as a function of τ, which is a difference of two terms, one term involving G_2, a second term involving G_3 (see 152), plays an important role in the theory. It is a modular form of weight 12. It turns out that Δ, as a function of τ, is a factor involving π and q multiplied by an infinite product over positive integers of a simple function of q and an integer (see 153), in which $q = \exp(2\pi i\tau)$. On the other hand, the function

(137) $$E(x) = \sum_{n \leq N} e^{2\pi i f(n)x}$$

(138) $$f(y) = \begin{cases} 0 & \text{if } y \text{ is not a prime} \\ 1 & \text{if } y \text{ is a prime} \end{cases}$$

(139) $$f(y) = t \log y$$

(140) $$|E(x)| \leq N$$

(141) $$\lim_{N \to \infty} \frac{|E(x)|}{N} = 0$$

(142) $$\begin{cases} \text{Let } r \geq 46, \quad 0 \leq w \leq 1, \quad t \geq 1 \\ M \leq M' \leq 2M, \quad \theta = (266{,}000r^2)^{-1} \\ t^{1/r} \leq M \leq t^{1/(r-1)} \quad \text{then} \\ \left| \sum_{m=M}^{M'} e^{ti \log(m+w)} \right| \leq M^{1-\theta} \end{cases}$$

(143) $$\theta(x) = \prod_{n=1}^{\infty} (1 - x^n)$$

(144) $$\begin{cases} \theta(x) = 1 + \sum_{n=1}^{\infty} (-1)^n \left(x^{\frac{n(3n-1)}{2}} + x^{\frac{n(3n+1)}{2}} \right) \\ \quad = 1 - x - x^2 + x^5 + x^7 - x^{12} - x^{15} + \cdots \end{cases}$$

(145) $$\theta(z, q) = 1 + 2 \sum_{n=1}^{\infty} q^{n^2} \cos 2nz$$

(146) $$\theta(0, q) = 1 + 2 \sum_{n=1}^{\infty} q^{n^2}$$

(147) $$\prod_{n=1}^{\infty} (1 - x^{2n})(1 + zx^{2n+1})(1 + z^{-1}x^{2n-1}) = \sum_{k=-\infty}^{\infty} z^k x^{k^2}$$

(148) $$w = \frac{az + b}{cz + d}$$

(149) $$\begin{cases} w_1 = z + 1 \\ w_2 = -\dfrac{1}{z} \end{cases}$$

(150) $$f(z) = (cz + d)^{-2k} f\left(\frac{az + b}{cz + d}\right) \qquad (ad - bc = 1)$$

(151) $$G_k(z) = \sum_{m,n}{}' \frac{1}{(mz + n)^{2k}}$$

(152) $$\Delta(\tau) = (60G_2)^3 - 27(140G_3)^2$$

(153) $$\Delta(\tau) = (2\pi)^{12} q \prod_{n=1}^{\infty} (1 - q^n)^{24}$$

η, as a function of τ, that is the $\frac{1}{12}$ power of q times an infinite product extended over all positive integers of a simple function of q and an integer (see 154), appears in Riemann's works, and its properties were elucidated by Dedekind. It plays a fundamental role in the problem of partitions, and its relation with $\Delta(\tau)$ is immediate.

The function j, as a function of τ, that is equal to a ratio of a third power involving g_2 to Δ, as a function of τ (see 155), is a modular form of weight 0; it is invariant under the complete modular group. Moreover, if $f(\tau)$ is any function meromorphic in the upper half plane and is a modular form of weight 0, then $f(\tau)$ is a rational function of $j(\tau)$; *i.e.*, the ratio of two polynomials in $j(\tau)$.

It is most remarkable that Abelian extensions of an imaginary quadratic field can be obtained as subfields obtained by adjoining values of $j(\tau)$ to the quadratic field. This settled the *Jugendtraum* of Kronecker. Abelian extensions of the rational field $\mathfrak{Q}$ are subfields of fields obtained by adjoining to $\mathfrak{Q}$ values of the exponential function e^x. Thus the relation between modular forms and algebraic number fields is very profound. Mention should be made of the connection between the zeta functions of algebraic number fields and elliptic functions, connections that were discovered by Kronecker in 1875. Extensions of these results were made by Erich Hecke (German-born), Carl L. Siegel (U.S.), and others.

These connections were used by the mathematician K. Heegner in 1952 and later by the mathematician Harold M. Stark in 1967 and others to determine all imaginary quadratic fields that have unique factorization. The problem had remained unsolved since 1933 when the mathematicians H. Heilbronn and E. Linfoot showed that there could be at most 10 such fields, nine of which were identified.

Another connection between modular forms and number theory was discovered by Hecke as follows: $f(\tau)$ may be expanded in a Fourier series in q such that f, as a function of q, is a power series with coefficients a_n, q being a complex exponential with variable τ (see 156). If $a_0 = 0$, then $f(\tau)$ is called a cusp form. Hecke proved that if $f(\tau)$ is a cusp form of weight $2k$, then the absolute value of a_n is less than or equal to a constant times the k power of n (see 157).

Ramanujan investigated in great detail the function that is a sum over positive integers of the function τ times q raised to an integer power and that is equal to q times a product of terms involving q and an integer (see 158). He conjectured that if $(m, n) = 1$, then $\tau(m)\tau(n) = \tau(mn)$, a fact proved by the U.S.-born mathematician Louis Joel Mordell. He was led to the conjecture that if p is prime, then the absolute value of τ as a function of p is less than a simple function of p (see 159). This conjecture remains unproved though a proof was reported in the 1970s but not yet confirmed.

(154) $$\eta(\tau) = q^{1/12} \prod_{n=1}^{\infty} (1 - q^n)$$

(155) $$j(\tau) = \frac{(720 G_2(\tau))^3}{\Delta(\tau)}$$

(156) $$f(q) = \sum_{n=0}^{\infty} a_n q^n \qquad q = e^{2\pi i \tau}$$

(157) $$|a_n| \leq A n^k$$

(158) $$\sum_{n=1}^{\infty} \tau(n) q^n = q \prod (1 - q^n)^{24}$$

(159) $$|\tau(p)| < 2p^{\frac{11}{2}}$$

(160) $$f(s) = 1 - \frac{1}{2^s} + \frac{1}{3^s} - \cdots$$

(161) $$\left| \sum_{n \leq x} \mu(n) \right| \leq A x^{1/2 + \varepsilon}$$

(162) $$K(a, b) = \sum_{\substack{n \bmod p \\ n\bar{n} \equiv 1 (p)}} e^{\frac{2\pi i}{p}(an + b\bar{n})}$$

(163) $$|K(a, b)| \leq A\sqrt{p}$$

RESULTS OBTAINABLE FROM ELEMENTARY METHODS

A great deal of attention has been paid to the question of whether the results of analytic number theory can be derived without the use of complex analytic machinery, especially such results as those of the Goldbach and Waring problems, the statements of which do not involve analytic concepts. In fact, a number of results have been proved without complex function theory. Several results that have been proved by elementary methods are enumerated below.

(i) The prime number theorem—proved by the Norwegian-born U.S. mathematician Atle Selberg and the Hungarian mathematician Paul Erdös with later refinements by Bombieri.

(ii) Solution to the Waring problem obtained by the Soviet mathematician Yury Vladimirovich Linnik.

(iii) Dirichlet's theorem on primes in an arithmetic progression, proved by the German-born Canadian mathematician Hans Zassenhaus and Atle Selberg.

(iv) Calculation of the number of lattice points inside contours, performed by Vinogradov.

(v) Siegel's theorem on the class number of quadratic fields, proved by Linnik.

The term elementary is in some respects a misnomer. The methods are elementary primarily in avoiding the use of such results as Cauchy's theorem of complex function theory. In fact, they tend to be rather intricate.

Furthermore, there has also been a trend to avoid the use of analysis even in the statements of some of the theorems that on the face of it are analytic in character. For example, in the prime number theorem the function $\log x$ is replaced by $\ell(n) = \Sigma 1/k$, the summation for $k \leq n$, and then $\pi(n)/n$ is compared with $\ell(n)$: Thus, transcendental concepts are entirely removed.

SOME UNSOLVED PROBLEMS OF ANALYTIC NUMBER THEORY

1. The most famous is perhaps the Riemann hypothesis. An elementary formulation follows: If $s = \sigma + it$ and f as a function of s is equal to 1 plus negative and positive reciprocals of integers raised to the s power (see 160), which converges for $\sigma > 0$, then the Riemann conjecture is that if $f(s) = 0$, $0 < \sigma < 1$, then $\sigma = \frac{1}{2}$. Again, if $\mu(n)$ is the Möbius function described above, then the Riemann hypothesis is equivalent to the statement that there exists a constant A such that the absolute value of the function μ with integer arguments is less than or equal to a constant A times the $\frac{1}{2} + \varepsilon$ power of x (see 161), in which ε is any positive number.

2. The Goldbach conjecture for even integers: Every even integer $n \geq 4$ is the sum of two primes.

3. Hecke's hypothesis: If χ is a real character mod k, then $L(s, \chi) \neq 0$ if $0 < s < 1$.

4. Better estimates for exponential sums. In the case of K as a function of a and b—that is, a sum over integers defined in terms of p, summands being of exponential type (see 162)—the mathematician André Weil proved the Riemann hypothesis for function fields and using it showed that the absolute value of K is less than or equal to A times the square root of p (see 163).

Such sharp estimates in general are much to be desired.

5. The least quadratic non-residue. If p is a prime and $n(p)$ is the least quadratic non-residue, then the Riemann hypothesis implies, as shown by the mathematician N. Ankeny, that $|n(p)| \leq A \log^2 p$. A proof without hypothesis is yet to be found. The mathematician Burgess has shown that $|n(p)| \leq A p^{1/4\sqrt{e}}$.

6. If $d > 0$, and $h(d)$ the class number of the quadratic field $\mathfrak{Q}(\sqrt{d})$, then is $h(d) = 1$ infinitely often as Gauss conjectured?

7. More generally, are there infinitely many algebraic number fields that have unique factorization; *i.e.*, for which the class number is 1?

8. Are there infinitely many primes of the form $n^2 + 1$?

9. Artin's conjecture. Given an integer g, are there infinitely many primes p of which g is a primitive root?

10. Ramanujan conjectured that in $q(\prod_{n=1}^{\infty}(1-q^n))^{24} = \Sigma\tau(k)q^k$, the summation for $k = 1$ to ∞, if p is a prime, then $|\tau(p)| \leq 2p^{11/2}$. The mathematician Lehmer conjectured that for every n, $\tau(n) \neq 0$.

11. More general conjectures on coefficients of modular forms of a given weight.

12. Improvement of the error term in the prime number theorem is being sought.

13. Proofs of congruence properties of partitions; *e.g.*, Ramanujan conjectured if $24k \equiv 1 \pmod{m}$, in which $m = 5^a 7^b 11^c$, the function p, with argument expressed in terms of m, n, and λ, is equivalent to $0 \pmod{m}$ (see 164). Some special cases of this are known. For example, the mathematician A.O.L. Atkin has proved some results for $m = 11^c$. The mathematician Sarvadaman Chowla has pointed out that the conjecture is false for $m = 7^3$. The question remains to determine just what congruence properties are true.

14. If $d(n)$ is the number of divisors of n, and it is known that D, as a function of x, is a sum of two terms plus Δ, as a function of x, the first term involving the product of x and $\log x$, the second term involving γ and x (see 165), in which $\Delta(x)$ is an error term. It may be asked whether there exists a constant A such that the absolute value of Δ is less than or equal to A times the $\frac{1}{4} + \varepsilon$ power of x (see 166). A similar conjecture for the number of lattice points in a circle.

15. Do there exist infinitely many Fermat primes; *i.e.*, primes of the form $2^{2^k} + 1$?

16. Do there exist infinitely many regular primes; *i.e.*, primes p such that if h is the class number of the field $\mathfrak{Q}(\exp 2\pi i/p)$ then p is not a factor of h? This has a close connection with Fermat's last theorem.

17. Waring's problem. If $n = x_1^k + \cdots + x_s^k$, then it is known that if $s \geq 2^k + 1$, there exists n_0, such that if $n > n_0$ the equation has a solution in integers $x_1, x_2, \cdots, x_s$. What is the least value of s for which this statement remains valid; *e.g.*, is $s \geq 4k$ sufficient?

18. Are there infinitely many twin primes; *i.e.*, primes p such that $p + 2$ is also prime? (R.G.A.)

Geometric and probabilistic number theory

GEOMETRIC NUMBER THEORY

Geometric number theory, or, as it is sometimes misleadingly called, the geometry of numbers, is a branch of number theory that can be developed by the use of certain geometric methods. It centres on the arithmetical theory of quadratic and higher forms and the problems of the approximation of real numbers by rational numbers. If integers a, b, c satisfy the inequality $ac > b^2$, the (positive definite) binary quadratic form in x and y with coefficients a, b, and c (see 167) takes the value zero when the variables x, y are both zero, but takes only positive integral values for other integral values of the variables. A question of interest is what can be said about the positive integers that will be represented by the form, and, in particular, what will be the smallest positive value taken by the form for integral values of the variables. Lagrange developed an arithmetical technique for answering such questions in 1773, and, in particular, he proved that it is always possible to choose integers x, y such that a quadratic in x and y satisfies a double inequality, being greater than zero and less than a function of its coefficients (see 168). By 1831, the theory had been developed further, and the mathematician L.A. Seeber published an elaborate account of the corresponding arithmetical theory for ternary quadratic forms. In his review of Seeber's book, Gauss introduced a geometric point of view that greatly simplified the proofs of Seeber's results and led to a proof of an important conjecture of Seeber's. If Q as a function of x, y, and z, a ternary quadratic with six coefficients (see 169), is positive definite—*i.e.*, if it takes no negative values, and only takes the value zero when the variables x, y, z are all zero—then it can be expressed, in many ways, as the sum (see 170) of the squares of the three linear forms that are written ξ_1, ξ_2, ξ_3 (see 171). The coefficients will satisfy a determinental identity involving a third-order determinant that is positive with elements symmetric about the diagonal (see 172). Gauss recognized that as x, y, z vary independently over the integers, the corresponding points (ξ_1, ξ_2, ξ_3) will vary over the points of a regular discrete array of points, the points of a lattice. Indeed, the lattice is simply obtained by applying the linear transformation (see 173) already defined (see 171) to the standard lattice Λ_0 of all points (x, y, z) with integral coordinates. A previous formula (see 170) then expresses the value of the quadratic form when the variables x, y, z are integers, as the square of the Euclidean distance of the corresponding lattice point (ξ_1, ξ_2, ξ_3) from the origin. This very simple idea enabled Gauss to re-interpret much of Seeber's work geometrically and to simplify and extend it.

(164) $$p(mn + \lambda) \equiv 0 \pmod{m}$$

(165) $$D(x) = x \log x + (2\gamma - 1)x + \Delta(x)$$

(166) $$|\Delta(x)| \leq Ax^{1/4+\varepsilon}$$

(167) $$ax^2 + 2bxy + cy^2 \qquad (ac > b^2, \qquad a > 0)$$

(168) $$0 < ax^2 + 2bxy + cy^2 < [\tfrac{4}{3}(ac - b^2)]^{\frac{1}{2}}$$

(169) $$Q(x, y, z) = ax^2 + by^2 + cz^2 + 2fyz + 2gzx + 2hxy$$

(170) $$Q(x, y, z) = \xi_1^2 + \xi_2^2 + \xi_3^2$$

(171) $$\begin{cases} \xi_1 = \alpha_1 x + \beta_1 y + \gamma_1 z \\ \xi_2 = \alpha_2 x + \beta_2 y + \gamma_2 z \\ \xi_3 = \alpha_3 x + \beta_3 y + \gamma_3 z \end{cases}$$

(172) $$\begin{vmatrix} a & h & g \\ h & b & f \\ g & f & c \end{vmatrix} = \begin{vmatrix} \alpha_1 & \beta_1 & \gamma_1 \\ \alpha_2 & \beta_2 & \gamma_2 \\ \alpha_3 & \beta_3 & \gamma_3 \end{vmatrix}^2 > 0$$

(173) $$T{:}(x, y, z) \rightarrow (\xi_1, \xi_2, \xi_3)$$

The lattices introduced here are essentially the same as the lattices used in crystallography but are quite distinct from the abstract mathematical structures that the U.S. mathematician Garrett Birkhoff named lattices later.

Although Dirichlet made some use of these geometrical ideas, the next major development was due to Minkowski, who published his *Geometrie der Zahlen* ("Geometry of Numbers") giving a detailed account of his methods in 1896.

Minkowski's application to a complex body

Minkowski introduced the idea of a convex body. A set K is said to be convex, if whenever A and B are points of K, the line segment AB is also in K. A set K is said to be a convex body if it is closed (*i.e.*, contains all the points of its boundary or surface), bounded (*i.e.*, does not extend to infinity), and contains an inner point (*i.e.*, a point not on its boundary). A set is said to be symmetrical in the origin $O = (0, 0, 0)$ if, with each point A in the set, the reflection A' of A in O is also in the set. In three-dimensional space, Minkowski's first fundamental theorem, in its simplest form, asserts that if a convex body is symmetrical in the origin and has volume 8 or more units, then it contains, in addition to the origin, a symmetrical pair of points (x, y, z), $(-x, -y, -z)$ with integral coordinates, not all zero. On applying the transformation T (see 171, 173) the general form is obtained: If a convex body K is symmetrical in the origin and has volume at least eight times the absolute value of

the determinant Δ—that is, a determinant with columns composed of $\alpha_i\beta_j\gamma_k$ (see 174)—giving the volume of the unit cell of the lattice, then K contains, in addition to the origin, a symmetrical pair of points (ξ_1, ξ_2, ξ_3), $(-\xi_1, -\xi_2, -\xi_3)$ of the lattice defined by the transformation (see 171).

Applying this result to the cube of points (ξ_1, ξ_2, ξ_3), satisfying the condition that the absolute value of a typical variable ξ_i is less than or equal to the one-third power of the absolute value of Δ (see 175), Minkowski obtained Dirichlet's 1842 result on linear forms: given the forms referred to (see 171), there will always be integers x, y, z, not all zero satisfying the stated inequalities (see 175). This result leads quite simply to the result that, if θ, φ are any real irrational numbers, there are infinitely many pairs of rational numbers u/w, v/w, u, v, w being integers that are good simultaneous approximations to θ and φ in that the absolute differences between θ and the ratio u to w and between φ and the ratio v to w are each less than or equal to a simple function of w (see 176).

Minkowski obtained many other important results, in particular, his second fundamental theorem, and his improvements of the French mathematician Charles Hermite's estimate for the arithmetic minima of positive definite quadratic forms. To discuss these results and the subsequent development of the subject the limitations of three-dimensional space must be discarded.

The Cartesian coordinates

The 17th-century French mathematician René Descartes had revolutionized three-dimensional geometry by introducing a rectangular Cartesian coordinate system, so that each point of space is represented by a set of three coordinates, and by showing that the whole of geometry could be reduced to the algebra and analysis of the sets of coordinates. For example, if a and b are points with coordinates (a_1, a_2, a_3) and (b_1, b_2, b_3), the distance between these is the square root of the sum of squares of differences between b_1 and a_1 (see 177), and the points on the line segments joining a to b are those with coordinates forming a vector with three components, each of which is expressed in terms of θ and $\{a_k\}$ and $\{b_k\}$ (see 178) when $0 \leq \theta \leq 1$. Furthermore, $\mathbf{a} = (a_1, a_2, a_3)$, $\mathbf{b} = (b_1, b_2, b_3)$ are regarded as vectors, the distance between the points a and b is the distance between vectors $\mathbf{b}$ and $\mathbf{a}$ and is the square root of the sum of squares of differences of components (see 179) of the vector $\mathbf{b} - \mathbf{a}$, and the points on the line segments joining $\mathbf{a}$ to $\mathbf{b}$ are $(1 - \theta)\,\mathbf{a} + \theta\mathbf{b}$, with $0 \leq \theta \leq 1$.

When vector algebra was developed, partly as a result of the Irish mathematician Sir William Rowan Hamilton's attempts to develop applied mathematics in terms of quaternions, there were immediate advantages in developing an algebra for vectors with any arbitrary fixed number n of components. The basic formula for combination of vectors is that if $\mathbf{a} = (a_1, a_2, \cdots, a_n)$, $\mathbf{b} = (b_1, b_2, \cdots, b_n)$ are vectors, and λ, μ are scalars (in this context, just another name for real numbers), then $\lambda\mathbf{a} + \mu\mathbf{b}$ denotes the vector with n components each of which is expressed in terms of λ, μ and a_k and b_k with length of the vector $\mathbf{a}$ determined in the usual way (see 180). The geometrical term length has already been used to describe a purely algebraic object, the square root of the sum of the squares of the components of a vector, which has no realization in ordinary geometry. Mathematicians soon took this process much further, calling the vectors $\mathbf{x}$ that have n components (see 181) the points of n-dimensional Euclidean space, and using geometrical language to describe the vector algebra. For example, the distance between the points $\mathbf{a}$ and $\mathbf{b}$ is defined to be $\|\mathbf{b} - \mathbf{a}\|$, and the line segment joining the points $\mathbf{a}$ and $\mathbf{b}$ is defined to be the set of points $(1 - \theta)\mathbf{a} + \theta\mathbf{b}$ with $0 \leq \theta \leq 1$. While Descartes regarded the sets of coordinates and their algebra as a means of describing the geometric realities of space, the geometrical language used in the study of n-dimensional Euclidean space E^n is now regarded as describing the realities of the corresponding vector algebra of vectors with n real components.

The set Λ_0 of all points $\mathbf{x} = (x_1, x_2, \cdots, x_n)$ with integral coordinates is taken to be the standard lattice in E^n; the general lattice Λ in E^n is obtained by applying a nonsingular linear transformation to Λ_0, and so is the set of points possessing n coordinates ξ_i (see 182) with ξ_i a linear combination of x_j (see 183), $x_1, x_2, \cdots, x_n$ taking all integral values, and the determinant $\det(\alpha_{ij})$ of the coefficient matrix being nonzero. The determinant $d(\Lambda)$ of Λ is defined to be the absolute value of the determinant of the coefficients (see 184).

Using the natural generalizations of the ideas of a convex body symmetrical in $O = (0, 0, \cdots, 0)$, Minkowski's first theorem takes the form: If a convex body K in E^n is symmetrical in O and has volume $2^n d(\Lambda)$ or more, then K contains, in addition to the origin, a symmetrical pair of points of Λ.

Minkowski's second fundamental theorem is explained by introducing the successive minima of a convex symmetrical body for a lattice. For any set S of points and any positive number λ, the set of all points of the form $\lambda\mathbf{s}$ with $\mathbf{s}$ a point of S is called the expansion of S by the factor λ and is denoted by λS. The expansion of S will be a genuine expansion or enlargement of S from the origin if λ exceeds unity; but it will be a contraction or reduction of S toward the origin if λ is less than unity. The first minimum λ_1 of a convex body K in E^n, symmetrical in O, for a lattice Λ is defined to be the least positive number λ_1 such that the expanded body $\lambda_1 K$ contains points of Λ other that O. More generally, if $1 \leq r \leq n$, the rth minimum λ_r is defined to be the least positive number λ_r such that the set of points of Λ in $\lambda_r K$ does not lie in any substance of E^n of dimension less than r. It follows immediately from these definitions that λ_1, $\lambda_2, \cdots, \lambda_n$ is a non-decreasing sequence of positive numbers. Using $V(K)$ to denote the volume of the convex body, Minkowski's first theorem reduces to the inequality setting the nth power of λ_1 multiplied by $V(K)$ less than or equal to $d(\Lambda)$ times the nth power of 2 (see 185); his second theorem makes the stronger and more sophisticated assertion that there is a similar inequality involving the product of n of the $\{\lambda_k\}$ (see

$$(174)\quad \Delta = \begin{vmatrix} \alpha_1 & \beta_1 & \gamma_1 \\ \alpha_2 & \beta_2 & \gamma_2 \\ \alpha_3 & \beta_3 & \gamma_3 \end{vmatrix}$$

$$(175)\quad |\xi_i| \leq |\Delta|^{\frac{1}{3}}, \qquad i = 1, 2, 3,$$

$$(176)\quad \left|\theta - \frac{u}{w}\right| \leq \frac{1}{w^{3/2}}, \qquad \left|\varphi - \frac{v}{w}\right| \leq \frac{1}{w^{3/2}}$$

$$(177)\quad [(b_1 - a_1)^2 + (b_2 - a_2)^2 + (b_3 - a_3)^2]^{\frac{1}{2}}$$

$$(178)\quad ((1 - \theta)a_1 + \theta b_1, (1 - \theta)a_2 + \theta b_2, (1 - \theta)a_3 + \theta b_3)$$

$$(179)\quad \|\mathbf{b} - \mathbf{a}\| = [(b_1 - a_1)^2 + (b_2 - a_2)^2 + (b_3 - a_3)^2]^{\frac{1}{2}}$$

$$(180)\quad \begin{cases} (\lambda a_1 + \mu b_1, \lambda a_2 + \mu b_2, \cdots, \lambda a_n + \mu b_n) \\ \text{further, the length of the vector } \mathbf{a} \text{ is} \\ \|\mathbf{a}\| = [a_1^2 + a_2^2 + \cdots + a_n^2]^{\frac{1}{2}} \end{cases}$$

$$(181)\quad \mathbf{x} = (x_1, x_2, \cdots, x_n)$$

$$(182)\quad \boldsymbol{\xi} = (\xi_1, \xi_2, \cdots, \xi_n)$$

$$(183)\quad \xi_i = \sum_{j=1}^{n} \alpha_{ij} x_j$$

$$(184)\quad d(\Lambda) = |\det(\alpha_{ij})|$$

$$(185)\quad \lambda_1^n V(K) \leq 2^n d(\Lambda)$$

186). While the first theorem is easy to prove, there is no very simple proof of the second theorem; the simplest is probably that due to the mathematician Harold Davenport (1939), despite the fact that he leaves one not quite obvious step to the reader.

Minkowski's first fundamental theorem can be restated in yet another form. If K is any convex body symmetrical in O, there may exist a largest real number $\Delta(K)$ with the property that all lattices Λ with determinant $d(\Lambda)$ less than $\Delta(K)$ have a pair of symmetrical points, not O, in K. Minkowski's theorem asserts that there is such a number $\Delta(K)$ for each K, and that $\Delta(K) \geq 2^{-n}V(K)$. The number $\Delta(K)$ is called the critical determinant of K, and any lattice Λ with $d(\Lambda) = \Delta(K)$ and with no point, other than O, in the interior of K is called a critical lattice of K. These definitions of critical determinant and critical lattice extend immediately to the case in which K is replaced by a star-body S. A star-body is defined to be a closed set, not necessarily bounded, symmetrical in O, containing O as an inner point, and with the star property that, whenever a point X lies in S, then the line segment OX lies wholly in S. Although Minkowski's theorem does not apply, the critical determinant $\Delta(S)$ will exist as a positive real number or will have the conventional value plus infinity. In order to obtain refined applications to number theory, much effort has been devoted to determining or estimating the critical determinants of a whole range of convex bodies and of star-bodies. Although many of the most interesting star-bodies are of infinite volume, some, and in particular all the convex bodies, have finite volume. Minkowski stated that he had proved that, for any star-body S of finite volume $V(S)$, the critical determinant satisfies the inequality involving a function of n, the function being a sum over positive integers of reciprocals of nth powers of the integers (see 187). Minkowski never published a proof (except in the very special case of a sphere), and the first published proof was accomplished by the mathematician Hlawka in 1944.

(186) $$\lambda_1\lambda_2 \cdots \lambda_n V(K) \leq 2^n d(\Lambda)$$

(187) $$\begin{cases} \Delta(S) \leq \dfrac{1}{2\zeta(n)} V(S) \\ \text{with} \\ \zeta(n) = \displaystyle\sum_{r=1}^{\infty} \frac{1}{r^n} \end{cases}$$

(188) $$\frac{n+2}{2}\left(\frac{1}{\sqrt{2}}\right)^n$$

(189) $$\tau_n \sim \frac{n}{e\sqrt{e}}, \quad \text{as } n \to \infty$$

(190) $$n \log n + n \log \log n + 5n$$

(191) $$n^{\log_2 \log_e n + c}$$

Lattice packing

Minkowski's proof of his first theorem was based on the idea of a lattice packing. If K is a convex body (not necessarily symmetrical in O), then the system of translates $\{K + \mathbf{x}_i\}$, $i = 1$ to $i = \infty$, of K by a sequence of vectors $\{\mathbf{x}_i\}$, $i = 1$ to $i = \infty$, is said to be a packing of K, if no two of the translates have any point in common. The packing is said to be a lattice packing with lattice Λ if the sequence of vectors $\{\mathbf{x}_i\}$, $i = 1$ to $i = \infty$, is an enumeration of the vectors of the lattice Λ. The density of a packing is defined to correspond to the intuitive notion of that proportion of the whole of space covered by the sets of the packing. Minkowski gave a very simple proof, using only a few lines of vector algebra, that if K is a convex body symmetrical in O, and if Λ is a lattice, then the translates by the vectors of Λ of the contracted body $\frac{1}{2}K$ will form a packing, if, and only if, O is the only point of Λ in K. His first result follows immediately from this observation and the intuitively obvious remark that the density of a packing cannot exceed unity. The densities of packings and lattice packings of convex bodies have been extensively studied. The work of Lagrange, Seeber, and Gauss already implies that the densest lattice packings of circular discs in the plane and of spherical balls in 3-space are given by the hexagonal lattice and by the body centred cubic lattice. The densest lattice packings of spherical balls were determined by the Russian mathematicians Aleksandr Nikolayevich Korkin and Zolotarev in 1872 and 1877 for dimensions 4 and 5 and by the mathematician Blichfeldt in 1925, 1926, and 1934 for dimensions 6, 7, and 8. Although the densest lattice packings of spherical balls are not known in higher dimensions, some quite dense examples are known. In particular, the Canadian mathematician Harold Scott MacDonald Coxeter (1951) gives a dense packing of spherical balls in 12 dimensions and the U.S. mathematician Joseph Leech (1967) one in 24 dimensions. The mathematician Conway's study (1969) of the remarkable symmetry group of Leech's packing lattice led him to the discovery of three new large simple groups. Much less is known about general packings that are not necessarily lattice packings. The mathematician L. Fejes Tôth has developed a beautiful and almost completely satisfactory theory of the packing of convex domains in the plane. Very little is known, however, even in 3-space, in which mathematicians have been unable to show that the density of a non-lattice packing of spherical balls cannot exceed the density $0.7404 \cdots$ of the densest lattice packing, despite convincing physical evidence that this must be true. Blichfeldt in 1914 showed that in n-dimensions the density of a packing of spherical balls cannot exceed the bound composed of the product of one-half the quantity $n + 2$ multiplied by the nth power of the reciprocal of the square root of 2 (see 188), a bound that has only been very slightly improved since by the mathematician Rogers (1958). A very elementary argument shows that spheres, or indeed any symmetrical convex bodies, can be packed with density at least $(\frac{1}{2})^n$, and this lower bound for the density has only been slightly improved by the mathematician Schmidt (1963). The factor of ignorance between the upper and lower bounds for the density of the densest packing, or lattice packing, of spherical balls remains essentially of the order $(\sqrt{2})^n$.

Lattice coverings

There is a parallel theory of coverings and lattice coverings of space with convex bodies. The condition that no two translates have any common point is dropped and replaced by the condition that each point of space is in at least one of the translates. The density of a covering is now the average number of times a point of space is covered by the system of translates. The thinnest, or least dense, lattice covering of n-dimensional space by spherical balls has been determined for $n = 2$ by the mathematician Kerschner (1939), for $n = 3$ by the mathematician Bambah (1954), and for $n = 4$ by the mathematicians Delauney and Rushkov (1963). Tôth has solved the general covering problem for symmetrical convex plane domains; but again there are virtually no precise results for general coverings in three or more dimensions. Coxeter, Few, and Rogers (1959) have shown that coverings of n-dimensional space by spherical balls must have density that is at least a certain bound τ_n; τ_n becomes asymptotically equal to a constant times n, as n approaches infinity, the constant being the reciprocal of the product of e and the square root of e (see 189). Rogers (1957, 1959) has shown that, for any convex body K, in n-dimensional space, there is a covering with K with density not greater than an expression composed of n times the logarithm of n and the logarithm of the logarithm of n (see 190) and that there is a lattice covering with K with density not greater than an expression involving n raised to a power, the log to the base two of the natural logarithm of n (see 191). Rogers' results, however, are nonconstructive, and there is no effective method of finding coverings with densities that are at all reasonably small.

The results about lattice packings of spherical balls can all be translated to yield results about the arithmetical minima of positive definite quadratic forms. The n linear forms are considered, each ξ_i being a linear combination of n variables such as x_j (see 192) with determinant having absolute value written Δ and being composed of coefficients $\alpha_{i,j}$ (see 193). For each n, there will be a least positive number γ_n such that the inequality, the sum of the squares of the $\{\xi_k\}$ being less than or equal to γ_n times a power of Δ (see 194), always has a solution in integers $x_1, x_2, \cdots, x_n$, not all zero. This constant γ_n was discussed by Hermite (1850) and is sometimes known as Hermite's constant. By the results quoted for the closest lattice packings of spherical balls, its precise value is known up to $n = 8$; for large values of n, it lies between the asymptotic bounds $n/(2\pi e)$ and $n/(\pi e)$. Many results of geometric number theory can be stated in terms of such inequalities. An early result of the Russian mathematician Andrey Andreyevich Markov (1878) asserts that integers x_1, x_2, not both zero, can be found to satisfy the absolute value of the product of ξ_1 and ξ_2 being less than or equal to one-third of Δ (see 195), unless the ratios α_{11}/α_{12}, α_{21}/α_{22} belong to a countable sequence of pairs of conjugate quadratic irrationals, and that in these cases only slightly weaker inequalities hold. Similar results, involving only a finite sequence of exceptions, have been obtained by Markov (1903) and by the mathematician Venkov (1945) for the quadratic form $|\xi_1^2 - \xi_2^2 - \xi_3^2|$ and by Davenport (1938, 1943) and Swinnerton-Dyer (1971) for the product of three linear forms (see 196). It is conjectured that for any $\varepsilon > 0$, it is always possible, for each choice of the $+$ and $-$ signs, to find integers x_1, x_2, x_3, x_4, x_5, not all zero, to satisfy the absolute value of certain sums and differences of squares of ξ_k being less than ε times a power of Δ (see 197), but this conjecture seems to be far beyond the reach of known methods. The best result in this direction is due to combined work of the mathematicians Bryan John Birch (English), Harold Davenport, and Ridout (1958) and asserts the corresponding result for indefinite quadratic forms in 21 or more variables.

(192) $$\xi_i = \sum_{j=1}^{n} \alpha_{ij} x_j$$

(193) $$\Delta = |\det(\alpha_{ij})|$$

(194) $$\xi_1^2 + \xi_2^2 + \cdots + \xi_n^2 \leqslant \gamma_n \Delta^{2/n}$$

(195) $$|\xi_1 \xi_2| \leqslant \tfrac{1}{3}\Delta$$

(196) $|\xi_1 \xi_2 \xi_3|$ Davenport (1939) has also discussed the form $|\xi_1|(\xi_2^2 + \xi_3^2)$ and Oppenheim (1929, 1931) the forms $|\xi_1^2 + \xi_2^2 - \xi_3^2 - \xi_4^2|$ and $|\xi_1^2 - \xi_2^2 - \xi_3^2 - \xi_4^2|$.

(197) $$|\xi_1^2 - \xi_2^2 \pm \xi_3^2 \pm \xi_4^2 \pm \xi_5^2| < \varepsilon \Delta^{\frac{2}{5}}$$

(198) $$\Delta(K \cap (2\mathbf{a} - K)) \leqslant d(\Lambda)$$

(199) $$\xi_i = \sum_{j=1}^{n} \alpha_{ij} x_j, \qquad i = 1, 2, \cdots, n$$

(200) $$\Delta = \det(\alpha_{ij})$$

(201) $$|(\xi_1 + c_1)(\xi_2 + c_2) \cdots (\xi_n + c_n)| \leqslant (\tfrac{1}{2})^n |\Delta|$$

(202) $$|(\xi_1 + c_1)(\xi_2 + c_2)| < \tfrac{1}{128} |\Delta|$$

Nonhomogeneous forms

The examples discussed so far all involve homogeneous forms. When lattice covering problems are expressed in terms of the theory of forms, they involve nonhomogeneous forms. There are a number of further results in geometric number theory that are of this nonhomogeneous nature. One of the few general results of the subject is the remarkable result of the mathematician Macbeath (1952) that: if a closed convex set K in n-dimensional space contains no complete straight line but does contain in its interior a point of a lattice Λ, then there is a point $\mathbf{a}$ of Λ that is reasonably close to the boundary of the convex set K, in that, the volume $V(K \cap (2\mathbf{a} - K))$ of the intersection $K \cap (2\mathbf{a} - K)$ of the convex body K with its reflection $2\mathbf{a} - K$ in the point $\mathbf{a}$ does not exceed $2^n d(\Lambda)$; indeed, Macbeath proves the stronger inequality involving Δ, K, the vector $\mathbf{a}$, and $d(\Lambda)$ (see 198). A much more special nonhomogeneous problem stems from the work of Minkowski, who proved, in the case $n = 2$, and who conjectured, in general, that: given linear forms establishing ξ_i as a linear combination (see 199) of integral variables $x_1, x_2, \cdots, x_n$ with Δ being the determinant of the coefficients (see 200) and given real numbers $c_1, c_2, \cdots, c_n$, then there are integral choices for the variables that ensure that the absolute value of the product of n factors, each the sum of ξ_i and c_i, is less than or equal to the nth power of ½ times the absolute value of Δ (see 201). The conjecture has been proved by R. Remak (1923) in the case $n = 3$ and by Freeman J. Dyson (1948) in the case $n = 4$. B.F. Skubenko (1937) has proved it for $n = 5$. The mathematician N. Tschebotareff (1940), in the general case, proved that weaker inequality with $(\tfrac{1}{2})^n|\Delta|$ replaced by $(\tfrac{1}{2})^{n/2}|\Delta|$. The problem of deciding whether or not Minkowski's conjecture is true or false remains one of the most challenging problems of the subject. Davenport (1950), in the case $n = 2$, has proved a partial converse to Minkowski's result. He proves that, if the forms ξ_1, ξ_2 are given, then constants c_1, c_2 can be found, namely, the absolute value of products of two sums, each of a ξ_i with a c_i, is less than the reciprocal of 128 times the absolute value of Δ (see 202), has no solution in integers x_1, x_2. This enabled Chatland and Davenport (1950) to show that Euclid's algorithm holds in no real quadratic field beyond the last known example (that generated by $\sqrt{73}$). A little later Davenport (1950), by a refinement of his method, showed that Euclid's algorithm holds only a finite number of cubic fields with negative discriminant, and only in a finite number of complex quartic fields with complex conjugate fields.

Critical determinants and critical lattices

The German-born mathematician Kurt Mahler has investigated the problems of geometric number theory from a general point of view. The formal definitions of critical determinant and critical lattice are his. By regarding the lattices in n-dimensional space as points of a larger space of n^2 dimensions and establishing the compactness of certain sets of lattices, Mahler (1946) was able to show that every star-body with finite critical determinant has at least one critical lattice. This general result at once enabled many difficult special results to be proved by rather simpler methods. Mahler's work also led to a much clearer understanding of the circumstances under which one can assert that infinitely many points of a lattice lie in a star-body. In the case of a convex body, symmetrical in O, Swinnerton-Dyer (1953) established the useful result that a critical lattice must have at least $n(n + 1)$ points on the boundary of the convex body.

The classical applications of geometric number theory were mainly to the theory of algebraic numbers and included Dirichlet's theory of units and Minkowski's proof that the absolute value of the discriminant of an algebraic number field exceeds unity. Davenport's contribution to the problem of Euclid's algorithm has already been mentioned. Two striking applications are now mentioned. Davenport, partly in collaboration with Birch and Donald J. Lewis (U.S.), has used results of John William Scott Cassels (English) and Minkowski's successive minima results together with an extension of the Hardy–Littlewood–Vinogradov method to obtain striking new results about the values taken by quadratic and cubic forms with sufficiently many variables.

Schmidt (1971) has used results of Mahler on the geometric number theory of compound convex bodies, to-

gether with the method of Roth, to prove that if $\theta_1, \theta_2, \cdots, \theta_n$ are algebraic numbers and $1, \theta_1, \theta_2, \cdots, \theta_n$ are linearly independent over the rationals and ε is positive, then there is a constant k, depending only on n and ε, such that the inequalities composed by setting the absolute difference between θ_i and the ratio p_i to q less than an expression involving k, q, n, and ε (see 203) have only finitely many solutions in integers $h_1, h_2, \cdots, h_n, q$. The case $n = 1$ is, of course, Roth's theorem. **(C.A.Ro.)**

PROBABILISTIC NUMBER THEORY

The theory of probability, conceived originally in connection with gambling, provides an insight into some remarkable multiplicative properties of the integers—properties that in the absence of probability theory would remain undiscovered or at best inexplicable. Probability illuminates aspects of arithmetic because certain divisibility phenomena reflect the probabilistic idea of independence.

The number of prime divisors. According to the fundamental theorem of arithmetic, each positive integer can be factored in just one way into a product of primes. From the multiplicative point of view then, an integer is built up of primes, and it is natural to ask how many of them it contains. Probability is useful here.

The number $r(n)$ is taken to be the number of distinct primes appearing in the factorization of n. For example, $r(12) = 2$ because $12 = 2 \cdot 2 \cdot 3$ contains the two prime divisors 2 and 3 (the divisor 2 is counted only once, even though it appears twice in the factorization). It is simple to list the values of $r(n)$ for, say, the first 35 values of n (see 204). The case $n = 1$ is special because 1 is not counted as a prime, $r(1) = 0$, whereas $r(n)$ is a positive integer in all other cases.

The number $r(n)$ grows very slowly. The smallest integer containing two distinct prime factors is $6 = 2 \cdot 3$, and the smallest containing three of them is $30 = 2 \cdot 3 \cdot 5$; the smallest containing four will be $210 = 2 \cdot 3 \cdot 5 \cdot 7$, which is off the table. On the other hand, the fact that there is an infinite number of primes implies that $r(n)$ takes on arbitrarily large values: $r(n) = k$ if n is the product of k distinct primes, and, however large k may be, there do exist k distinct primes. The same fact implies that $r(n)$ takes on the value 1 infinitely often: $r(n) = 1$ if n is a prime. As n increases through the integers, $r(n)$ thus fluctuates in an irregular way, taking on arbitrarily large values but also dropping back infinitely many times to 1 (and infinitely many times to 2, and so on).

Distribution of prime factors

In view of this irregularity, it is of mathematical interest to see how $r(n)$ behaves on the average and how its values are distributed. It is supposed that n varies from 1 through N, in which N is a large integer, and the average (see 205) of $r(n)$ over this range is considered. If $N = 100$, this average is 1.71: the typical integer under 100 has rather less than two prime divisors. If $N = 100{,}000{,}000$, the average is not much greater, being approximately 2.9: the typical integer under 100,000,000 has just under three prime divisors—remarkably few. Even for $N = 10^{100}$, sometimes called a googol, the average is only about 5.4. Still, the average does go to infinity with N—but very slowly indeed: if N is the 10^{100}-th power of 10, a considerable number called a googolplex, the average is around 23.9. Except for the case $N = 100$, these figures come from an approximation specifying that the arithmetic average of N values of r is asymptotically equal to the logarithm of the logarithm of N (see 206).

A more detailed description of the way in which the number of prime factors varies is provided by the complete distribution of the values of $r(n)$ as n runs from 1 to N. A sorting of all the cases for $N = 100$ gives the distribution (see 207). That is, of the integers up to 100, 35 contain exactly one prime factor, 56 contain two, and 8 contain three, the corresponding proportions thus being .35, .56, and .08 in a table relating integer values from zero to three (inclusive) with frequency and with proportional frequency (see 207).

The high frequencies are near the average, and they taper off on either side of this central value. The problem probability helps solve is that of describing this distribution in more detail—for large values of N, for which an actual counting out of cases would be impossible. The probabilistic analysis that follows reveals the striking fact that this distribution, when converted to the proper scale, approximately follows the normal law of errors; that is, the proportion of integers n—in the range $1 \leq n \leq N$—for which $r(n)$, properly scaled, lies between specified limits will be approximately the area between those limits under the standard normal density—the famous bell-shaped curve that describes the distribution of various quantities (see PROBABILITY, THEORY OF).

(203) $$\left|\theta_i - \frac{h_i}{q}\right| < \frac{k}{q^{1+(1/n)+\varepsilon}}, \qquad i = 1, 2, \cdots, n$$

(204)

n	$r(n)$	n	$r(n)$	n	$r(n)$	n	$r(n)$	n	$r(n)$
1	0	8	1	15	2	22	2	29	1
2	1	9	1	16	1	23	1	30	3
3	1	10	2	17	1	24	2	31	1
4	1	11	1	18	2	25	1	32	1
5	1	12	2	19	1	26	2	33	2
6	2	13	1	20	2	27	1	34	2
7	1	14	2	21	2	28	2	35	2

(205) $$\frac{r(1) + r(2) + \cdots + r(N)}{N}$$

(206) $$\frac{r(1) + r(2) + \cdots + r(N)}{N} \approx \log \log N$$

(207)

Value	Frequency	Proportional Frequency
0	1	.01
1	35	.35
2	56	.56
3	8	.08

(208) $$\frac{r(n) - \log \log N}{\sqrt{\log \log N}}$$

The appropriate scale is a ratio composed of $r(n)$ and the negative value of and the square root of the logarithm of the logarithm of N (see 208); the centre log log N of the scale comes from the approximation (see 206) for the average value of $r(n)$, and the division by $\sqrt{\log \log N}$ properly counteracts the ever-increasing dispersion about this central value. The proportion of integers up to N for which this ratio (see 208) lies between x and y is for large N approximately the area of the shaded region in Figure 1. For example, if $x = -.2$ and $y = +1.2$, this area is about .46. If N is a googol, so that log log $N \approx 5.4$, then the stated ratio lies between these two limits if $r(n)$ itself lies between 5 and 8. If N is a googolplex, so that log log $N \approx 23.9$, the ratio is in this range if $r(n)$ is between 23 and 29. Thus something like half the integers under a googol have 5 to 8 prime divisors, and something like half the integers under a googolplex have 23 to 29 prime divisors.

Independence. The right-hand column of the table constructed with the values 0, 1, 2, 3 and corresponding frequencies 1, 35, 56, 8 (see 207) can be viewed as a set of probabilities. If an integer is drawn at random from among the integers $1, 2, 3, \cdots, 99, 100$, each having probability $\frac{1}{100}$ of being drawn, then there is probability .56 that it has exactly two prime divisors. Probability theory applies, leading to the distribution law just described, not merely because proportional frequencies can be viewed as probabilities, but because it is possible to bring to bear one of the basic ideas of probability, that of independence.

If Peter throws a pair of balanced dice, there is probability $\frac{6}{36} = \frac{1}{6}$ that the total number of dots showing is seven. If Paul throws a pair, there is probability $\frac{4}{36} = \frac{1}{9}$ that his total is five. The probability that Peter throws a seven *and* that Paul throws a five is the product $\frac{1}{6} \cdot \frac{1}{9} = \frac{1}{54}$, and this is because the two events in question are independent. Mathematically, two events are independent if their probabilities satisfy this product rule, and the definition reflects a phenomenon common in nature: the occurrence of the one event has no influence on the occurrence of the other.

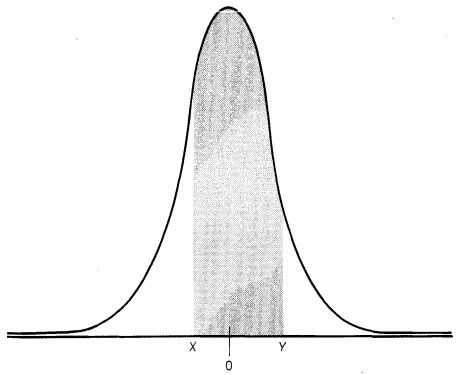


Figure 1: The normal distribution curve, showing two real values x and y and the corresponding shaded area beneath the curve that gives a proportion or a probability (see text).

To see how independence ties in with the behaviour of $r(n)$, the set of multiples of k is considered for each integer k. If this set is denoted A_k, A_2 consists of even integers, A_3 consists of those divisible by 3, and so on. Of these sets (see 209), in which the subscript ranges over the primes, some contain a given integer n and some do not, and $r(n)$ is precisely the number of sets in the list (209) that do contain n. To ask after the statistical behaviour of $r(n)$ is to ask for the probability that $r(n)$ takes on a specified value, say 4, and this is to ask for the probability that n lies in exactly 4 of the sets in the list (209). This probability can in principle be computed and can in fact be understood via the probabilities that n lies in A_2, in A_3, in A_2, and A_3 simultaneously, in A_2 and A_5 and A_{11} simultaneously, and so on for all combinations.

If n is drawn at random from among 1, 2, 3, $\cdots$, 99, 100, the probability that it lies in A_2 is $P(A_2) = \frac{50}{100} = \frac{1}{2}$, because there are 50 even integers up to 100. The probability that n lies in A_3 is $P(A_3) = \frac{33}{100}$, because there are 33 multiples of 3 up to 100. That $P(A_3)$ is approximately $\frac{1}{3}$ reflects the fact that every third integer is a multiple of 3; that $P(A_3)$ is not exactly $\frac{1}{3}$ is due to the fact that 3 does not divide evenly into 100.

Now n lies in the set $A_2 \cap A_3$ (the intersection of A_2 and A_3); that is to say, n lies both in A_2 and in A_3, when it is divisible both by 2 and by 3. And n is divisible both by 2 and by 3 exactly when it is divisible by their product 6; this is so because 2 and 3 have no common factor, and it represents a fundamental arithmetic fact (if n is divisible both by 2 and by 4, say, it does not follow that it is divisible by their product 8, $n = 12$ providing a contrary instance). In brief, the intersection of A_2 and A_3 is A_6 (see 210). The probability that n is divisible by 2 and by 3 is therefore $P(A_2 \cap A_3) = P(A_6) = \frac{16}{100}$, because there are 16 multiples of 6 in the range 1 to 100 (dividing 6 into 100 gives a quotient of 16—and an irrelevant remainder of 4); $P(A_6)$ is approximately $\frac{1}{6} = 0.1666\ldots$ because every sixth integer lies in A_6.

What is essential is the approximation stating that the probability of the intersection of A_2 and A_3 is approximated by the product of the corresponding probabilities (see 211), and the validity of this approximation rests on the fact that 2 and 3 individually divide n if, and only if, their product 6 does, together with the fact that A_2, A_3, and A_6 have respective probabilities approximately $\frac{1}{2}$, $\frac{1}{3}$, and $\frac{1}{6}$. Now this approximation (see 211) can advantageously be viewed as asserting the approximate independence of two events: An integer is drawn at random between 1 and 100; the event that it is divisible by 2 is approximately independent of the event that it is divisible by 3.

The rest of the events in the list (209)—divisibility by 5, divisibility by 7, and so on—are also approximately independent of one another if n is chosen at random between 1 and a large integer N. This is the key idea of probabilistic number theory. Now $r(n)$, the quantity of interest, is the number among these events that actually occur, and it is just this kind of random quantity (mainly for cases in which the events in question are strictly independent) that has received much attention in the classical literature of probability. Under appropriate conditions, such a number of occurrences will, when measured on the proper scale, approximately follow the normal distribution curve, and this is true of $r(n)$ rescaled as above (see 208). This theorem (it becomes a precise mathematical theorem, a form of the central limit theorem, when stated in terms of limits [see 216]) is plausible on probability grounds and can be given a rigorous proof by means of a proper combination of probability theory and number theory—there are difficulties, owing to the fact that the events in the list (209) are only approximately independent.

Euler's function

The Euler φ-function. A different application of probabilistic reasoning explains some properties of Euler's function, the number $\varphi(n)$ of integers less than or equal to n and relatively prime to n (that is, having no factors in common with n). The integers not exceeding 6 and relatively prime to it are 1 and 5, so that $\varphi(6) = 2$. The concern here is with $\varphi(n)/n$, the proportion of integers up to n that share no factors with it. It has a convenient expression as a product of factors of the form 1 minus the reciprocal of p (see 212).

The right side of this equation stands for the product of the factors $(1 - 1/p)$ for all primes p that divide n; if $n = 6$, for example, it is $(1 - \frac{1}{2})(1 - \frac{1}{3})$, which checks with $\varphi(6)/6 = 2/6$.

As before, n is chosen at random between 1 and N. If 3 divides n, it contributes to the product (see 212) the factor $(1 - \frac{1}{3})$; as indicated before, the probability of this is about $\frac{1}{3}$ for large N. If 3 does not divide n, then $(1 - \frac{1}{3})$ does not appear in the product (see 212), and in this case 3 may conveniently be regarded as contributing to the product the factor 1; the probability of this is about $\frac{2}{3}$. So the expected value of the factor contributed by 3, the amount 3 contributes on the average, is about 1 minus the reciprocal of 3 squared (see 213). The product of this and the analogous expected values for the other primes is an infinite product of factors of the type 1 minus the reciprocal of a prime squared (see 214), an infinite product known to have value $6/\pi^2$.

Now the expected value of a product of independent random factors is the product of their individual expected values; the random factors here are approximately independent (see 211). This is indeed true in the limit: the arithmetic average of N terms of the type $\varphi(k)$ divided by k converges to 6 over the square of π (see 215).

If m is chosen at random between 1 and n, the chance that it is relatively prime to n is $\varphi(n)/n$. And if n is chosen at random between 1 and N, and then m is chosen at random between 1 and n, the chance that m and n are relatively prime is the average referred to above (see 215). Because m and n play symmetric roles in this argument, it follows that, if m and n are chosen independently and randomly between 1 and N, the probability that they are relatively prime is near $6/\pi^2$ for large N.

In addition to its expected value, it is possible to describe the distribution of $\varphi(n)/n$. Figure 2 gives this distribution for $N = 100$. The height of the curve over a given point x on the horizontal scale gives the proportion of n, $1 \leq n \leq 100$, for which $\varphi(n)/n \leq x$. (The distribution actually gives discrete points, linked in the diagram by line segments.) The curve is seen to be irregular, rising slowly over most of its range but very sharp-

ly in some places. A probabilistic analysis shows that the corresponding curve for general N tends, as N increases to infinity, toward a limiting curve that is very irregular indeed. The limiting curve has horizontal tangent at almost all (in the technical sense of measure theory) of its points, so that its rate of increase is 0 almost everywhere, but it nonetheless contrives to climb continuously from 0 to 1 over its course (it represents what is called a singular function).

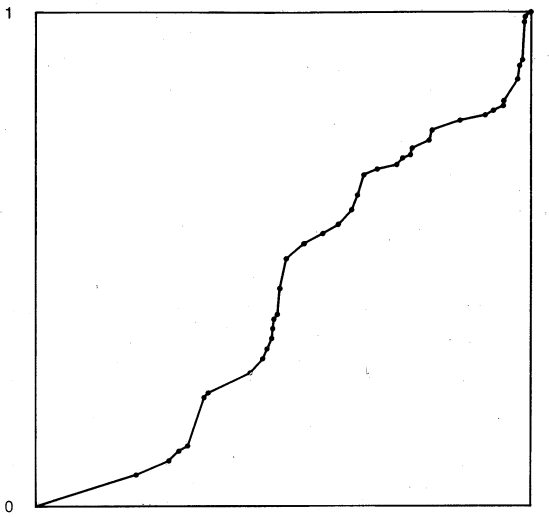


Figure 2: Probability distribution, such that the height of the curve over a given point x on the horizontal axis gives the proportion of n for which $\phi(n)/n \leq x$.

The theorems. The exact theorem governing the behaviour is of the above ratio (208) that identifies the limit of $P_N(A)$ as N approaches infinity as the integral of the normal density and in which $P_N(A)$ for a set A of integers is $1/N$ times the number of integers n that satisfy $1 \leq n \leq N$ and at the same time lie in A [P_{100} was denoted P above; in (216) $\{n: \cdots\}$ denotes the set of integers satisfying the condition following the colon].

It should be noted that the number of multiples of k up to N is $[N/k]$, the integer part of N/k, so that $P_N(A_k)$ is approximately the reciprocal of k (see 217), the approximation valid for large N because the two quantities in question differ by less than $1/N$. It is a consequence of the fundamental theorem of arithmetic that distinct primes p and q each divide n if and only if their product does. Therefore, $A_p \cap A_q = A_{pq}$ and P_N of the intersection of A_p and A_q is approximately a product of two factors, one involving A_p and one involving A_q (see 218). If $\delta_p(n)$ is 1 if the prime p divides n, and 0 otherwise, then r as a function of n is the sum over primes of $\delta_p(n)$ (see 219). The approximation (218) shows that the summands here are approximately independent random variables if n is random, $n \leq N$. The expected value of $r(n)$ is a sum of values of P_N that has two approximations (see 220), in which the first approximation comes from above reasoning and the second is a basic fact of number theory (the combined error is bounded, so the percentage error goes to 0). The variance of $r(n)$ is essentially what it would be if the summands were truly independent, and this turns out to be asymptotically log log N as well. The normalizing constants (see 216) are thus the asymptotic mean and standard deviation of the sum (see 219) and the summands behave sufficiently like independent random variables that the central limit theorem (see 216) does hold. This was first proved by sieve methods; simpler proofs based on the method of moments have since been found.

Error estimates for the central limit theorem (216) have been found; these coincide with the estimates for independent random variables.

It is a consequence of the central limit theorem that, if x_N goes to infinity faster than $\sqrt{\log \log n}$, then the limit, as N approaches infinity, of P_N is 1 when applied to the set of integers n for which the absolute difference between $r(n)$ and the log log N is less than x_N (see 221). This stands to the previous result as the law of large numbers stands to the central limit theorem. From this limit (221) it follows that, for most values of n under N, $r(n)$ is near log log N; the central limit theorem (216) gives very detailed information about how $r(n)$ is distributed around this central value.

(209) $A_2, A_3, A_5, A_7, A_{11}, \cdots,$

(210) $A_2 \cap A_3 = A_6$

(211) $P(A_2 \cap A_3) = .16 \approx P(A_2) \cdot P(A_3) = .165$

(212) $$\frac{\varphi(n)}{n} = \prod_{p|n}\left(1 - \frac{1}{p}\right)$$

(213) $$\frac{1}{3} \cdot \left(1 - \frac{1}{3}\right) + \frac{2}{3} \cdot 1 = 1 - \frac{1}{3^2}$$

(214) $$\left(1 - \frac{1}{2^2}\right)\left(1 - \frac{1}{3^2}\right)\left(1 - \frac{1}{5^2}\right) \cdots$$

(215) $$\frac{1}{N}\left(\frac{\varphi(1)}{1} + \frac{\varphi(2)}{2} + \cdots + \frac{\varphi(N)}{N}\right) \to \frac{6}{\pi^2}$$

(216) $$\lim_{N \to \infty} P_N\left\{n{:}x \leq \frac{r(n) - \log \log N}{\sqrt{\log \log N}} \leq y\right\} = \frac{1}{\sqrt{2\pi}} \int_x^y e^{-\frac{1}{2}u^2}\, du$$

(217) $$P_N(A_k) = \frac{1}{N}\left[\frac{N}{k}\right] \approx \frac{1}{k}$$

(218) $$P_N(A_p \cap A_q) = \frac{1}{N}\left[\frac{N}{pq}\right] \approx \frac{1}{pq} \approx \frac{1}{N}\left[\frac{N}{p}\right] \cdot \frac{1}{N}\left[\frac{N}{q}\right] = P_N(A_p) \cdot P_N(A_q)$$

(219) $$r(n) = \sum_p \delta_p(n)$$

(220) $$\sum_p P_N(A_p) = \sum_{p \leq N} \frac{1}{N}\left[\frac{N}{p}\right] \approx \sum_{p \leq N} \frac{1}{p} \approx \log \log N$$

(221) $$\lim_{N \to \infty} P_N\{n{:}|r(n) - \log \log N| < x_N\} = 1$$

(222) $$f(n) = \sum_p f(p)\, \delta_p(n)$$

Additive functions

A function $f(n)$ of integers is called additive if $f(mn) = f(m) + f(n)$ whenever m and n are relatively prime, and it is called completely additive if, in addition, $f(p^\alpha) = f(p)$ for prime powers p^α. If f is completely additive, then it can be written as a sum over primes of products $f(p)\delta_p(n)$ (see 222), a special case of which was previously given (see 219)—$r(n)$ is completely additive with $r(p) = 1$. The summands (see 222) being approximately independent, as in the case of $r(n)$, it is possible to establish for completely additive functions a general central limit theorem, similar to the one that was already stated (see 216), under conditions that parallel those of the classical Lindeberg theorem. It is even possible to treat the case in which the limiting normal distribution is replaced by a more general infinitely divisible distribution.

Similar results hold if $f(n)$ is additive but not completely additive. An example of such a function is the number of prime divisors of n with multiplicity counted, so that

$f(12) = 3$ because 2 appears twice in the factorization $12 = 2 \cdot 2 \cdot 3$ and 3 appears once. The central limit theorem (216) is unaltered if $r(n)$ is replaced by this function.

The central limit theorem and the law of large numbers are represented in probabilistic number theory by the above limits (216) and (221) and their generalizations. There are also representatives of certain probability theorems that concern the convergence of random series, and the limit theorem for $\varphi(n)/n$ is one of these.

The logarithm of the ratio of $\varphi(n)$ to n is a sum over primes of products involving $\delta_p(n)$ and a logarithmic function (see 223), and this is a completely additive function. The $\delta_p(n)$ behave very much as do independent random variables that assume the values 1 and 0 with probabilities p^{-1} and $1 - p^{-1}$. If the $\delta_p(n)$ were exactly like such variables, it would follow from probability theory that, since $\Sigma_p p^{-1} \log (1 - p^{-1})$ converges, the series (see 223) converges with probability 1; moreover, the distribution function $F(x)$ of the total sum could be identified as the one having as characteristic function (Fourier transform; see 224) the product of the characterstic functions of the various summands. As N increases, the $\delta_p(n)$ mimic ever more closely the independent variables, and in the limit (223) does have the distribution function $F(x)$ just derived (see 225). Results on random series imply $F(x)$ is a singular function; the curves corresponding to the one in Figure 2 converge to $F(\log x)$, which must also be singular.

$$\log\frac{\varphi(n)}{n} = \sum_p \delta_p(n) \log\left(1 - \frac{1}{p}\right) \tag{223}$$

$$\int_{-\infty}^{\infty} e^{itx}\, dF(x) = \prod_p \left\{1 - \frac{1}{p} + \frac{1}{p}\exp\left[it \log\left(1 - \frac{1}{p}\right)\right]\right\} \tag{224}$$

$$\lim_{N\to\infty} P_N\left\{n: \log\frac{\varphi(n)}{n} \leq x\right\} = F(x) \tag{225}$$

$$\lim_{N\to\infty} P_N\left\{n: \log\frac{\sigma(n)}{n} \leq x\right\} = G(x) \tag{226}$$

$$D\left\{n: \log\frac{\sigma(n)}{n} \leq x\right\} = G(x) \tag{227}$$

A generalization of this result covers a broad class of completely additive functions for which the series $\Sigma_p |f(p)|/p$ and $\Sigma_p |f(p)|^2/p$ converge (if these diverge, as for $r(n)$, central limit theory applies instead). It is possible also to treat functions that are only additive (not completely additive), such as $\log \{\sigma(n)/n\}$, in which $\sigma(n)$ is the sum of all the divisors (prime or not) of n. In this case, as N increases, the distribution function converges to a function $G(x)$ whose characteristic function can be identified (see 226).

All these results can be restated in terms of density. A set A of integers has density d if $\lim_N P_N(A) = d$, which is expressed by $D(A) = d$. The limit relation that was just derived can be restated in terms of D (see 227). The distribution function $G(x)$ is continuous for all x, as follows from an analysis of its characteristic function, which is interesting because it shows that the set of deficient numbers (those satisfying $\sigma(n) < n$) and the set of abundant numbers (those satisfying $\sigma(n) > n$) have densities and that the set of perfect numbers (those satisfying $\sigma(n) = 2n$) has density 0 (it is conjectured that the last set is finite).

Except for one (see 215), all the results here were discovered in the present century. Hardy and Ramanujan proved the law of large numbers (see 221) in 1920 by nonprobabilistic methods, and Paul Turán in 1934 gave it a probablistic proof. The mathematicians Paul Erdös (Hungarian) and Mark Kac (U.S.) proved the central limit theorem in 1939, extensions and refinements of which as well as other results (see 226) are associated with the names Davenport, Delange, H. Halberstam, J. Kubilius, I.J. Schoenberg, and A. Wintner. (P.P.B.)

BIBLIOGRAPHY. The following works are concerned with elementary topics: HAROLD T. DAVIS (comp.), *Tables of the Higher Mathematical Functions*, vol. 2 (1935, reprinted 1963), contains extensive tabular and bibliographical material on Bernoulli numbers. Additional references on Bernoulli numbers include: H.S. VANDIVER, "An Arithmetical Theory of the Bernoulli Numbers," *Am. Math. Soc. Transl.*, 51:502–531 (1942); JAMES V. USPENSKY and MAXWELL A. HEASLET, *Elementary Number Theory*, ch. 9 (1939); ALAN FLETCHER, J.C.P. MILLER, and LOUIS ROSENHEAD, *An Index of Mathematical Tables*, pp. 40–80 (1946); and H.S. VANDIVER, "On Developments in an Arithmetic Theory of the Bernoulli and Allied Numbers," *Scr. Math.*, 25:273–303 (1961). For information on Fermat's last theorem, see LOUIS J. MORDELL, *Three Lectures on Fermat's Last Theorem* (1921); H.S. VANDIVER and G.E. WAHLIN, "Report of the Committee on Algebraic Numbers," *Bull. Natn. Res. Coun., Wash.*, no. 62, ch. 2 (1928); and H.S. VANDIVER, "Fermat's Last Theorem: Its History and the Nature of the Known Results Concerning It," *Am. Math. Mon.*, 53:555–578 (1946). Several treatises on algebra have chapters on the arithmetic theory of continued fractions. HUBERT S. WALL, *Analytic Theory of Continued Fractions* (1948, reprinted 1967), deals with questions of convergence and the function-theoretic aspects of continued fractions. The paper of J.S. MACNERNEY, "Investigation Concerning Positive Definite Continued Fractions," *Duke Math. J.*, 26:663–677 (1959), deals with continued fractions whose partial numerators and denominators are matrixes.

The reader interested in algebraic number theory may wish to consult some of these works: DAVID HILBERT, *Die Theorie der algebräischen Zahlkörper* (*Jahresbericht der deutschen Mathematiker-Vereinigung*, vol. 4, 1897), reprinted in vol. 1 of Hilbert's *Gesammelte Abhandlungen*; this work (the *Zahlbericht*) is not available in English. It contains an extensive bibliography covering the literature up to 1897. The "Report of the Committee on Algebraic Numbers," *Bull. Natn. Res. Coun., Wash.*, no. 28 (1923), and no. 62 (1928), is an updating of the *Zahlbericht*. HARRY POLLARD, *The Theory of Algebraic Numbers* (1950), is a brief, elementary account in classical style. The following take the modern p–adic view: EMIL ARTIN, *Algebraic Numbers and Algebraic Functions* (1967); EDWIN WEISS, *Algebraic Number Theory* (1963); and Z.I. BOREVICH and I.R. SHAFAREVICH, *Number Theory* (1966; orig. pub. in Russian, 1964), also contains a good treatment of analytic number theory. The following also incorporate a treatment of class field theory: EMIL ARTIN and JOHN TATE, *Class Field Theory* (1967); J.W.S. CASSELS and ALBRECHT FROHLICH (eds.), *Algebraic Number Theory* (1967); SERGE LANG, *Algebraic Number Theory* (1970). ANDRE WEIL, *Basic Number Theory* (1967), also includes class field theory as well as some material on analytic number theory.

The following works are especially useful as references on analytic number theory. SARVADAMAN CHOWLA, *The Riemann Hypothesis and Hilbert's Tenth Problem* (1965); HAROLD DAVENPORT, *Multiplicative Number Theory* (1967); HELMUT HASSE, *Vorlesungen über Zahlentheorie*, 2nd rev. ed. (1963); KARL PRACHAR, *Primzahlverteilung* (1957); JEAN-PIERRE SERRE, *Cours d'arithmetique* (1970); CARL L. SIEGEL, *Analytische Zahlentheorie* (1963–64); EDWARD C. TITCHMARSH, *The Theory of the Riemann Zeta-Function* (1951); IVAN M. VINOGRADOV, *The Method of Trigonometrical Sums in the Theory of Numbers* (1954; orig. pub. in Russian, rev. ed., 1953), are recent works, suitable for an introduction to topics in analytic number theory, and are by authors who have contributed significantly to the subject. Other good introductory works include RAYMOND AYOUB, *An Introduction to the Analytic Theory of Numbers* (1963); JOSEPH LEHNER, *A Short Course in Automorphic Functions* (1966); and HANS RADEMACHER, *Lectures on Analytic Number Theory* (1954–55).

Some works concerning geometric number theory and related topics are J.W.S. CASSELS, *An Introduction to Diophantine Approximation* (1957) and *An Introduction to the Geometry of Numbers*, 2nd ed. (1971); L. FEJES TOTH, *Lagerungen in der Ebene auf der Kugel und im Raum* (1953) and *Regular Figures* (1964); CORNELIUS G. LEKKERKERKER, *Geometry of Numbers* (1969); HERMANN MINKOWSKI, *Geometrie der Zahlen* (1896, reprinted 1969), *Diophantische Approximationen* (1907, reprinted 1961), and *Gesammelte Abhandlungen*, 2 vol. (1911, reprinted 1967); and CLAUDE A. ROGERS, *Packing and Covering* (1964).

On probabilistic number theory, see GODFREY H. HARDY and EDWARD M. WRIGHT, *An Introduction to the Theory of Numbers*, 4th ed. (1960), an excellent introduction to the theory of numbers, touching incidentally on probabilistic number theory; MARK KAC, *Statistical Independence in Probability, Analysis and Number Theory* (1959), which contains a very readable introduction to probabilistic number theory and is highly recommended; and JONAS KUBILIUS, *Probabilistic Methods in the Theory of Numbers* (1964; Eng. trans. from the 2nd Russian ed., 1962), a full-scale treatment of the subject with a large bibliography.

Some additional references on various aspects of number theory are as follows: LEONARD E. DICKSON, *History of the Theory of Numbers*, 3 vol. (1919–23), a monumental topical history; and *Modern Elementary Theory of Numbers* (1939); H.J.S. SMITH, *Report on the Theory of Numbers* (1859–65, reprinted 1965); GEORGE B. MATHEWS, *Theory of Numbers* (1892); ROBERT D. CARMICHAEL, *The Theory of Numbers* (1914) and *Diophantine Analysis* (1915), reprinted together as *The Theory of Numbers and Diophantine Analysis* (1959); OYSTEIN ORE, *Number Theory and Its History* (1948); BURTON W. JONES, *The Arithmetic Theory of Quadratic Forms* (1950); TRYGVE NAGELL, *Introduction to Number Theory*, 2nd ed. (1964); WILLIAM J. LEVEQUE, *Elementary Theory of Numbers* (1962) and *Topics in Number Theory*, 2 vol. (1956); HAROLD DAVENPORT, *The Higher Arithmetic*, 3rd ed. (1968); HANS RADEMACHER and OTTO TOEPLITZ, *The Enjoyment of Mathematics: Selections from Mathematics for the Amateur* (1957); HANS RADEMACHER, *Lectures on Elementary Number Theory* (1964); ERNST TROST, *Primzahlen* (1953); and THEODOR ESTERMANN, *Introduction to Modern Prime Number Theory* (1952).

(I.K./R.G.A./C.A.Ro./P.P.B.)

Numerical Analysis

The term numerical analysis, which has been used since 1947 when the Institute of Numerical Analysis was founded at the University of California, describes the branch of mathematics concerned with methods of finding numerical solutions to problems.

To illustrate the use of numerical analysis, the concept of a function and the procedure known as interpolation will be described. The idea of a function in mathematics involves an association of each object in one set of objects with each object in another set. In particular, if the two sets considered are sets of real numbers, then each number can have an associated number given by some rule (see below). For example, the number 2 may have 4 associated with it, 3 may have 9, 4 may have 16, and, in general, x may have x^2. If $y = x^2$, y is said to be a function of x. It is usual to denote a function of x by $f(x)$. The correspondence does not have to be given by a formula (as $y = x^2$) but could be given by a tabulation.

Now a given $f(x)$ has values that depend on the values of the argument x; *i.e.*, independent variants upon the value of which that of a function depends. These values may be found for particular values of x by using the formula that defines the function. It is straightforward, for instance, to find the values of x^2 corresponding to values of x. In cases in which the formula is cumbersome, however, it is often easier to interpolate; *i.e.*, to calculate several values for the function and use numerical analysis to obtain intermediate values of the function. The procedure is to calculate widely spaced values with more decimal places than are ultimately required and then to interpolate values by an approximate process and round them off to the desired number of figures. An English mathematician named Henry Briggs used a similar procedure in 1624 to construct his celebrated table of logarithms (exponents [raised numbers] indicating the process to which a number is raised to produce a given number; *e.g.*, the logarithm of 100 to the base 10 is 2). The interpolation problem was first taken up by Briggs and was solved by the celebrated 17th-century English mathematician Sir Isaac Newton and James Gregory, a 17th-century Scottish mathematician and astronomer. Newton gave a great amount of attention to the interpolation formulas and investigated their application in detail; his directions for the construction of tables are very practical.

Early work on interpolation

Numerical analysis is also used for obtaining values of the differential coefficients of a function, solving differential equations, performing integrations, and solving systems of linear equations. In spite of the trend of modern mathematics toward greater abstraction and generality, many scientific and technological fields require quantitative results rather than qualitative ones. Thus, space flight would not be possible without the precise numerical solution of the underlying equations of motion. The modern development of numerical analysis has been strongly influenced by the development of the electronic computer.

A constructive method is formulated mathematically as an algorithm; that is, a prescribed sequence of algebraic, arithmetical, or logical operations. Special algorithmic languages, such as Fortran and Algol, have been developed for the description of algorithms. These languages can be read by the computing machine without further translation. Problems in algebra can usually be solved by terminating algorithms with a finite number of steps. An example is the Euclidean algorithm for finding the greatest common divisor of two integers. Problems in analysis usually require nonterminating algorithms that have an infinite number of steps. Only a finite number of steps can, of course, be carried out in any concrete application, but for a convergent algorithm the accuracy of the answer should increase indefinitely with the number of steps performed.

The prototype of a nonterminating algorithm is the principle of iteration, which in the simplest case may be described as follows. If f is a real function of a real variable, x_0 is chosen, and $x_1 = f(x_0)$, $x_2 = f(x_1)$, and generally $x_{n+1} = f(x_n)$ are formed where n runs through the positive integers. Under certain conditions the sequence of numbers $\{x_n\}$ tends to a solution of the equation $x = f(x)$. A trivial example is $f(x) = 1 + 0.1x$, $x_0 = 1$, $x_1 = 1.1$, $x_2 = 1.11$, $x_3 = 1.111$, and it is evident that $x_n \to 1.111111 \cdots = 10/9$, the unique solution of $x = 1 + 0.1x$. The same principle may be used to solve systems of equations and functional equations.

Use of numerical analysis to study numerical errors

In addition to the invention of efficient algorithms for the solution of specific problems, research in numerical analysis is concerned with the study of the numerical errors occurring in the application of these algorithms. Two kinds of errors are distinguished, commonly called truncation error and rounding error. Truncation errors are caused by terminating an essentially infinite algorithm after a finite number of steps. By a careful assessment of this error, it may be possible to estimate in advance the number of steps necessary to attain a predetermined accuracy, to speed up the convergence of the algorithm, and thus to attain the desired accuracy more quickly. The rounding error is caused by the fact that even in an electronic computer the results of most arithmetical operations have to be rounded off. Although the individual rounding errors are small, their cumulative effect can, in view of the large number of arithmetical operations performed, grow very rapidly and under unfavourable conditions invalidate the final result. In order to be sound, an algorithm must remain immune to the accumulation of rounding errors. This immunity is called numerical stability. (Ed.)

FINITE DIFFERENCES

Successive differencing of functions; interpolation. If a function $f(x)$ be given in a table for the values a, $a + 1$, $a + 2$, $\cdots$ of the argument, then the excess of any term $f(a + 1)$ above that which immediately precedes it, or the function $f(a + 1) - f(a)$, is called the first difference of the function $f(a)$ and is written $\Delta f(a)$ in which Δ is the Greek capital delta. Likewise, the first difference of $f(a + 1)$ is $f(a + 2) - f(a + 1)$, which is written $\Delta f(a + 1)$. The differencing procedure may be applied again to the new list of numbers $\Delta f(a)$, $\Delta f(a + 1)$, $\cdots$. The difference of the function $\Delta f(a)$ (see Box, equation 1), which is usually written $\Delta^2 f(a)$, is called the second

(1) $$\Delta(\Delta f(a)) = \Delta f(a+1) - \Delta f(a)$$

difference of $f(a)$. By repeating the process it is possible to form the third, fourth, $\cdots$, nth differences (see 2) by means of the relation that the nth difference is the difference of the $(n-1)$th (see 3). It is possible to list these differences in order to display the algebraic structure (see 4) in which n is a positive integer and the binomial coefficient notation is used. The procedure may be reversed to express a value of $f(a+n)$ in terms of the initial value and the differences. It follows that f at $(a+1)$ can be so expressed (see 5) and f at $(a+2)$ can be so expressed (see 6). In general f at $(a+n)$ is a succession of increasing differences of f at a, each difference being weighted with a binomial coefficient (see 7). Differences are arranged in tabular form, in what is called a difference table (see Table 1). The differences are usually written on a line between the two numbers for which

Table 1: Differences of $f(a)$

	entry	1st	2nd	3rd
$a-2$	$f(a-2)$			
		$\Delta f(a-2)$		
$a-1$	$f(a-1)$		$\Delta^2 f(a-2)$	
		$\Delta f(a-1)$		$\Delta^3 f(a-2)$
a	$f(a)$		$\Delta^2 f(a-1)$	
		$\Delta f(a)$		$\Delta^3 f(a-1)$
$a+1$	$f(a+1)$		$\Delta^2 f(a)$	
		$\Delta f(a+1)$		
$a+2$	$f(a+2)$			

they represent the differences. It will be observed that the even differences fall on the same line as the argument and function, while the odd differences lie between the lines. Thus, for the function $a^3 - a^2$ the table is Table 2. The third differences of this function are constant, and

Table 2: Differences of $f(a) = a^3 - a^2$

a	$f(a)$	Δ	Δ^2	Δ^3	Δ^4
-2	-12				
		10			
-1	-2		-8		
		2		6	
0	0		-2		0
		0		6	
1	0		4		
		4			
2	4				

all the following differences are zero. The difference of x^n is a polynomial of the $(n-1)$th degree (see 8). Therefore the first difference of a polynomial of the nth degree is a polynomial of degree $n-1$. Hence the nth difference of a polynomial of degree n is a constant, and all higher differences will be zero. Products of x and factors composed of x after subtraction by increasing integer values (see 9) are called factorials. It is then possible to take a difference of factorials (see 10). The rule for differencing factorials $x^{(n)}$ with respect to x is analogous to the rule for differentiating x^n with respect to x. In like manner, second differences of factorials are possible (see 11). Moreover, the nth difference of a factorial (see 12) is the product of the integers 1, 2, 3, $\cdots$ n. A polynomial can always be represented by a series of factorials. If $f(x)$ is a polynomial of degree n, then its representation in differences at the point $(a+x)$ (see 13) is the form analogous to Taylor's theorem for a polynomial.

Divided differences. The case in which the values of the argument, for which the function is known, are unequally spaced will now be considered. It is supposed that $f(x)$ is to be tabulated for $x = x_0, x_1, x_2, \cdots x_n$, in which the intervals $x_1 - x_0$, $x_2 - x_1$, $\cdots$ need not be equal. For abbreviation, this may be written in square bracket notation (see 14). The first-order divided differences are then defined by ratios of differences of square brackets to differences in the corresponding x values (see 15). The order inside the brackets is immaterial (see 16). The difference between two first-order differences divided by the interval of the independent variable that they span is called the second-order divided difference (see 17). The divided differences of higher orders are formed in the same way. In general (see 18), it is possible to form the divided difference of the nth order. As an example, the entries in Table 3 are the cubes of the integers. It is easily seen that $[x_0, x_1, \cdots, x_n]$ is a linear function of $f(x_0)$, $f(x_1)$, $\cdots$ $f(x_n)$. From the definition it follows that specific calculations can be made showing this to be true for two points and three points (see 19), and in general it is true for n points (see 20), as can be proved by induc-

(2) $\Delta^3 f(a), \Delta^4 f(a), \cdots, \Delta^n f(a)$

(3) $\Delta^n f(a) = \Delta(\Delta^{n-1} f(a))$

(4)
$$\begin{cases} \Delta f(a) = f(a+1) - f(a) \\ \Delta^2 f(a) = f(a+2) - 2f(a+1) + f(a) \\ \Delta^n f(a) = \sum_{s=0}^{n} (-1)^{n-s} \binom{n}{s} f(a+s) \end{cases}$$

(5) $f(a+1) = f(a) + \Delta f(a)$

(6) $f(a+2) = f(a) + 2\Delta f(a) + \Delta^2 f(a)$

(7) $f(a+n) = \sum_{s=0}^{n} \binom{n}{s} \Delta^s f(a)$

(8) $\Delta x^n = (x+1)^n - x^n = \sum_{s=0}^{n-1} \binom{n}{s} x^s$

(9) $x^{(n)} = x(x-1)(x-2)\cdots(x-n+1)$

(10) $\Delta x^{(n)} = (x+1)^{(n)} - x^{(n)} = n x^{(n-1)}$

(11) $\Delta^2 x^{(n)} = n(n-1) x^{(n-2)}$

(12) $\Delta^n x^{(n)} = n!$

(13) $f(a+x) = \sum_{s=0}^{n} \frac{x^{(s)}}{s!} \Delta^s f(a)$

(14) $[x_s] = f(x_s), \qquad s = 0, 1, 2, \cdots, n$

(15)
$$\begin{cases} [x_0, x_1] = ([x_0] - [x_1])/(x_0 - x_1) \\ [x_1, x_2] = ([x_1] - [x_2])/(x_1 - x_2) \end{cases}$$

(16) $[x_0, x_1] = [x_1, x_0], \qquad [x_1, x_2] = [x_2, x_1]$

(17) $[x_0, x_1, x_2] = ([x_0, x_1] - [x_1, x_2])/(x_0 - x_2)$

(18)
$$[x_0, x_1, \cdots, x_n] = \frac{[x_0, x_1, \cdots, x_{n-1}] - [x_1, x_2, \cdots, x_n]}{x_0 - x_n}$$

(19)
$$\begin{cases} [x_0, x_1] = \dfrac{f(x_0)}{x_0 - x_1} + \dfrac{f(x_1)}{x_1 - x_0} \\ [x_0, x_1, x_2] = \dfrac{f(x_0)}{(x_0 - x_1)(x_0 - x_2)} + \dfrac{f(x_1)}{(x_1 - x_0)(x_1 - x_2)} + \\ \qquad + \dfrac{f(x_2)}{(x_2 - x_0)(x_2 - x_1)} \end{cases}$$

(20)
$$\begin{cases} [x_0, x_1, \cdots, x_n] = \dfrac{f(x_0)}{(x_0 - x_1)(x_0 - x_2)\cdots(x_0 - x_n)} + \\ \qquad + \dfrac{f(x_1)}{(x_1 - x_0)(x_1 - x_2)\cdots(x_1 - x_n)} + \\ \qquad + \dfrac{f(x_n)}{(x_n - x_0)(x_n - x_1)\cdots(x_n - x_{n-1})} \end{cases}$$

Table 3: Divided Differences of $f(x) = x^3$

x	$f(x)$	first order	second order	third order
−1	−1			
		1		
0	0		1	
		4		1
2	8		7	
		39		1
5	125		17	
		175		
10	1,000			

tion. It follows that the divided difference $[x_0, x_1, \cdots, x_n]$ is a symmetric function of its arguments $x_0, x_1, \cdots, x_n$ so that the order in which these are taken is indifferent. It appears from the definition of the divided differences that if $x_s = s$ $(s = 0, 1, 2, \cdots, n)$, then the nth divided difference is expressed in terms of nth differences (see 21). If $f(x) = 1/x$, the nth divided difference is a simple expression (see 22). As another example, if $f(x) = x^n$, then $[x_1, x_2, \cdots, x_n] = x_1 + x_2 + \cdots x_n$ and $[x_0, x_1, \cdots, x_n] = 1$. The divided differences of order n of a polynomial of the nth degree are constant and all the following differences are zero. If in the definitions of the divided differences x_0 is replaced by x, this system of equations may be written for f at x in terms of divided differences (see 23). From this there is obtained in succession a general expression for f at x including a remainder term (see 24) in which the remainder is expressed in a formula (see 25). This formula, discovered by Newton, is called Newton's interpolation formula with divided differences. It is one of the fundamental propositions of the calculus of finite differences. Other interpolation formulas used by the Scottish mathematician James Gregory, the 18th-century British mathematician James Stirling, the 19th-century German mathematician Carl Friedrich Gauss, and the 19th-century German astronomer Friedrich Wilhelm Bessel all follow as special cases from Newton's formula, and Newton himself investigated these special cases. If $f(x)$ is a polynomial of degree $p < n$, the remainder term R_n vanishes. In more general cases, an approximate value of R_n may be obtained. If $f(x)$ has a continuous derivative of order n in the interval considered, it follows from Rolle's theorem that the remainder equals the nth derivative of the function evaluated at the point ξ (see 26), ξ being situated in the smallest interval containing the numbers $x_1, x_2, \cdots x_n, x$. If all the differences diminish rapidly, so that from and after the mth order they are not worth taking into account in practice, it denotes that a polynomial of degree m will be a sufficient representation of the function in the considered interval. As an example of the use of Newton's formula from Table 3, the 3rd power of x can be expressed (see 27). Here the divided differences have been taken on the descending diagonal line.

Newton's interpolation formula

If in (20) x_0 is replaced by x, after a slight modification Lagrange's interpolation formula results (see 28), in which $\phi(x) = (x - x_1)(x - x_2) \cdots (x - x_n)$, and the remainder term R_n is the same as in Newton's formula (24). It is seen that for $x = x_s$ $(s = 1, 2, \cdots n)$ the right-hand side of (28) takes the value $f(x_s)$. If $f(x)$ is a polynomial of degree $p < n$, the remainder term vanishes. Lagrange's interpolation formula shows how a polynomial can be passed through any number of the tabulated values and the polynomial used for approximating the function between these entries. The Lagrange formula, however, has the defect that if another point of interpolation were added, the new higher degree interpolation polynomial could not be obtained by easily modifying the previous ones. Instead, it is necessary to start the computation of the new polynomial all over again. Newton's formula (24), on the other hand, has the advantage that the next higher degree interpolation polynomial is found simply by adding a new term.

Thus, for example, if it is desired to construct a polynomial of the second degree in x that shall have the values 3, 1, 3, the values of x being −1, 0, 2, respectively, then from (28) a simple calculation shows that the polynomial has coefficients 1, −1, and 1 (see 29).

Reciprocal differences. A method of interpolation will now be described by means of rational functions. It is more general than polynomial interpolation. Reciprocal differences, introduced by Thiele, led to the approximate representation of a function by a continued fraction. The reciprocal difference of $f(x)$ of the first order is defined (see 30) to be the reciprocal of the divided difference. The reciprocal difference of second order is defined in terms of the first-order reciprocal differences (see 31). It is seen that this difference is symmetric in the arguments x_0, x_1 and x_2, that is, its value does not change with interchange of the arguments. Proceeding to reciprocal difference of the third order, (see 32) and generally for the nth order (see 33), it can be proved that reciprocal differences of any order are symmetric functions of their

$$(21)\quad [x_0, x_1, \cdots, x_n] = \frac{\Delta^n f(0)}{n!}$$

$$(22)\quad [x_0, x_1, \cdots, x_n] = \frac{(-1)^n}{x_0 x_1 \cdots x_n}$$

$$(23)\quad \begin{cases} f(x) = f(x_1) + (x - x_1)[x, x_1] \\ [x, x_1] = [x_1, x_2] + (x - x_2)[x, x_1, x_2] \\ \qquad \cdot \\ \qquad \cdot \\ \qquad \cdot \\ [x, x_1, \cdots, x_{n-1}] = [x_1, x_2, \cdots, x_n] + \\ \qquad\qquad + (x - x_n)[x, x_1, x_2, \cdots, x_n] \end{cases}$$

$$(24)\quad \begin{cases} f(x) = f(x_1) + (x - x_1)[x_1, x_2] + \\ \qquad + (x - x_1)(x - x_2)[x_1, x_2, x_3] + \cdots + \\ \qquad + (x - x_1)(x - x_2) \cdots \\ \qquad\quad (x - x_{n-1})[x_1, x_2, \cdots, x_n] + R_n \end{cases}$$

$$(25)\quad R_n = (x - x_1)(x - x_2) \cdots (x - x_n)[x, x_1, x_2, \cdots, x_n]$$

$$(26)\quad R_n = (x - x_1)(x - x_2) \cdots (x - x_n)\frac{f^{(n)}(\xi)}{n!}$$

$$(27)\quad x^3 = -1 + (x + 1) + (x + 1)x + (x + 1)x(x - 2)$$

$$(28)\quad f(x) = \sum_{s=1}^{n} \frac{f(x_s)}{x - x_s}\frac{\phi(x)}{\phi'(x_s)} + R_n$$

$$(29)\quad f(x) = \frac{3x(x-2)}{(-1)(-3)} + \frac{(x+1)(x-2)}{1(-2)} + \frac{3(x+1)x}{3 \cdot 2} = x^2 - x + 1$$

$$(30)\quad \begin{cases} \rho_1(x_0 x_1) = \dfrac{x_0 - x_1}{f(x_0) - f(x_1)} \\ \rho_1(x_1 x_2) = \dfrac{x_1 - x_2}{f(x_1) - f(x_2)} \end{cases}$$

$$(31)\quad \rho_2(x_0 x_1 x_2) = \frac{x_0 - x_2}{\rho_1(x_0 x_1) - \rho_1(x_1 x_2)} + f(x_1)$$

$$(32)\quad \rho_3(x_0 x_1 x_2 x_3) = \frac{x_0 - x_3}{\rho_2(x_0 x_1 x_2) - \rho_2(x_1 x_2 x_3)} + \rho_1(x_1 x_2)$$

$$(33)\quad \begin{cases} \rho_n(x_0 x_1 \cdots x_n) \\ \quad = \dfrac{x_0 - x_n}{\rho_{n-1}(x_0 x_1 \cdots x_{n-1}) - \rho_{n-1}(x_1 x_2 \cdots x_n)} + \\ \quad + \rho_{n-2}(x_1 x_2 \cdots x_{n-1}) \end{cases}$$

arguments. In equations (30–33) x is written for x_0 and $\rho_1(xx_1)$, $\rho_2(xx_1x_2)$, $\cdots$ is eliminated. This leads to a sequence of identities for f at x (see 34). This continued fraction is called Thiele's interpolation formula. If in the continued fraction $x = x_1, x_2, \cdots, x_n$ is stated successively, then $f(x_1), f(x_2), \cdots, f(x_n)$ is obtained. It follows that Thiele's formula gives a method of obtaining a rational function that agrees in value with a given function at any number of prescribed points. Augustin-Louis Cauchy, a French mathematician, has given another approximate representation by a rational function in which the numerator and denominator are formed in a way similar to Lagrange's formula (28).

Interpolation formulas with ordinary differences. The cases in which the points of interpolation are equally spaced will now be considered. If in Newton's formula (24) x_s is set equal to $a + s - 1$ and x is replaced with $a + x$, then a formula (see 35) for a function evaluated at $(a + x)$ in terms of ξ is obtained, in which ξ is some number in the range spanned by 0, $n - 1$, and x. This is the Gregory-Newton formula for forward interpolation. It was discovered by Gregory in 1670. The remainder term was given by Cauchy. It uses the differences on a downward diagonal line, and it is best suited for interpolation near the beginning of a table of differences.

The Gregory-Newton formula

If in (24) x_s is set equal to $a - s + 1$ and x is replaced with $a - x$, then Newton's backward interpolation formula is obtained for a function evaluated at $(a - 2)$ (see 36). This formula is used when it is desired to find values of the function near the end of a table. It uses the differences on an upward diagonal line.

A formula will now be derived that employs values on a horizontal line of the difference table. If in (24) elementary substitutions (see 37) are made and x is replaced with $(a + x)$, then there follows an expression for f at $(a + x)$ in terms of f at a and differences of ascending orders (see 38). This formula, often called the Newton-Gauss formula of interpolation, is convenient for use when the value of the argument, for which the function is required, is near the middle of the table of differences. From Newton's formula (28) others can be derived.

Numerical differentiation. By means of a table of a function, the successive differential coefficients of the function can be approximated, and this process is called numerical differentiation. From the forward difference formula (35), differentiating with respect to x and letting $x \to 0$, it follows that a formula for the first derivative of f is obtained (see 39). From formula (36) it follows in a similar way that a difference expression for the first derivative of a function is obtained (see 40). Another example is furnished by differentiating the Newton-Gauss formula of interpolation (38). The higher derivatives may be obtained in the same manner, but the error in their evaluation increases with their order.

Numerical integration. The approximate value of an integral can be expressed as a linear function of a certain number of values of the function to be integrated, and this process is called numerical integration or mechanical quadrature. Integrating from a to b, the function $f(x)$ given by Lagrange's formula (28) gives an approximation formula for the integral of a function (see 41) in which a remainder term is expressible as an integral (see 42). The coefficients of $f(x_s)$ depend upon the interpolation points $x_0, x_1, \cdots, x_n$ but are independent of the form of $f(x)$. If in this formula equidistant arguments $x_s = a + sh$, in which $s = 0, 1, \cdots n$ are taken and $(b - a)/n$ is equal to h, the 18th-century English mathematician Roger Cotes's formula for an integral results (see 43), in which integration is reduced to the evaluation of coefficients (see 44). The value of the coefficients $c_{n,s}$ were calculated by Cotes for $n = 1, 2, 3, \cdots 10$. There is a large number of other numerical integration formulas.

Summation of series. To evaluate the sum of a function ϕ evaluated at integer points x to $x + (n - 1)$ (see 45) it is sufficient to find a function $f(x)$ so that $\Delta f(x)$ is equal to $\phi(x)$. Then by summation the desired sum is found to be a simple difference between the function at $(x + n)$ and the function at x (see 46). Thus from the difference formula for factorials (see 47) it follows that the

$$
(34)\quad \begin{cases}
f(x) = f(x_1) + \dfrac{x - x_1}{\rho_1(x, x_1)} \\
f(x) = f(x_1) + \dfrac{x - x_1}{\rho_1(x_1, x_2) + \dfrac{x - x_2}{\rho_2(x, x_1, x_2) - f(x_1)}} \\
f(x) = f(x_1) + \dfrac{x - x_1}{\rho_1(x_1, x_2) + \dfrac{x - x_2}{\rho_2(x_1, x_2, x_3) - f(x_1) + R_1}} \\
R_1 = \dfrac{x - x_3}{\rho_3(x, x_1, x_2, x_3) - \rho_1(x_1, x_2)} \\
\cdot \\ \cdot \\ \cdot
\end{cases}
$$

$$
(35)\quad f(a + x) = \sum_{s=0}^{n-1} \binom{x}{s} \Delta^s f(a) + \binom{x}{n} f^{(n)}(a + \xi)
$$

$$
(36)\quad f(a - x) = \sum_{s=0}^{n-1} (-1)^s \binom{x}{s} \Delta^s f(a - s) + \\
+ (-1)^n \binom{x}{n} f^{(n)}(a - \xi)
$$

$$
(37)\quad x_{2s} = a - s, \qquad x_{2s+1} = a + s
$$

$$
(38)\quad \begin{cases}
f(a + x) = f(a) + \dfrac{x}{1} \Delta f(a - 1) + \dfrac{x(x + 1)}{2!} \Delta^2 f(a - 1) + \\
\quad + \dfrac{x(x^2 - 1)}{3!} \Delta^3 f(a - 2) + \\
\quad + \dfrac{x(x^2 - 1)(x + 2)}{4!} \Delta^4 f(a - 2) + \cdots
\end{cases}
$$

$$
(39)\quad f'(a) = \sum_{s=1}^{n-1} (-1)^{s-1} \frac{\Delta^s f(a)}{s} + \frac{(-1)^{n-1}}{n} f^{(n)}(a + \xi)
$$

$$
(40)\quad f'(a) = \sum_{s=1}^{n-1} \frac{\Delta^s f(a - s)}{s} + \frac{1}{n} f^{(n)}(a - \xi)
$$

$$
(41)\quad \int_a^b f(x)\, dx = \sum_{s=0}^{n} \frac{f(x_s)}{\phi'(x_s)} \int_a^b \frac{\phi(x)}{x - x_s}\, dx + R
$$

$$
(42)\quad \begin{cases}
R = \displaystyle\int_a^b \phi(x)\, [x, x_0, x_1, \cdots, x_n]\, dx \\
\phi(x) = (x - x_0)(x - x_1) \cdots (x - x_n)
\end{cases}
$$

$$
(43)\quad \int_a^b f(x)\, dx = (b - a) \sum_{s=0}^{n} c_{n,s} f(a + sh)
$$

$$
(44)\quad c_{n,s} = \frac{1}{n} \int_0^n \binom{t}{s} \binom{n - t}{n - s}\, dt
$$

$$
(45)\quad \sum_{s=0}^{n-1} \phi(x + s)
$$

$$
(46)\quad \sum_{s=0}^{n-1} \phi(x + s) = f(x + n) - f(x)
$$

$$
(47)\quad \Delta x^{(m+1)} = (m + 1) x^{(m)}
$$

sum of a factorial can be calculated (see 48). If $m = 2$, this gives a known arithmetic expression (see 49). For a factorial expression of the form with negative index (see 50) ϕ can be easily calculated (see 51). Thus, by (46), if $n \to \infty$, the summation of a new series is achieved (see 52). Special cases follow (see 53).

Functional equations. A relation that is characteristic of a function and more or less determines its form is called a functional equation. Differential and difference equations are examples of functional equations. The functional equation in which a function f is restricted so that the sum of two of its values equals its value at the sum of the arguments (see 54) has the solution $f(x) = cx$, in which c is an arbitrary constant (see 55). Cauchy has proved that if a solution is continuous it will always be of the above form. Another 19th-century French mathematician, Gaston Darboux, showed that if a solution is only bounded in an interval it will still be of the same form, but a 20th-century German mathematician, Georg Karl Wilhelm Hamel, and a 19th–20th-century French mathematician, Henri-Léon Lebesgue, proved the existence of discontinuous solutions of a complicated form.

Similarly, it is seen that the equation in which a function f is restricted so that the product of two of its values equals its value at the sum (see 56) has the solution $f(x) = c^x$, in which again c is an arbitrary constant, for if the logarithm of both sides is taken, this equation is referred to the preceding one.

The functional equation involving a specific linear combination of functional values (see 57) has a solution expressed in logarithms and c_1 and c_2 (see 58), in which c_1 and c_2 are functions of x unaffected by the change of x into $2x$ and $1/c_3$ is the constant 8 $(\log 2)^2$. This can easily be verified by substitution in the equation.

Difference equations. First a linear equation of order n with constant coefficients will be considered (see 59). Using a previous expression (see 4), it can always be written in the form of a linear combination of functional values of a function u, with coefficients q_i, the linear combination set equal to 0 (see 60), in which $q_0, q_1, \cdots q_n$ are constants, and it is supposed that q_0 and q_n do not equal zero. Setting $u(x)$ equal to ρ^x gives a linear combination in terms of ρ (see 61). The polynomial whose product with the x power of ρ is restricted by the equation to vanish (see 62) is called the characteristic function of the given equation. If ρ_s is a zero of this function, ρ_s^x will be a solution of the difference equation. If the characteristic function has n unequal zeros $\rho_1, \rho_2, \cdots \rho_n$, there are n particular solutions $\rho_1^x, \rho_2^x, \cdots \rho_n^x$, and the general solution of the difference equation (60) is a linear combination of these (see 63) in which coefficients $\pi_1(x) \cdots \pi_n(x)$ are periodic functions of period one. For example, a particular difference equation involving three terms (see 64) has the solutions 2^x and 3^x, and the general solution is a weighted average of these (see 65).

If the characteristic function has one or more multiple zeros, the function u is written in two factors (see 66). It follows that the value of u at the general point $x + n$ can be computed (see 67). Substitution in a previous formula (60) yields a sum (see 68).

Letting ρ_1 be a zero of multiplicity m, then it will solve a basic equation (see 69). If $s < m$ and ρ equals ρ_1 in a previous equation (68) an expansion is obtained (see 70). The right-hand side vanishes if $v(x)$ is a polynomial of degree $r < m$. The difference equation (60) thus has the solution $\rho_1^x p_1(x)$, in which $p_1(x)$ is a polynomial of degree $m - 1$, and consequently the m solutions $\rho_1^x x^s$ $(s = 0, 1, 2, \cdots m - 1)$ so that a zero of multiplicity m gives a set of m particular solutions. Thus, if the characteristic function $f(\rho)$ has the distinct zeros $\rho_1, \rho_2, \cdots \rho_r$, there is the general solution (see 71) in which the coefficients of the polynomials $p_s(x)$ are periodic functions of unit period. An exemplary equation (see 72) has the solutions 2^x, 2_2^x, and 3^x and the general solution is a linear combination of these (see 73).

Next the important difference equation identifying u at $(x + 1)$ with the product of x and $u(x)$ (see 74) is

(48) $$\sum_{s=0}^{n-1} (x+s)^{(m)} = \frac{(x+n)^{(m+1)} - x^{(m+1)}}{m+1}$$

(49) $$1 \cdot 2 + 2 \cdot 3 + 3 \cdot 4 + \cdots + n(n+1) = \frac{n(n+1)(n+2)}{3}$$

(50) $$f(x) = x^{(-m)} = \frac{1}{(x+1)(x+2)\cdots(x+m)}$$

(51) $$\phi(x) = \Delta f(x) = -mx^{(-m-1)}$$

(52) $$\sum_{s=0}^{\infty} (x+s)^{(-m-1)} = \frac{x^{(-m)}}{m}$$

(53) $$\sum_{s=0}^{\infty} \frac{1}{(x+s)(x+s+1)(x+s+2)} = \frac{1}{2x(x+1)}$$

(54) $$f(x) + f(y) = f(x+y)$$

(55) $$cx + cy = c(x+y)$$

(56) $$f(x)f(y) = f(x+y)$$

(57) $$f(4x) - 4f(2x) + 4f(x) = x$$

(58) $$c_1 x + c_2 x \log x + c_3 x (\log x)^2$$

(59) $$\sum_{i=0}^{n} p_i \Delta^i u(x) = 0$$

(60) $$Q(u(x)) = \sum_{i=0}^{n} q_i u(x+i) = 0$$

(61) $$Q(\rho^x) = \rho^x \sum_{i=0}^{n} q_i \rho^i$$

(62) $$f(\rho) = \sum_{i=0}^{n} q_i \rho^i$$

(63) $$\sum_{s=1}^{n} \pi_s(x)\, \rho_s^x$$

(64) $$u(x+2) - 5u(x+1) + 6u(x) = 0$$

(65) $$\pi_1(x) 2^x + \pi_2(x) 3^x$$

(66) $$u(x) = \rho^x v(x)$$

(67)
$$\begin{cases} u(x+1) = \rho^{x+1}(v(x) + \Delta v(x)) \\ u(x+2) = \rho^{x+2}(v(x) + 2\Delta v(x) + \Delta^2 v(x)) \\ u(x+n) = \rho^{x+n} \sum_{s=0}^{n} \binom{n}{s} \Delta^s v(x) \end{cases}$$

(68) $$Q(\rho^x v(x)) = \rho^x \sum_{s=0}^{n} \frac{\rho^s}{s!} f^{(s)}(\rho)\, \Delta^s v(x)$$

(69) $$f^{(s)}(\rho_1) = 0$$

(70) $$Q(\rho_1^x v(x)) = \sum_{s=m}^{n} \frac{\rho_1^{x+s}}{s!} f^{(s)}(\rho_1) \Delta^s v(x)$$

(71) $$u(x) = \sum_{s=1}^{r} \rho_s^x\, p_s(x)$$

considered. It is satisfied by the Eulerian integral (see 75) for, if x is greater than zero, integration by parts shows that this is the case (see 76). This solution of (74) is called the gamma (Γ) function, and it is denoted by $\Gamma(x)$. Putting $x = 1$ in (75) gives $\Gamma(1) = 1$. If n be a positive integer, it then follows from (74) that $\Gamma(n+1) = n!$ The first attempts to define the gamma function originated in the problem of finding, by a certain interpolation, a function of a real variable x that is continuous when x is positive and reduces to $x!$ when x is a positive integer. When x is large, $\Gamma(x)$ is given approximately by Stirling's formula (see 77). By repeated application of (74) it follows that a recursion relation for Γ is obtained (see 78). The relation can be written in a slightly different form (see 79). It follows from Stirling's formula that the second fraction on the right-hand side tends to 1 as n tends to infinity. Therefore, a limiting expression for Γ is obtained (see 80). This formula is also due to Euler. From this a product formula follows (see 81). A comparison of this with the infinite product for $\sin x$ (see 82) at once gives the functional equation that relates the product of two gamma functions with the sine function (see 83). Inserting $x = \frac{1}{2}$, the result $\Gamma(\frac{1}{2}) = \sqrt{\pi}$ is obtained.

The use of E, Δ, and D

Operational methods in the solution of difference and differential-difference equations. The symbol E is used to denote the operation of giving to x in a subject function the increment 1, and D is used to denote the operation of differentiation (see 84). With this notation, the first difference is expressed in terms of E (see 85). It is convenient to abbreviate by writing $\Delta = E - 1$, and similarly $E = 1 + \Delta$ and $E = e^D$. The meaning of these relations is that the operation indicated by the symbol on the left is equivalent to the series of operations on the right. The operators E, Δ, and D each combine with constant quantities and with each other by the rules of algebra, and they may be treated as if they were mere symbols of quantity. The difference equation (60) may be written symbolically with f and E (see 86) in which $f(x)$ is the characteristic function. It can be solved by the operational method of the 19th-century English logician and mathematician George Boole as follows. By letting $f(x)$ have the zeros $\rho_1, \rho_2, \cdots, \rho_n$, then f can be written as a specific product of factors (see 87). The difference equation (86) can be written as a product of successive operations operating upon u with the result set equal to 0 (see 88), in which the order of the successive operations is indifferent. If the equation constructed from the first such operation (see 89) is solved, a particular solution of the difference equation results because the operator that remains (see 90) when performed on zero must produce zero. If the zeros are all different, n particular solutions ρ_r^x ($r = 1, 2, \cdots, n$) are attained by solving n separate equations (see 91). If the zero ρ_r is repeated m times, however, it is necessary to solve the equation that includes the corresponding operator term raised to the mth power (see 92). An observation is necessary (see 93). A consequence follows (see 94) making possible a simple equation (see 95). The difference equation (86) then has m solutions (see 96). A zero of multiplicity m gives a set of m particular solutions. In all cases the general solution of (86) can be obtained. A more general equation will now be considered (see 97), in which $\phi(x)$ is a given polynomial of degree m. Writing $1 + \Delta$ for E, the symbolic solution of equation (97) is available (see 98).

It is supposed that $f(\rho)$ does not admit the root 1 and $1/f(1+\Delta)$ is expanded in powers of Δ (see 99), stopping at the term in Δ^m, for $\phi(x)$ is nullified by higher powers of Δ. Hence this operator expansion operating upon ϕ (see 100) satisfies the equation (97). To the solution thus obtained the general solution of the associated homogeneous equation (86) is added. The general solution of (97) is then the sum of the solution already obtained plus products of the $\{\pi_r\}$ and the $\{\rho_r^x\}$ (see 101), in which $\pi_1 \cdots \pi_n$ are arbitrary periodic functions of period one. For example, the equation in which values of u at four points are weighted with coefficients and set equal to x (see 102) has a characteristic function that is

(72) $u(x+3) - 7u(x+2) + 16u(x+1) - 12u(x) = 0$

(73) $\pi_1 2^x + \pi_2 2^x x + \pi_3 3^x$

(74) $u(x+1) = xu(x)$

(75) $u(x) = \int_0^\infty e^{-t} t^{x-1}\, dt$

(76) $\int_0^\infty e^{-t} t^x\, dt = x \int_0^\infty e^{-t} t^{x-1}\, dt$

(77) $\Gamma(x) \sim x^x e^{-x} \sqrt{\frac{2\pi}{x}}$

(78) $\Gamma(x) = \frac{\Gamma(x+n)}{x(x+1)\cdots(x+n-1)}$

(79) $\Gamma(x) = \frac{n^x(n-1)!}{x(x+1)\cdots(x+n-1)} \frac{\Gamma(x+n)}{n^x \Gamma(n)}$

(80) $\Gamma(x) = \lim_{n\to\infty} \frac{n^x(n-1)!}{x(x+1)\cdots(x+n-1)}$

(81) $\Gamma(x)\Gamma(1-x) = \frac{1}{x} \prod_{n=1}^{\infty} 1 \Big/ \left(1 - \frac{x^2}{n^2}\right)$

(82) $\sin x = x \prod_{n=1}^{\infty} \left(1 - \frac{x^2}{n^2\pi^2}\right)$

(83) $\Gamma(x)\Gamma(1-x) = \frac{\pi}{\sin \pi x}$

(84) $Eu(x) = u(x+1), \qquad Du(x) = \frac{du(x)}{dx}$

(85) $\Delta u(x) = Eu(x) - u(x) = (E-1)u(x)$

(86) $f(E)u(x) = 0$

(87) $f(E) = (E-\rho_1)(E-\rho_2)\cdots(E-\rho_n)$

(88) $(E-\rho_1)(E-\rho_2)\cdots(E-\rho_n)u(x) = 0$

(89) $(E-\rho_n)u(x) = 0$

(90) $(E-\rho_1)(E-\rho_2)\cdots(E-\rho_{n-1})$

(91) $(E-\rho_r)u(x) = 0$

(92) $(E-\rho_r)^m u(x) = 0$

(93) $f(E)\rho^x u(x) = \rho^x f(\rho E)u(x)$

(94) $\begin{cases} (E-\rho_r)^m u(x) = \rho_r^x (\rho_r E - \rho_r)^m \rho_r^{-x} u(x) \\ \quad = \rho_r^{x+m} \Delta^m \rho_r^{-x} u(x) = 0 \end{cases}$

(95) $\Delta^m \rho_r^{-x} u(x) = 0$

(96) $\rho_r^x x^s, \qquad s = 0, 1, 2, \cdots, m-1$

(97) $f(E)u(x) = \phi(x)$

(98) $\frac{1}{f(1+\Delta)} \phi(x)$

easily determined (see 103) and can be evaluated formally at $1 + \Delta$ (see 104). It follows that a linear expression in x (see 105) is a particular solution, and the general solution is obtained from this by adding three products of the general type described (see 106).

Differential-difference equations, it has been noted, are functional equations involving derivatives and differences. There is a very close similarity between the theory of differential-difference equations and the theory of difference equations. It has been shown that every solution of a linear homogeneous difference equation can be written as a linear combination of a finite number of particular solutions. Solutions of linear homogeneous differential-difference equations with constant coefficients can likewise be written as sums of independent particular solutions, but there are infinitely many such solutions that must be found by transcendental methods, and the entire theory is accordingly more complicated. As an example, an equation that is first order and inhomogeneous (see 107) may be considered in which q_0 and q_1 are given constants, and $q_0 + q_1 \neq 0$, $q_1 \neq 0$. The operator $f(D)$ is defined (see 108). The equation is expressible in operator form (see 109). A symbolic solution is obtained by performing the inverse operation (see 110). Expanding in powers of D gives the reciprocal of the operator $f(D)$ (see 111). If $p(x)$ is a polynomial of degree n, $D^s p(x) = 0$, if $s > n$, and thus (see 112) in terms of powers of D operating upon p a particular solution is obtained. Now the differentiation properties of the exponential (see 113) are relevant. Then $\exp(\rho_s^x)$ is a solution of the homogeneous equation (see 114) if and only if ρ_s is a zero of the characteristic function (see 115). There are, in general, infinitely many zeros $\rho_1, \rho_2, \rho_3, \cdots$. The homogeneous equation thus has solutions of the form of a linear combination of exponentials (see 116) in which $c_1, c_2, c_3, \cdots$ are arbitrary constants decreasing sufficiently to ensure convergence, and the equation to be solved (107) therefore has a general solution (see 117) composed of the sum of u and the stated combination of exponentials. (N.E.N.)

(99) $$\frac{1}{f(1+\Delta)} = a_0 + a_1\Delta + a_2\Delta^2 + \cdots + a_m\Delta^m$$

(100) $$(a_0 + a_1\Delta + \cdots + a_m\Delta^m)\phi(x)$$

(101) $$(a_0 + a_1\Delta + \cdots + a_m\Delta^m)\phi(x) + \sum_{r=1}^{n} \pi_r \rho_r^x$$

(102) $$u(x+3) - 4u(x+2) + u(x+1) + 6u(x) = x$$

(103) $$f(E) = E^3 - 4E^2 + E + 6 = (E+1)(E-2)(E-3)$$

(104) $$f(1+\Delta) = 4 - 4\Delta - \Delta^2 + \Delta^3$$

(105) $$\frac{1+\Delta}{4}x = \frac{x+1}{4}$$

(106) $$\frac{x+1}{4} + \pi_1(-1)^x + \pi_2 2^x + \pi_3 3^x$$

(107) $$u'(x+1) + q_0 u(x+1) + q_1 u(x) = p(x)$$

(108) $$f(D) = De^D + q_0 e^D + q_1$$

(109) $$f(D)u(x) = p(x)$$

(110) $$u(x) = \frac{1}{f(D)}p(x)$$

(111) $$\frac{1}{f(D)} = b_0 + b_1 D + b_2 D^2 + \cdots$$

(112) $$u(x) = \sum_{s=0}^{n} b_s D^s p(x)$$

(113) $$f(D)e^{\rho x} = f(\rho)e^{\rho x}$$

(114) $$f(D)u(x) = 0$$

(115) $$f(\rho) = \rho e^{\rho} + q_0 e^{\rho} + q_1$$

(116) $$\sum_{s=1}^{\infty} c_s e^{\rho_s x}$$

(117) $$\sum_{s=0}^{n} b_s D^s p(x) + \sum_{s=1}^{\infty} c_s e^{\rho_s x}$$

(118) $$[x_i, x_{i+1}], \qquad i = 0, 1, 2, \cdots, (n-1)$$

(119) $$f(x_i) = P_1(x_i), \qquad i = 0, 1, 2, \cdots, n$$

(120) $$D^k f(x_i) = D^k P_3(x_i) \qquad k = 0, 1;\ i = 0, 1, 2, \cdots, n$$

(121) $$D^k f(x_i) = D^k P_{2m-1}(x_i), \qquad \begin{cases} 0 \leq k \leq m-1 \\ 0 \leq i \leq n \end{cases}$$

APPLICATIONS OF NUMERICAL ANALYSIS

Approximation. Approximation of a function $f(x)$ by a simpler function $A(x)$ is one of the most widely used ideas in numerical analysis. The difference $f(x) - A(x)$ is the error of the approximation, and this must be kept reasonably small over the range of x considered. The interpolating functions mentioned previously are examples of approximating functions in which the error is zero at selected points along the x-axis. An interpolating function, however, may not give a good fit between these selected points (nodes), and the first derivatives of the interpolating function may bear little resemblance to the first derivatives of the original function $f(x)$. Accordingly, it is often advantageous to break up the range of x, over which an approximation is required, into a number of non-overlapping subintervals and to fit each subinterval with a polynomial of low degree. For example, if there are n subintervals (see 118) the piecewise linear approximating function, denoted by $P_1(x)$, satisfies $n + 1$ restrictions (see 119). This approximating function may well give a good fit between the nodes, but of course contributes nothing toward obtaining some agreement between the first derivatives of the function and its approximation. Consequently, the possibility of constructing a piecewise approximating function that has the same values of function and first derivative as $f(x)$ at the nodes, x_i $(i = 0, 1, 2, \cdots, n)$ is considered. Such a function is a piecewise cubic polynomial, denoted by $P_3(x)$, in which its function values and first derivatives are matched to those of f at $n + 1$ points (see 120). In general, the piecewise polynomial $P_{2m-1}(x)$ in which functional values and $m - 1$ derivatives are matched at $n + 1$ points (see 121) interpolates the function and its first $(m - 1)$ derivatives at the nodes x_i $(i = 0, 1, 2, \cdots, n)$. Such approximating functions are known as piecewise Hermite interpolates, named for a 19th-century French mathematician, Charles Hermite.

Hermite interpolates

In problems in which the approximate has to match $f(x)$ but not $Df(x)$, $D^2f(x)$, etc., it is undesirable to introduce derivatives of $f(x)$ as additional parameters (arbitrary constants). Consequently, a desirable property in piecewise functions might be continuity of the derivatives at the nodes without specifying the values of the derivatives there. Piecewise approximates of this type are known as splines, the most popular form of which is the cubic spline. Given the values of f_i $(i = 0, 1, 2, \cdots, n)$ a cubic polynomial is fitted between successive pairs of nodes and continuity of both first and second derivatives is required at all internal nodes.

The natural form of the cubic spline for equally spaced nodes distance h apart in the range $[0, b]$ is a cubic in

(x/h) (see 122) in which a specialized notation is used (see 123). Applying the condition that the spline should equal the function at each of $n + 1$ points (see 124), there are $(n + 1)$ linear equations for the $(n + 3)$ coefficients, and so there are two free parameters, which can be fixed arbitrarily.

Approximation of functions of two space variables over rectangular regions can be carried out by splitting the region up using a rectangular grid, and using piecewise bivariate Hermite functions or splines. These are tensor products of the univariate Hermite functions or splines described in this section. When the region is not rectangular, approximation is very difficult even with piecewise approximating functions.

So far the approximates have agreed with $f(x)$ at $(n + 1)$ points in the range. This of course is not necessary, and two frequently used measures of the error function $e(x)$ are the integral of the square of the error function, integrated on $\mathfrak{R}$, and the maximum over $\mathfrak{R}$ of the absolute value of the error function (see 125, 126) in which $\mathfrak{R}$ is the range of x considered. The approximating function $A(x)$ is chosen to make one or other of these expressions as small as possible. The former is called the least squares measure and the latter the maximum measure of the error. With regard to the latter measure a very important result exists. It is that among all polynomials of degree n or less, a unique polynomial exists that minimizes, for $x \in \mathfrak{R}$, max $|f(x) - P_n(x)|$. This minimum value of size E is taken with alternating sign at $(n + 2)$ arguments in $\mathfrak{R}$.

Finally, rational functions are mentioned as a possible class of approximating functions. These functions are ratios of polynomials. They are particularly suitable when $f(x)$ has one or more singularities (or poles) within R. The denominator of the rational function is chosen to have zeros coinciding with the poles of the original function $f(x)$.

Differentiation and integration. Numerical differentiation involves finding the derivatives of functions that are either given as a table of values or are given analytically in such a complicated manner that the exact derivatives are exceedingly tedious to evaluate. The standard procedure is to obtain a polynomial approximation $p(x)$ to the function $f(x)$ and to accept derivatives of the approximation (see 127) as approximations to the derivatives of the function (see 128). In the case of the function given as a table of values, the polynomial $p(x)$ is an interpolating polynomial, and examples of these for equally and unequally spaced data were given earlier. Alternatively, the function with a complicated form $f(x)$ may be replaced, for example, by a cubic spline, its having the same values of function and first two derivatives as $f(x)$ at selected points in the range. It should be pointed out that numerical differentiation is unquestionably inaccurate in many cases because two functions, say $f(x)$ and $p(x)$, can be very close together over a range of x and nevertheless have very different slopes in this range. The accuracy also diminishes with increasing order of the derivatives.

Numerical integration is one of the most successful operations in numerical analysis. In most integration formulas the integrand $f(x)$ is replaced by an approximating polynomial, and it is integrated exactly over the required range. The approximating polynomial is usually a high degree interpolating polynomial over the complete range or a piecewise polynomial in which the range has been divided into several equal parts with a simple polynomial over each part. Very popular examples of the latter are the trapezoidal rule (piecewise linear; see 129) and the rule of the 18th-century English mathematician Thomas Simpson (piecewise quadratic; see 130), each involving h, in which h is the length of a part.

So far the numerical integration formulas considered have been of the form of a sum of products of coefficients and functional values evaluated at arguments (see 131), in which the arguments x_i have been equally spaced distance h apart. The possibility of choosing the arguments x_i and the coefficients A_i so that the numerical integration formula is exact for $f(x)$ being any polynomial of degree $\leq 2n + 1$ is worth consideration. This in fact can be done and leads to the development of Gaussian quadrature formulas. These formulas, as might be expected, are more accurate than formulas with equally spaced arguments, but they require extra work because the location of the arguments must be found in advance.

(122) $$\begin{cases} S_{0,b}\left(\frac{x}{h}\right) = \alpha_0 + \alpha_1\left(\frac{x}{h}\right) + \alpha_2\left(\frac{x}{h}\right)^2 + \alpha_3\left(\frac{x}{h}\right)^3 + \\ \qquad + \sum_{s=1}^{n-1} \beta_s \left(\frac{x}{h} - s\right)_+^3 \end{cases}$$

(123) $$\left(\frac{x}{h} - s\right)_+ = \begin{cases} 0 & \frac{x}{h} \leq s \\ \left(\frac{x}{h} - s\right) & \frac{x}{h} > s \end{cases}$$

(124) $$S_{0,b}\left(\frac{x_i}{h}\right) = f_i, \qquad i = 0, 1, 2, \cdots, n$$

(125) $$\int_{x \in R} [e(x)]^2 \, dx$$

(126) $$\max_{x \in R} |e(x)|$$

(127) $$p'(x), \; p^{(2)}(x), \; p^{(3)}(x), \; \ldots$$

(128) $$f'(x), \; f^{(2)}(x), \; f^{(3)}(x), \; \ldots$$

(129) $$\int_{x_0}^{x_n} f(x)\,dx \doteq \tfrac{1}{2}h[f_0 + 2f_1 + \cdots + 2f_{n-1} + f_n]$$

(130) $$\int_{x_0}^{x_n} f(x)\,dx \doteq \tfrac{1}{3}h[f_0 + 4f_1 + 2f_2 + 4f_3 + \cdots + 2f_{n-2} + 4f_{n-1} + f_n]$$

(131) $$\sum_{i=0}^{n} A_i f_i$$

Computational methods in linear algebra. The solution of many linear problems in science and engineering reduces to the mathematical problem of solving a set of linear equations. In matrix form, the set of equations can be represented by $A\mathbf{x} = \mathbf{b}$, in which A is a square matrix of order n with real elements, and $\mathbf{x}$, $\mathbf{b}$ are column matrices (that is, matrices composed of one column), each with n real elements. The matrix A is assumed to be nonsingular; that is $|A| \neq 0$, in which $|A|$ is the determinant of A. The problem is to find the elements of the vector $\mathbf{x}$, given the elements of A and $\mathbf{b}$. This simple looking problem can be deceptively difficult.

There are two principal types of method for solving linear systems, direct and iterative. The direct method reaches the solution after a finite number of different operations. The iterative method begins with a guess at the solution and by repeated use of the same operation guides the original guessed value toward the required solution. Iterative methods are usually employed when the matrix A has a particular structure incorporating large blocks of zero elements.

Direct method of solution

The most popular direct method of solution is Gaussian elimination, which is now described. Rows and columns of A are interchanged until the maximum modulus element is in the top left-hand position. This is the first pivot and is called a_{11}. The first equation is divided by the first pivot. Now the first equation is multiplied by a_{21}, the leading element in the second row, and subtracted from the second equation; then the first equation is multiplied by a_{31}, the leading element in the third row, and subtracted from the third equation; and so on. This annihilates the

first column except for unity in the first equation. The first equation is thus obtained after being divided by a_{11} together with a system of $(n-1)$ equations. The maximum modulus element is then moved up to the top left-hand corner of the $(n-1)$ block by interchanging rows and columns, and the process repeated. If $(n-1)$ such eliminations are performed, the resulting linear system, which is triangular, is equivalent to the original linear system (*i.e.*, it has the same solution). The triangular system is easily solved by back substitution and so the solution of the original system is obtained. If A is symmetric and positive definite, the method described can be considerably simplified. The modified method is known as Choleski's method.

Iterative methods

The iterative methods used originally for solving linear systems are those of the 19th-century German mathematician Karl Gustav Jacob Jacobi and Gauss-Seidel (the 19th-century German astronomer P.L. Seidel). In recent years, however, the method of successive overrelaxation (S.O.R.) has been universally accepted as the best iterative method. In this method, the matrix A is written as $(I\text{-}L\text{-}U)$, in which I is the unit matrix and L, U are lower and upper triangular matrices, respectively, with zero entries on the main diagonals. The S.O.R. method is given by a relation between successive values of the vector that involves $\mathbf{x}_0$ and w (see 132) in which $\mathbf{x}_0$ is the guessed value of the solution, and w is a parameter to be determined, which probably lies in the range $1 \leq w \leq 2$. The value of w that gives the fastest convergence of the iterative method in a particular problem is usually obtained empirically.

Finding eigenvalues and eigenvectors

Another computational problem in linear algebra, which is just as important as solving a set of linear equations, is finding the eigenvalues and associated eigenvectors of a matrix A with real elements. In mathematical terms this reduces to finding the numbers λ (written as Greek letters lambda) for which the linear system $(A - \lambda I)\,\mathbf{x} = 0$ has a solution other than $\mathbf{x} = \mathbf{0}$. Each such number λ, which can be real or complex, is called an eigenvalue of A, and the corresponding solution vector $\mathbf{x}$ ($\neq \mathbf{0}$) is called the associated eigenvector. In general, a matrix A of order n has n eigenvalues and n associated eigenvectors. If only a few eigenvalues are required, iterative methods are usually employed, but if all the eigenvalues are required, direct methods involving similarity transformations are often used. These transformations reduce the matrix A to simpler form without altering the eigenvalues. The direct methods in common use for symmetric matrices are those of the U.S. mathematicians Wallace Givens and Alston S. Householder, both of which reduce A to tridiagonal form. This is a form with zero elements everywhere except on the principal diagonal and the two subdiagonals adjacent and on opposite sides of the main diagonal. Standard procedures exist for finding the complete eigenvalue system for a tridiagonal matrix.

Nonlinear equations. In solving a set of linear equations $A\mathbf{x} = \mathbf{b}$ in which A is a square matrix, it was accepted that a unique solution existed provided $|A| \neq 0$; that is, n linearly independent linear equations are required to find n unknowns. This simple guideline no longer applies with nonlinear equations, for it is seen immediately that one nonlinear equation may have any number of solutions. For example, $x^2 - 2 = 0$ has two solutions, $x = \pm\sqrt{2}$, and $\sin x = 0$ has an infinity of solutions $x = r\pi$, in which r is any integer. It is a nontrivial matter to solve one nonlinear equation in one unknown, and this case is examined first, writing the single equation in the form $f(x) = 0$. Although x is used for the unknown quantity, in fact x can lie anywhere in the complex plane. The equation is now rewritten as $x = g(x)$, which is not a unique form in any particular example. The second degree polynomial equation (see 133), for example, can be rewritten as any one of three or more forms (see 134). Any of the forms motivate the iterative method in which successive values of x, designated by symbols with superscripts, are related through some function g (see 135) with the appropriate form of g, in which x_0 is an initial guess at a root of the equation. The method is successful if the iteration converges to the required root. In the above example, which has $x = 2$ as a solution, the iterative process (see 136) diverges for any starting value other than 2, whereas another possible process (see 137) converges to the required solution for a wide range of possible starting values. The general convergence condition for the first order (or linear) iterative method described is given in the following theorem. Letting $x = X$ be a root of $f(x) = 0$, and I be an interval containing the point $x = X$ and also the initial approximation x_0, and if the absolute value of the derivative is bounded by a number less than 1 (see 138) for all points in I, then the iteration expressed in terms of a general function g (see 139) converges to the root X.

The absolute value of the derivative (see 140) determines the rate of convergence of the process, and it is advantageous to make it as small as possible, say zero. Such an iterative procedure is second order (or quadratic), and the most popular method of this class owes its origin to Newton in which a specific form of g (see 141) leads to the iterative method in which the $(k+1)$th point is obtained from the kth by subtracting the ratio of the function to its derivative (see 142) measured at the kth point.

The practical solution of systems of nonlinear equations is significantly more difficult than the solution of a single equation, although most of the methods for a single equation extend in an obvious manner to systems. The importance of finding a good method for solving nonlinear systems cannot be stressed enough, since many nonlinear problems in science and engineering reduce to the mathematical problem of solving a set of nonlinear equations.

The extension of Newton's iterative method to a system is indicated because this extension forms the starting point for many practical methods of solving nonlinear systems. The system of equations is written as $f_i(x_1, x_2, \cdots x_n) = 0$ $(i = 1, 2, \cdots n)$, and the square matrix, the Jakobian J, is introduced (see 143). Newton's

(132) $$\begin{cases}\mathbf{x}_{i+1} = (I - wL)^{-1}[wU + (1-w)I]\mathbf{x}_i + w(I - wL)^{-1}\mathbf{b} \\ i = 0, 1, 2, \cdots\end{cases}$$

(133) $$f(x) \equiv x^2 - x - 2 = 0$$

(134) $$x = x^2 - 2, \qquad x = 1 + \frac{2}{x}, \qquad x = (2+x)^{1/2}$$

(135) $$x^{k+1} = g(x^k), \qquad k = 0, 1, 2, \cdots$$

(136) $$x^{k+1} = (x^k)^2 - 2, \qquad k = 0, 1, 2, \cdots$$

(137) $$x^{k+1} = (2 + x^k)^{1/2}$$

(138) $$\left|\frac{dg(x)}{dx}\right| \leq K < 1$$

(139) $$x^{k+1} = g(x^k), \qquad k = 0, 1, 2, \cdots$$

(140) $$\left|\frac{dg(x)}{dx}\right|$$

(141) $$g(x) = x - \frac{f(x)}{\dfrac{df(x)}{dx}}$$

(142) $$x^{k+1} = x^k - \left(\frac{f(x)}{\dfrac{df(x)}{dx}}\right)^k, \qquad k = 0, 1, 2, \cdots$$

$$(143)\quad \begin{cases} J(x_1, x_2, \cdots, x_n) = \left[\dfrac{\partial f_i(x_1, x_2, \cdots, x_n)}{\partial x_j}\right] \\ i, j = 1, 2, \cdots, n \end{cases}$$

$$(144)\quad \mathbf{x}^{k+1} = \mathbf{x}^k - J^{-1}(\mathbf{x}^k) f(\mathbf{x}^k), \qquad \begin{cases} k = 0, 1, 2, \cdots \\ |J(\mathbf{x})| \neq 0 \end{cases}$$

$$(145)\quad f_i(\mathbf{x}^k) + \sum_{j=1}^{n} f_{ij}(\mathbf{x}^k)(x_j^{k+1} - x_j^k) = 0, \qquad \begin{cases} i = 1, 2, \cdots, n \\ k = 0, 1, 2, \cdots \end{cases}$$

$$(146)\quad x_1^{k+1}, x_2^{k+1}, \cdots, x_n^{k+1}$$

$$(147)\quad \begin{cases} x_i^{k+1} = x_i^k - w\dfrac{f_i(x_1^{k+1}, \cdots, x_{i-1}^{k+1}, x_i^k, \cdots, x_n^k)}{f_{ii}(x_1^{k+1}, \cdots, x_{i-1}^{k+1}, x_i^k, \cdots, x_n^k)} \\ i = 1, 2, \cdots, n \qquad k = 0, 1, 2, \cdots \end{cases}$$

$$(148)\quad y^n = f(x, y, y', \cdots, y^{(n-1)})$$

$$(149)\quad y(x_0) = A_0, \qquad \cdots, \qquad y^{(n-1)}(x_0) = A_{n-1}$$

$$(150)\quad y' = p, \qquad p' = f(x, y, p)$$

$$(151)\quad \mathbf{Y}' = \mathbf{F}(x, \mathbf{Y}), \qquad \mathbf{Y} = [y, p]^T$$

$$(152)\quad \mathbf{F} = [F_1, F_2]^T$$

$$(153)\quad \mathbf{Y}(x_0) = [A_0, A_1]^T$$

$$(154)\quad y^{(2)} = f(x, y, y')$$

$$(155)\quad y(a) = y_a, \qquad y(b) = y_b$$

$$(156)\quad \begin{cases} k_1 = hf(x, y) \\ k_2 = hf(x + \frac{1}{2}h, y + \frac{1}{2}k_1) \\ k_3 = hf(x + \frac{1}{2}h, y + \frac{1}{2}k_2) \\ k_4 = hf(x + h, y + k_3) \\ y(x + h) = y(x) + \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4) \end{cases}$$

$$(157)\quad \begin{cases} y_{k+1} = y_{k-3} + \dfrac{4h}{3}(2y'_{k-2} - y'_{k-1} + 2y'_k) \\ y_{k+1} = y_{k-1} + \dfrac{h}{3}(y'_{k+1} + 4y'_k + y'_{k-1}) \end{cases}$$

$$(158)\quad \tfrac{14}{45} h^5 y^{(5)} \quad -\tfrac{1}{90} h^5 y^{(5)}$$

$$(159)\quad \sum_{i=1}^{n} c_i e^{\lambda_i x}$$

method for the system then can be expressed in vector form (see 144). If the equation expressing Newton's method is multiplied through by J, this becomes (see 145) a linear system of equations of order n to be solved for the successive values of the vector exponents (see 146) at each iteration. A practical variant of this is the generalized Newton's method involving parameter w (see 147) in which w is to be determined empirically in order to speed up the convergence.

Ordinary differential equations. The numerical solution of ordinary differential equations is one of the most important branches of numerical analysis because many physical problems lead to ordinary differential equations that cannot be solved analytically. The two main types of problem are initial value and boundary value. The former requires a function $y(x)$ to be found for $x \geq x_0$ which satisfies an nth-order equation (148) and takes initial values (149) consisting of the function and its first $n - 1$ derivatives at $x = x_0$. When $n \geq 2$, the single differential equation can be replaced by a system of first order equations. For example, if $n = 2$, the second order equation becomes a first order system (see 150) with the initial conditions $y(x_0) = A_0$, $p(x_0) = A_1$. This can be written in vector form (see 151, 152) in which vector components are $F_1 = p$ and $F_2 = f(x, y, p)$. The initial condition is also expressible in vector form (see 153). A typical boundary value problem is given by a second order differential equation (see 154) together with the two boundary values at points a and b (see 155). The solution is required for $a \leq x \leq b$.

In the vast majority of numerical methods for ordinary equations the differential equation is replaced by a difference equation. A discrete set of equally placed arguments x_k distance h apart is chosen, and a sequence of corresponding values $y(x_k)$ is calculated from the difference formula.

The Runge-Kutta and Milne methods

The initial value problem for a single first order equation may be considered first. The most popular classes of methods here are Runge-Kutta (named for the 19th–20th-century German mathematician Carl David Tolmé Runge and the 20th-century German aerodynamicist M. Wilhelm Kutta) and predictor-corrector multistep methods. An example of the former is easily constructed (see 156) and an example of the latter (see 157) is Milne's method (named for the 20th-century U.S. mathematician and astrophysicist William Edmund Milne) in which the predictor is an explicit formula, and the corrector is a more accurate implicit formula. The accuracy of these methods is measured by the local truncation obtained after Taylor expansions (named for the 18th-century English mathematician Brook Taylor) have been carried out for the various terms in the difference formula. These are positive and negative constants times fifth order terms (see 158) for the predictor and corrector formulas of Milne's method, respectively.

There is little point in a method being locally accurate, however, if it is not convergent and stable. A method is convergent if as the grid size h tends to zero, the approximate solutions converge to the exact solution of the problem. Convergence is concerned with truncation error only. A method is stable if a round-off error, committed anywhere in the calculation, propagates in such a way that it does not significantly influence the true solution. Necessary and sufficient conditions for the stability of a multistep method can be given in terms of the roots of the characteristic polynomial. For example, in the corrector formula of Milne's method, if $h = 0$, and $y_k = C\lambda^k$, in which C is a constant, then the characteristic polynomial is $\lambda^2 - 1$ with roots ± 1. This method is stable in the sense that the roots are distinct and lie on the unit circle.

In any calculation, however, the grid size h is not zero, and if the true solution is decreasing, one of the characteristic roots is greater than unity, and so Milne's method is relatively unstable. An exact stability analysis of a multistep method for finite h is a difficult and tedious task. On the other hand, Runge-Kutta methods are always stable for sufficiently small grid size, and so convergence theorems are much more straightforward than for multistep methods.

Many of the numerical methods for solving a single equation carry over to a system. A complication in the calculation of the solution arises, however, if the system is "stiff." To illustrate stiffness in a system, the linear system $\mathbf{y}' = A\mathbf{y}$, of order n, in which the eigenvalues of A are distinct, is considered. The general solution of this system is given by a linear sum of exponentials (see 159) in which $\{\lambda_i\}$ appearing in the exponent are the eigenvalues of A and $\{c_i\}$ are constants. If the real parts of some of the eigenvalues of A are negative and very much larger than the real parts of others, the terms corresponding to the large eigenvalues die away quickly and the system is said to be stiff. When difference methods are applied to the linear system, stability restrictions on the grid size take the form $\{|h\lambda i| < d\}$, in which d is a fixed quantity. If some of the $|\lambda_i|$ are large, the permissible step forward is small even when the corresponding contribution to the true solution is negligible. A simple method that is very satisfactory with stiff equa-

tions is the trapezoidal rule. For the linear system this takes a form (see 160) that is easily expressed. For a nonlinear system $y^1 = f(x, y)$, the stiffness analysis above applies except that the eigenvalues are now those of the Jacobian matrix (see 161).

Finally, boundary value problems are often solved numerically by shooting methods. In these, additional starting values are assumed to enable initial value methods to be used. Once the right-hand boundary is reached, the calculated value is compared with the boundary condition there, and the additional initial conditions modified until agreement is reached at the right-hand boundary. If, for example, the problem $y^2 = f(x, y, y')$ with $y(a) = y_a$ and $y(b) = y_b$ is solved as an initial value problem using a two-step difference method, the additional condition $y'(a) = c$, in which c is a parameter, must be assumed. The parameter c is modified until the solution at $x = b$ agrees with the value y_b.

Partial differential equations. The rapid growth of high speed computers has led to intense development in the numerical solution of partial differential equations. A method universally applicable to linear and nonlinear problems that give rise to elliptic, parabolic, and hyperbolic partial differential equations is the method of finite differences, and most of what follows will be devoted to this method, followed by the finite element method as a means of obtaining numerical solutions of elliptic problems.

The method of finite differences

The independent variables in a partial differential equation are either all space variables (elliptic case), or the time variable together with one or more space variables (parabolic and hyperbolic cases). The region over which the solution is required is usually closed in an elliptic problem, and open otherwise. The elliptic or steady problem has boundary conditions over the perimeter or surface enclosing the region, whereas the parabolic and hyperbolic or unsteady problems have initial condition(s) at $t = 0$ and boundary conditions on the sides of the open region. In the finite difference method, the region over which the solution is required is covered by a rectangular grid with grid lines parallel to the coordinate axes. The grid spacing in the direction of the space variables is h, whereas the grid spacing in the direction of the time axis (parabolic and hyperbolic cases) is k. If the original region is rectangular in shape, then the rectangular grid can be made to fit the region exactly. If the original region is not rectangular, then complications arise with a rectangular grid in the vicinity of the boundary of the original region.

Initially the elliptic problem is considered, and at each grid point inside the region, each partial differential in the equation is replaced by an equivalent difference replacement. Thus, for each internal grid point there is formed, depending on the equation, a linear or nonlinear expression connecting a finite number of unknown function values. At grid points on the boundary of the region, either the function values (Dirichlet conditions, named for the 19th-century German mathematician Peter Gustav Lejeune Dirichlet) or the normal derivatives are known (Neumann conditions, named for the 19th-century German Karl Neumann), which on discretization produce linear expressions at each boundary grid point. The totality of equations at the internal and boundary grid points leads to the matrix equation $A\mathbf{U} = \mathbf{b}$ in the case of a linear equation and to the nonlinear system of equations $f_i(U_1, U_2, \cdots, U_n) = 0$ $(i = 1, 2, \cdots, n)$ in the case of a nonlinear equation, in which n is the number of grid points at which the solution is required. The iterative method of successive overrelaxation (S.O.R.) has already been described for solving the linear system and the generalized Newton's method for solving the nonlinear system. As the grid spacing h is reduced in an elliptic problem, the number of grid points, and hence the number of equations (linear or nonlinear), to be solved increases. In the limit as $h \to 0$, the theoretical solution of the system of equations should tend to the real solution of the problem. This is the problem of convergence, which is usually not a serious problem with elliptic equations.

(160) $$\mathbf{y}_{r+1} - \mathbf{y}_r = \tfrac{1}{2}h[(A\mathbf{y})_{r+1} + (A\mathbf{y})_r], \qquad r = 0, 1, 2, \cdots$$

(161) $$\frac{\partial(f_1, f_2, \cdots, f_n)}{\partial(y_1, y_2, \cdots, y_n)}$$

(162) $$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}$$

(163) $$U_m^{n+1} = (1 + r\delta_x^2)U_m^n$$

(164) $$(1 - \tfrac{1}{2}r\delta_x^2)U_m^{n+1} = (1 + \tfrac{1}{2}r\delta_x^2)U_m^n$$

(165) $$(1 - \tfrac{1}{2}r\delta_x^2)(1 - \tfrac{1}{2}r\delta_y^2)U_m^{n+1} = (1 + \tfrac{1}{2}r\delta_x^2)(1 + \tfrac{1}{2}r\delta_y^2)U_m^n$$

(166) $$\begin{cases} (1 - \tfrac{1}{2}r\delta_x^2)U_m^{n+\frac{1}{2}} = (1 + \tfrac{1}{2}r\delta_y^2)U_m^n \\ (1 - \tfrac{1}{2}r\delta_y^2)U_m^{n+1} = (1 + \tfrac{1}{2}r\delta_x^2)U_m^{n+\frac{1}{2}} \end{cases}$$

(167) $$\frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial x^2}$$

(168) $$U_m^{n+1} = 2(1 - p^2)U_m^n + p^2(U_{m+1}^n + U_{m-1}^n) - U_m^{n-1}$$

Use in time-dependent problems

In time-dependent problems that lead to parabolic and hyperbolic equations, the finite difference approximation is set up as in the elliptic case. The method of solution of the set of difference equations (linear or nonlinear), however, is entirely different. The calculation starts with the initial condition(s) at $t = 0$ and proceeds in a step-by-step manner in the direction of increasing time. For example, in the parabolic heat conduction equation in one space variable (see 162), two possible difference replacements are an explicit formula (see 163) and the implicit Crank–Nicolson formula (see 164) involving r, δ_x, and U_m^n, in which $r = k/h^2$, δ_x is the standard central difference operator, and U_m^n is the exact difference solution at the grid point $x = mh$, $t = nk$. These formulas advance the calculation from $t = 0$ to $t = k$ $(n = 0)$ then from $t = 1$ to $t = 2k$ $(n = 1)$, and so on. The solution by the explicit formula is straightforward, whereas the implicit formula requires the solution of a set of linear equations at each time step. The principal computational hazard with a calculation advancing in time is that a round-off error will grow as the calculation advances and swamp the true solution. This is the problem of instability. With the two difference methods described above the stability conditions are $0 < r \leq \frac{1}{2}$ for the explicit method and $r > 0$ for the Crank–Nicolson method. With the heat conduction equation in two space dimensions, the unconditionally stable counterpart of the Crank–Nicolson method can be written down (see 165). For ease of calculation, this can be split into two formulas (see 166). Advancing the calculation from $t = nk$ to $t = (n + \frac{1}{2})k$ requires the solution of sets of equations along lines parallel to the x-axis, followed by an advancement from $t = (n + \frac{1}{2})k$ to $t = (n + 1)k$, which requires the solution of sets of equations along lines parallel to the y-axis. This method was first introduced by Peaceman and Rachford and is known as an alternating direction implicit (A.D.I.) method.

In calculations involving hyperbolic equations, extra care must be taken to ensure that the rate of advancement of a calculation, as given by the difference formula, is not in excess of the rate governed by the characteristic curves. For example, in the case of the wave equation in one space dimension (see 167) the standard explicit difference replacement includes the parameter p (see 168) in which $p = k/h$. The characteristic curves of the wave equation consist of two families of straight lines inclined at angles $\pm \pi/4$ respectively to the t-axis. The lines of advancement or wave fronts of the finite difference calculation must make angles of less than $\pi/4$ with the

$$\alpha(x)y(x) = F(x) + \lambda \int_a^b K(x, \xi; y(\xi))\, d\xi \tag{169}$$

$$\alpha(x)y(x) = F(x) + \lambda \int_a^x K(x, \xi; y(\xi))\, d\xi \tag{170}$$

$$y(a),\ y(a+h),\ y(a+2h),\ \cdots,\ y(b) \tag{171}$$

positive and negative directions of the t-axis, and this leads to the condition $0 < p \leq 1$ on the grid ratio. This is the Courant-Friedrichs-Lewy condition, which is all important in an explicit difference calculation of a hyperbolic equation.

There are, of course, methods other than the finite difference method for obtaining numerical solutions of partial differential equations. Such a method, used extensively, particularly in elliptic problems, is the finite element method. Here the region under consideration is divided up into a number of parts or elements. In each element a function of the space variables is chosen in such a way that continuity of the overall function is maintained along the interelement curves or surfaces. For example, in a region in two dimensions covered by a triangular grid, linear functions in the triangles can give continuity along the internal sides. Thus in this case the overall approximating function is a piecewise linear function, involving parameters that are the function values at the vertices of the triangular grid. In many elliptic problems, the solution minimizes a certain integral involving derivatives of the unknown function. The piecewise approximating function obtained above is substituted into this integral formulation, and the values of the parameters found that minimize the integral. This is the approximate solution of the original elliptic problem. This technique, known as the Ritz method, makes use of a variational principle and involves the finite element method.

Integral equations. An integral equation is one in which the function to be determined appears under an integral sign, an abbreviation for the limit of a sum. Not unnaturally, the usual procedure for finding a numerical solution of an integral equation is to replace the integral by a suitable quadrature formula and to solve the discretized problem by an appropriate standardized procedure. There are two main types of integral equation: (A) the Fredholm equation (named for the 19th–20th-century Swedish mathematician Ivar Fredholm) that relates α, y, and F through an integral from a to b of λK (see 169) in which α, F, K are given functions, and λ, a, b are given constants; and (B) the Volterra equation, similar in structure to the Fredholm equation but with an indefinite integral (see 170). In these equations $y(x)$ is the function to be determined, and K is known as the kernel. If K can be written as $k(x, \xi)y(\xi)$, the integral equation is linear. Also when $\alpha = 0$ the integral equation is of the first kind and when $\alpha = 1$ the integral equation is of the second kind. From a numerical point of view, the distribution between Volterra and Fredholm equations corresponds to the distinction between initial and boundary value problems in ordinary differential equations.

The two types of integral equation

The usual numerical procedures for solving integral equations may be described as follows: in the Fredholm equation, the range $a \leq x \leq b$ is divided up equally into n parts: $b - a = nh$. A quadrature formula like the trapezoidal rule or Simpson's rule (see above *Numerical integration*) then replaces the integral leading to a set of simultaneous algebraic equations for $(n - 1)$ unknown quantities (see 171). These are solved by the methods described in the section on nonlinear equations, unless the system is linear when the methods described in the section on linear algebra are used. With equations of Volterra type, the upper limit of the integral is taken successively as $a + b$, $a + 2b$, $\cdots$, leading to a series of sets of equations, which is solved by a step-by-step procedure. The quantities $y(a + h)$, $y(a + 2h)$, $\cdots$, are calculated in turn. (A.R.Mi.)

BIBLIOGRAPHY. FRANCIS SCHEID, *Schaum's Outline of Theory and Problems of Numerical Analysis* (1968), provides excellent elementary, general coverage of numerical analysis. SAMUEL D. CONTE, *Elementary Numerical Analysis* (1965); CURTIS F. GERALD, *Applied Numerical Analysis* (1970); PETER A. STARK, *Introduction to Numerical Methods* (1970); and FRANCIS B. HILDEBRAND, *Introduction to Numerical Analysis* (1956), a group of four books that are reasonably elementary and well written, discuss the subject in greater detail than Scheid. Two advanced books with fairly wide coverage of numerical analysis are JOHN TODD (ed.), *Survey of Numerical Analysis* (1962); and EUGENE ISAACSON and HERBERT B. KELLER, *Analysis of Numerical Methods* (1966). Works devoted to a specific area of numerical analysis include RICHARD S. VARGA, *Matrix Iterative Analysis* (1962); JOHN R. RICE, *The Approximation of Functions*, vol. 1, *Linear Theory* (1964); LESLIE FOX, *An Introduction to Numerical Linear Algebra* (1964); JAMES H. WILKINSON, *The Algebraic Eigenvalue Problem* (1965); A.M. OSTROWSKI, *Solution of Equations and Systems of Equations*, 2nd ed. (1966); PETER HENRICI, *Discrete Variable Methods in Ordinary Differential Equations* (1962); THOMAS E. HULL, *The Numerical Integration of Ordinary Differential Equations* (1966); and ANDREW R. MITCHELL, *Computational Methods in Partial Differential Equations* (1969). Additional references on numerical analysis and related topics are LOUIS M. MILNE-THOMSON, *The Calculus of Finite Differences* (1960); JOHAN F. STEFFENSEN, *Interpolation*, 2nd ed. (1950); RICHARD E. BELLMAN and KENNETH L. COOKE, *Differential-Difference Equations* (1963); GEORGE BOOLE, *Calculus of Finite Differences* (1860; 4th ed., 1958); LOTHAR COLLATZ, *Funktionalanalysis und numerische Mathematik* (1964; Eng. trans., *Functional Analysis and Numerical Mathematics*, 1966); PETER HENRICI, *Elements of Numerical Analysis* (1964); HEINZ RUTISHAUSER, *Description of ALGOL 60* (1967); JOHN A. JACQUEZ, *A First Course in Computing and Numerical Methods* (1970); and JAMES B. SCARBOROUGH, *Numerical Mathematical Analysis*, 6th ed. (1966).

(N.E.N./A.R.Mi.)

Nurhachi

Nurhachi was the hereditary chieftain of a Juchen tribe in Manchuria who, in the 17th century, became the founder of the Manchu, or Ch'ing, dynasty in China (1644–1911). The Juchen (Jurchen) were a Tungus people who belonged to those border groups at the periphery of the Chinese Empire that normally were under the influence of the Chinese Imperial Court. Juchen leaders had earlier established the Chin dynasty (1122–1234) and controlled North China until they were conquered by the Mongols. In his own buildup of power among the Juchen, Nurhachi followed an Imperial tradition.

Nurhachi's tribe was the so-called Chien-chou Juchen, one of five Juchen tribes of Manchuria. The Chien-chou Juchen lived to the east of the Chinese border in the Long White Mountains north of the Yalu River. Four other Juchen tribes, the Hata, the Hoifa, the Ula, and the Yehe, were located farther north of the central forest and steppe region of Manchuria. These tribes were rivals for power in a frontier relationship of alternate fighting and cooperation, which included intermarriage. In this setting, Nurhachi established his career from very small beginnings. Born in 1559, he was called to leadership in his early 20s, after his father and grandfather had been killed in battle with rivals, supported in this case by China's Ming dynasty, which fostered rivalries among tribes on its borders as a way to make them less dangerous. At first, therefore, Nurhachi had to fight for survival in a situation of decline and disintegration of his own tribe. In 1586 he defeated Nikan Wailan, a rival in his own tribe, who was supported by the Chinese. From this basic success, Nurhachi moved on to destroy, one by one, the challenges of the other Juchen states. To isolate his Juchen opponents from the Chinese, Nurhachi invaded the Chinese-controlled part of Manchuria and moved, thus, on to a more ambitious military and political venture, the attack against the Chinese Empire.

Struggle for leadership of the Juchen

In preparation and while defeating Juchen rivals, Nurhachi established a Manchu state, which at first remained undefined in its political relationship of his Manchurian opponents as well as to the Chinese Empire. But its potential became clearer as the organization advanced. In 1599, under Nurhachi's direction, a Manchu nobleman

and scholar, Erdeni, knowledgeable in both Mongol and Chinese, created a Manchu system of writing that laid the foundation for a Manchu national literature. This was the year also in which the state of Hata, the first of the Juchen rivals, was defeated and incorporated into the Nurhachi state. In 1601 Nurhachi established what was to become the military organization of the Manchus, the banner system. Although basically military, the banners were also units of administration and taxation for the Manchu people and their families. Their commanders and administrators were appointed by Nurhachi, thus injecting an administrative structure into the Juchen tribal system. He assigned the four banners to three of his sons and one nephew, thereby preserving a part of the clan tradition without endangering his own authority. There were originally four banners; four more, established in 1615, were also given to reliable relatives.

This ingenious transformation of a tribal group into a military bureaucracy, which may have been inspired by the model of the military–political structure of Chinese frontier settlements in Manchuria and elsewhere, prepared the way for the Manchu conquest of China.

To provide an economic base for expansion, Nurhachi shrewdly used his position in Manchuria to amass a great fortune from his monopoly of mining in the area and the trade in pearls, fur, and ginseng (a medicinal root) from the area and from Korea. He even developed a new, profitable method of curing ginseng. He also accumulated silver reserves from his tribute missions to Peking, which combined tribute with the trading ventures.

The beginning of the conquest of China

Nurhachi launched his first attack against China in 1618. By that time, he already had defeated two more of the Juchen rivals, the Hoifa and the Ula, and incorporated them in his union, and the final showdown with the most dangerous opponent, the Yehe, and their Chinese supporters was at hand. The Chinese border city of Fu-shun was captured when its commander, Li Yung-fang, defected to the Manchu side. This defection was possible only because the Chinese official saw in the Manchu system the opportunity of serving a Manchu ruler without abandoning his Chinese cultural and political experience. He was only the first of a number of Chinese who surrendered or were captured and entered Manchu service in an administration that adapted many Chinese methods.

Nurhachi's relationship to the Ming emperor at Peking was at first ambiguous. He himself went several times at the head of tribute missions to Peking. In 1601, when the four banners were established, Nurhachi issued a vague claim of having founded a great "Yeh," a family realm or state. In 1616, before the attack against Fu-shun, Nurhachi proclaimed himself as khan ("emperor"), using the Chinese phrase T'ien Ming (Heavenly Mandated). He called his dynasty Chin or sometimes Hou Chin (Later Chin) to indicate a continuation of the Juchen dynasty of the 12th century. Even then, this assertion of Imperial authority did not necessarily imply a challenge to the supreme authority of the Ming, since the Chin dynasty of the 12th century had never ruled the whole of China. The attack against Imperial Chinese forces that followed in 1618 was justified by seven alleged grievances, accusations against the Chinese for their support of his enemies, their responsibility for the killing of Nurhachi's father and grandfather, and other complaints, all within the loyalty relationship between the Ming and his own state.

Nurhachi's ambition, however, clearly went further. He moved his capital into Chinese Manchuria, first to Liao-yang and finally to Shen-yang (Mukden), in 1625, and from there attempted to defeat the Chinese forces guarding the entrance to China proper. In February 1626 he was defeated for the first time by the Chinese at Ning-yuän and died September 30 of wounds.

Nurhachi thus never saw final success of his great political–military venture. On the foundation he established, however, his successors carried out his plans. As a tribal ruler who rose to khanship, Nurhachi had a harem of three wives and many concubines, mostly taken from the families of Juchen chieftains. He had 16 known sons, of whom one, Abahai (died 1643), succeeded him as khan, and another, Prince Dorgon, perhaps one of the most brilliant of the early Manchu leaders, as regent directed the final conquest of China and established the Manchu dynasty in Peking in 1644.

BIBLIOGRAPHY. ARTHUR W. HUMMEL (ed.), *Eminent Chinese of the Ch'ing Period, 1644–1912*, vol. 1, pp. 594–599 (1944), is the most useful biography of Nurhachi, written by FANG CHAO-YING, a well-known specialist in the field of history of the Manchus. It contains a detailed listing of Chinese, English, and Japanese sources on Nurhachi and his period. FRANZ MICHAEL, *The Origin of Manchu Rule in China* (1942, reprinted 1965), analyzes in detail the role of Nurhachi in creating the Manchu state and contains a bibliography of works on Manchu history.

(F.H.M.)

Nürnberg

A city of great importance in the historical development of Europe, as well as a modern industrial and commercial centre, Nürnberg (English Nuremberg) is situated in the forested uplands of the southern part of the Federal Republic of Germany. It is the foremost city in Franconia, the northern section of the *Land* (state) of Bavaria, and is exceeded in regional importance only by Munich (for related information see BAYERN).

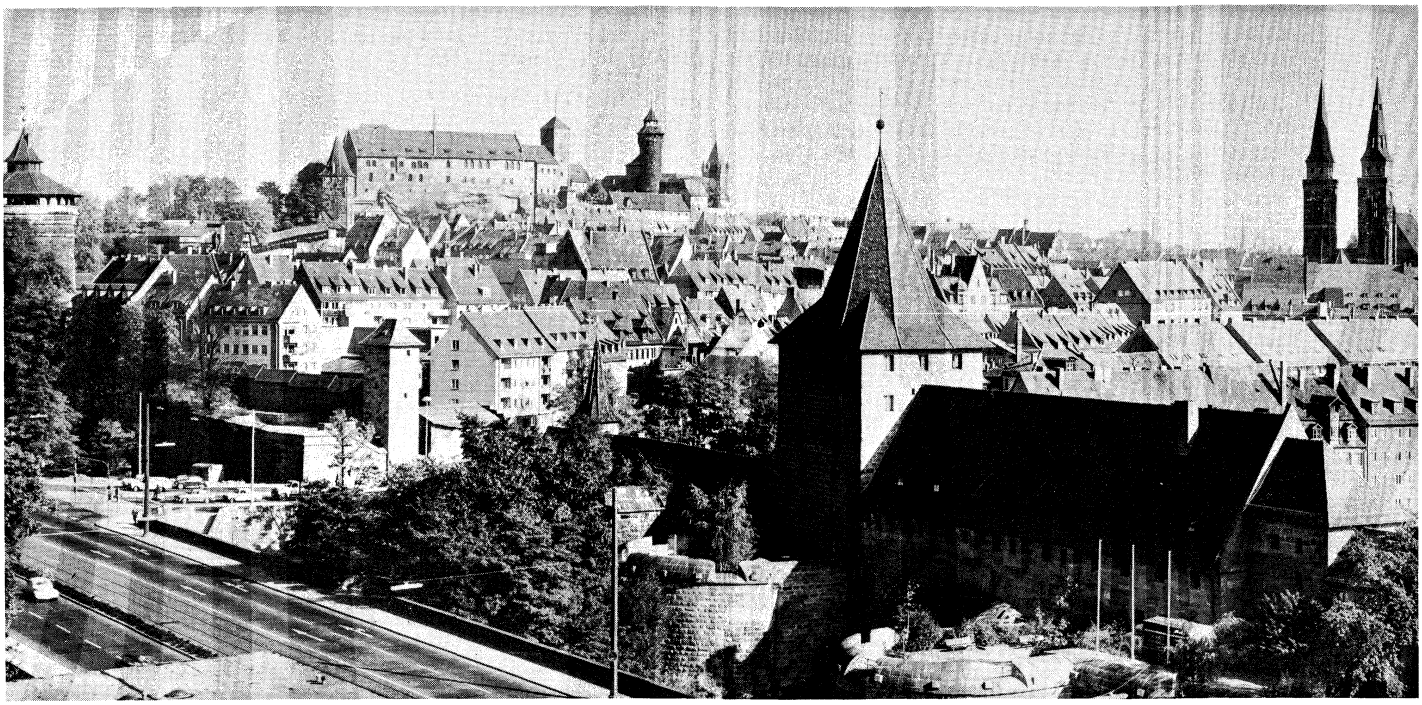
History. First mentioned in official records as Noremberg, in 1050, the city had its origin in a castle built about ten years earlier by the German emperor Henry III, duke of Bavaria. From a sandstone rock, rising 165 feet (50 metres) above the surrounding plain, the castle looked down on the Pegnitz River, commanding its strategic exit from the Jura Uplands, and thus dominating one of the ancient trade routes of Europe. At the foot of the rock, on the north bank of the Pegnitz, the vassals of the castle, together with artisans and merchants, soon established an embryonic settlement. Meanwhile, on the steeply rising southern bank of the river, another royal seat became the nucleus of a second settlement during the course of the 12th century, and, in 1219, the city was granted its first charter. The privileges this bestowed on the city were a direct result of the dearth of natural resources in the area: the soil was infertile, there were no vineyards, and the river was not navigable. In 1225 construction started on the city's first church, the St. Sebald, subsequently renowned as a pilgrims' shrine. A city council, consisting predominantly of members of wealthy patrician families, is first mentioned in 1256. The city soon gained full independence, becoming a free imperial city, with extensive land holdings and joining the Rheinische Städtebund (League of Rhineland Cities). A city wall was erected, and Nürnberg adopted its own seal.

The castle

A second church, the St. Lorenz, was built on the south side of the river in 1260, amid the growing community of Lorenzer *Stadt*. Nürnberg's character was now changing; no longer solely a fortified settlement, it was developing into a city of craftsmen and patricians, and manufacture and commerce became the foremost sources of income. In 1332 the city was granted the right to buy and sell free of tax in 72 different localities, and in the course of the 14th century Nürnberg concluded trade agreements with almost all the important cities and most of the larger territories of central Europe. From 1356 onward, the head of the Holy Roman Empire held his first imperial diet at Nürnberg, a significant indication of the city's growing influence; in 1424 the imperial crown jewels were transferred to the city, remaining in the Hospital Church of the Holy Ghost for over 350 years.

Early commercial expansion

In 1471 the painter Albrecht Dürer was born in Nürnberg. Thanks to Dürer and his famous contemporaries—the painter Michael Wohlgemuth (his teacher), the wood sculptor Veit Stoss, the brass founder Peter Vischer, the stonecutter and sculptor Adam Kraft, as well as the cobbler-poet Hans Sachs—the arts flourished in Nürnberg as never before or since. In 1525 the tenets of the Reformation were adopted by the city, and in 1526 the scholar and Protestant leader Philipp Melanchthon founded one of Germany's first *Gymnasiums*, an institution that continues to bear his name. Together with the humanist Willibald Pirkheimer, the astronomer Regiomontanus, and the cosmographer Martin Behaim, the designer of



Nürnberg, with a portion of the medieval wall surrounding the old city in the foreground and the imperial castle in upper left. St. Sebalduskirche is in the upper right.
By courtesy of the German Information Center

the first globe, Melanchthon laid the foundation for Nürnberg's reputation as a centre of learning in the developing Western world.

When the architect Jakob Wolff the Younger rebuilt the city hall in the style of a Renaissance palace in 1616, Nürnberg was at the pinnacle of its economic and cultural development; yet by 1806 it had lost its status as a free imperial city and, much indebted, became part of the kingdom of Bavaria. The shift of world trade routes from the land to the sea, following the discovery of America and of the seaway to India, and the devastations of the Thirty Years' War were the initial causes of this decline. Not until the dawn of the industrial age, when the first German railway (linking Nürnberg and Fürth) was opened, on December 7, 1835, did the city begin to flourish again, as a modern industrial centre.

The contemporary city. *The site.* Located at a height of between 948 and 1,230 feet (289 and 375 metres) above sea level, and lying to the west of the Jura Uplands, modern Nürnberg is situated on both banks of the Pegnitz River, at the point where it flows into a broad, level basin. The Pegnitz Valley is filled with geologically recent sand deposits, but there are occasional outcrops of older rock, the most noteworthy of which is the gray and reddish-brown sandstone eminence on which the Nürnberg castle is built.

The city owes its economic development to its location at the intersection of two natural trade routes running across Europe from north to south and from east to west. Sheltered by the hills of the Steigerwald to the west, Nürnberg has a relatively meagre rainfall, and its climate is almost continental.

Topography. The inner city, divided into two parts by the Pegnitz, is to this day encircled by the three-mile-long wall completed in 1452, and 60 of the original 128 towers are still intact. Even the older, inner, line of fortifications, dating from 1140 and 1320, can still be traced. Unfortunately, only a few historic buildings survived the massive bomb damage wrought toward the end of World War II, although some have been restored. The most important are the Gothic churches of St. Sebald and St. Lorenz, and, adjoining the market place, the Frauenkirche (Church of Our Lady). The Heilig Geist Spital (Hospital Church of the Holy Ghost), rising above the Pegnitz, is now an old people's home. In addition, there are the Mauthalle (customs house) on the Königstrasse, the Weinstadel (wine storage house), the Renaissance city hall, the Schöne Brunnen (a fountain), the Fembohaus (today a museum of the old city), and, finally, towering above them all, the imperial castle and its stables and granary, the latter now converted into a youth hostel.

Circling the city walls is the Ringstrasse, affording access to the main railroad station, the general post office, the Opera House, and the main commercial buildings. Extensions of the city, built in the 19th and early 20th centuries, lie beyond. These consist of blocks of workers' flats and large industrial plants, merging into modern garden suburbs and dormitory towns. Among the latter, the Langwasser community, to the south, has been planned for the needs of a population of 60,000. The uninterrupted expanse of woods framing Nürnberg to the north, east, and south, including the modern pine plantations of the Sebaldus and Lorenz Woods, serve as a reminder that the city was once a settlement in a clearing in the midst of the former imperial forest. The city covers a total area of about 52 square miles, of which only about 40 percent is built, with over 50 percent cultivated or forested.

Outlying regions

In the west, Nürnberg borders on the city of Fürth, which had 94,000 inhabitants by the early 1970s. The two communities have, in effect, become one twin city, with an uninterrupted built-up area linked by the axis of the Fürther Strasse, which runs along the former Nürnberg–Fürth railroad.

In the north, the so-called garlic land, an area under intensive vegetable cultivation, separates Nürnberg from the city of Erlangen, which had 86,000 inhabitants by the early 1970s. Close ties are also maintained between these two cities, Erlangen being important as a university town and a centre of the electrical industry. To the east and south, Nürnberg spreads into the industrial regions of the Pegnitz and Rednitz valleys.

Population. Numbering 40,000 in 1600, the city's population had shrunk to 25,000 by the beginning of the 19th century, reflecting Nürnberg's historic decline. The industrial age led to a rapid increase, and, by 1890, 143,000 people were living in the city. By 1914 the population had risen to 280,000, reaching 423,000 by 1939, but falling again to 195,000 by 1945, primarily as a result of war damage. Of 125,000 dwellings, about 60,000 were destroyed completely and about 51,000 were severely damaged. Eventually the city was extensively reconstructed, and by 1970 the population had climbed to 473,600. Al-

War damage

though 50 percent of the inhabitants are employed in industry, this is quite insufficient to meet the demand for labour, and consequently 80,000 people commute daily from the surrounding areas.

Economic life. For centuries, the metalworking crafts have been Nürnberg's chief industry. First there were the armorers, cannon founders, gunsmiths and sword smiths, and the pioneer compass and clock makers of the Middle Ages. They were followed by makers of gold foil, the manufacturers of wire, needles, and pins, the makers of metal toys, and, in the contemporary city, the electrical industry and the machine-building industries, together with bicycle and office-machine factories. About 45 percent of Nürnberg's labour force is employed in manufacturing, with about 40 percent of these making fine mechanical and optical goods and electrical apparatus (the latter of particular importance, as 10 percent of West Germany's total capacity is located there). Other major manufacturing employers include the motor vehicle, printing, chemical, wood and paper, and textile industries. The most important nonmanufacturing sectors are wholesale and retail trade, with 16 percent of the labour force; services, 12.5 percent; and transportation, 7 percent. When the International Toy Fair was established after World War II, Nürnberg became the world centre for this industry. To accommodate this and other exhibitions, a large area south of Nürnberg was converted into a centre for exhibits, sports, and recreation.

International exhibitions

Transportation. Nürnberg owes much of its development to its excellent situation near the important European trade routes. In the Middle Ages no fewer than 12 highways met in the city, making it a traffic centre for the whole of Europe. Today, it is a focal point not only for the old European highways, which now lead from The Netherlands via Vienna to the Balkans, from Italy to Scandinavia, from France to Bohemia, and from Switzerland to Poland, but also for modern Federal German and European highways; it is also connected to the Munich–Berlin and Frankfurt–Cologne autobahns. Already situated on the old Ludwigs-Danube-Main Canal, Nürnberg has built a modern harbour linked in the early 1970s with the new Europa-Kanal, which joins the Rhine, Main, and Danube rivers. Nürnberg's airport, on the north of the city, provides connections, particularly to the international airport in Frankfurt am Main.

Administration and services. Nürnberg is the regional headquarters for the postal service and for the Federal railroad and is the seat of a municipal court, a circuit court, and an appellate court. Although it lacks regional importance in the overall field of local government, it is the seat of the Federal Employment Institution.

There are a number of institutions of higher learning in and around the city, including the Faculty of Economic and Social Sciences of the University of Erlangen-Nürnberg, and the Ohm Polytechnic Institute for Applied Technology (founded in 1823 and named after the famous physicist Georg Simon Ohm, discoverer of Ohm's law, who taught from 1833–49). The Academy of Arts, founded in 1662, is the oldest in Germany. Other institutions, also the earliest of their kind in the nation, are the Pegnesische Blumenorden, a literary society founded in 1644, and the city's public library, which is over 600 years old. In addition to the museum of the old city located in the so-called Fembohaus—the only surviving patrician house, with a splendid Renaissance gable—the city is the home of the unique Germanisches Nationalmuseum, located in a former Carthusian monastery, which was chartered in 1852 to exhibit the history of German culture from its earliest time to the present. Its magnificent collections include a complete collection of Dürer's printed graphic works.

Cultural heritage

Ample provision has been made for cultural life in Nürnberg: there are an opera house and two theatres, as well as the modern Meistersingerhalle, which is available for the many congresses as well as for the concerts that take place in the city. Performances of annual organ concerts are usually held in the medieval churches. The design of the zoological gardens at the Schmausenbruck is of exceptional beauty.

BIBLIOGRAPHY. GERHARD PFEIFFER, *Nürnberg, Geschichte einer europäischen Stadt* (1971) and *Geschichte Nürnbergs in Bilddokumenten* (1970–71), provides two histories, the first textual, the second pictorial and with illustration notes in English. See also WILHELM SCHWEMMER, *Stadtführer, amtlicher Führer von Nürnberg*, 6th ed. (1970), the official city guide by an art specialist; and *Amt für Stadtforschung und Statistik Nürnberg: Statistische Nachrichten der Stadt Nürnberg*, a yearly statistical table. Other works include H.M. VON AUFSESS, *Nürnberg* (1967), a lavish picture book; and GERALD STRAUSS, *Nuremberg in the Sixteenth Century* (1966), illustrated and with maps; and WILHELM SCHWEMMER, *Nürnberg: A Guide Through the Old Town* (1960); ERICH MULZER, "Der Wiederaufbau der Altstadt von Nürnberg 1945–1970," *Mitt. Frank. Georg. Ges.*, Bd. 19 (1972).

(E.Wt.)

Nursing

The functions served by nursing are summarized in a statement from the Code of Ethics of the International Council of Nurses:

> Nurses minister to the sick, assume responsibility for creating a physical, social and spiritual environment which will be conducive to recovery, and stress the prevention of illness and promotion of health by teaching and example. They render health service to the individual, the family, and the community, and coordinate their services with members of other health professions. Service to mankind is the primary function of nurses and the reason for the existence of the nursing profession. Need for nursing service is universal. Professional nursing service is therefore unrestricted by considerations of nationality, race, color, political or social status.

Men and women engaged in nursing throughout the world comprise the largest single group of health workers, and the availability of effective nursing care in a country is a principal measure of its potential advancement in health. Modern nursing is essential not only to hasten the recovery of sick people but also to provide people with knowledge that will improve their health and productivity. Since health possesses humanitarian and economic values, nursing is an instrument of social progress. An analysis of nursing includes its history, the number and kinds of nursing personnel, the education that prepares nurses for their contribution to health, organizations among nursing personnel, and the relations of nursing with other professions.

HISTORY

Sensitivity to the suffering of others and desire to comfort have been found among human beings from earliest times. Attitudes toward childbearing, toward the aged, and toward death and ideas about the causes of sickness have influenced nursing and healing practices throughout time. Evil spirits were to be cast out; offended gods, placated. Ceremonials, fetishes, concoctions of extraordinary ingredients were used.

Early developments. Reports of early civilizations—Egyptian, Indian, Chinese, Aztec, Greek—refer to bleeding, cautery, applications of poultices and bandages, bonesetting, and trephining, or the removal of sections of bone from the skull. Hospices and hospitals were established to provide refuge for the weak and hospitality for sick strangers. The beginnings of medicine, surgery, psychiatry, midwifery, sanitation, and nursing can be traced in these records. Codes of ethics for those who give health care were formulated. The Oath of Hippocrates persists to this day (its text may be found in the article MEDICINE, HISTORY OF).

When Christianity came, with its teachings of love and brotherhood, healing and caring for the sick gained new impetus. Among the early converts to Christianity were highborn Roman matrons. Phoebe, a friend of St. Paul, "succored many"; Marcella made her palace into a monastery for women; Fabiola turned her home into a hospital. These women and others like them exemplified new intellectual freedom for women and their participation in religious and social action at a high level.

Contribution of Christianity

For several centuries thereafter the spread of health work moved slowly. Monastic orders grew, and from them some health services radiated; but large areas of the

world were untouched, and the monastic surge to undertake health work seemed to diminish. Later, military and chivalric orders were founded, supporting the Crusades and combining the making of war with charitable and hospital work. Some established companion orders for women. The Knights Hospitallers of St. John established two hospitals in Jerusalem and branches of the order in other countries, one in England around 1100. This order included among its functions the care of the insane; its influence on services during disasters and emergencies extends to modern times.

The monastic and chivalric orders established in various forms among health workers the beginnings of hierarchical organization, amateur participation that provided the roots of volunteerism, and a sense of calling or vocation.

Nursing can be seen to rise from two wellsprings, one scientific, the other religious and social.

Acceleration of scientific advancement in health began with the 16th century. A French surgeon, Ambroise Paré, a Flemish anatomist and physician, Andreas Vesalius, a Spanish theologian and physician, Michael Servetus, and an English physician, William Harvey, were leading contributors to progress. Servetus was burned at the stake, and Vesalius barely escaped a similar fate. The 19th century brought the "germ theory" of disease and ways of treating and preventing infectious diseases, then the largest cause of death. Anesthesia was discovered. Wars were times of advance.

It is often said that research produced more medical and health knowledge in the decades after World War II than in all previous centuries combined. Atomic and space sciences played a part. In the United States, governmental financial support of medical research grew to large proportions. Federal support of nursing research and research training in the United States from 1964 to 1971 approximated $20,000,000. The constantly growing mass of new knowledge to be applied in health services by health workers challenges the educational system for physicians, nurses, and others and strains the system of distribution of services to an awakened public.

Paralleling these scientific advances were advances made through the years in social and religious action. The work of St. Vincent de Paul and, later, of John Howard typify this kind of action. After making thorough studies of conditions, they recommended marked changes in hospitals and prisons. A new dimension, planning based on facts, thus was added. St. Vincent, with Saint Louise de Marillac, founded the Daughters of Charity (1633). He believed that the sisters should have general education, as well as training for their work. In the following century Howard revealed deplorable conditions in nursing—filth, stealing, and ill treatment of patients. Both men influenced reforms and emphasized citizens' responsibility for health and welfare of people everywhere.

The work of Florence Nightingale

19th-century reform. In the 19th century the movement for reform in nursing was led by Florence Nightingale, a woman of intellectual and moral power. Family contacts with humanitarian leaders and an education that included training in science, mathematics, and political economy were her preparation. She analyzed nursing as it was practiced in several countries, formulated her ideas, and wrote extensively.

In 1854 Florence Nightingale was asked by the British secretary of war to go to the Crimea, where absence of sewers and of laundering facilities, lack of supplies, poor food, disorganized medical service, and absence of nursing led to a death rate of more than 50 percent among wounded soldiers. Her work and that of the nurses whom she recruited brought sufficient improvement to lower the death rate to 2.2 percent. Thus, British military medicine was reorganized and the role of good nursing established. A gift of £44,000 was raised by popular subscription, and Florence Nightingale used it to establish a school of nursing at St. Thomas's Hospital.

Florence Nightingale believed nursing to be suitable as an independent career for capable, trained women, that nursing services should be administered by nurses with special preparation, and that relationships between physicians and nurses should be professional. She maintained that schools of nursing should be independent of hospitals and should be administered by nurses with physicians as paid teachers as needed, with proximity to hospitals for student experience. She believed that there was a substantial body of knowledge and skills to be learned in nursing. Nurses were to be prepared for hospital nursing and care of the sick at home, and they were to teach good health practices to patients and families.

Within 25 years after the Nightingale school was established, the old system of nursing in England by low-paid, untaught women or pauper inmates had almost disappeared, and patients began to receive effective nursing care to accompany the constantly improving medical care. This kind of revolution has taken place in most developed countries. In newer countries, efforts are made to bypass some of these steps in evolution and to bring nursing along with other health professions to the point of meeting modern challenges more quickly.

20th-century developments. In the 20th century the number of nurses of all kinds has increased rapidly, as has the number of kinds of nursing personnel. Recently, some nurses have specialized in clinical areas.

The development of physicians' assistants sometimes overlaps with the development of nurse specialists, though efforts are made to differentiate the two categories. Pediatric-nurse practitioners, for example, undertake functions previously performed only by physicians—taking histories, performing physical examinations on children, and running clinics for mothers and babies. Experimentation with shifts in functions from medicine to nursing is growing rapidly.

Planning aimed at bringing nursing resources into balance with needs for nursing personnel to implement health programs is a recent movement. Almost every state in the United States is involved in continuous planning for this purpose. Several other countries are similarly involved. The World Health Organization (WHO) has developed a guide for such planning published in English, French, Russian, and Spanish; this guide helps countries to set goals and avoid disadvantages of haphazard growth. Coordinated planning among the several health professions and occupations is a desirable new trend.

Historians may characterize the 1970s as a time of planning, research, expansion of functions of nursing, improvements in delivery of health and nursing services, development of specialities, changes in utilization of existing types of nurses and development of new types, and reorganization of nursing education.

NURSING PRACTICE, EDUCATION, AND ORGANIZATION

Methods of classifying nurses

Kinds of nursing. Nursing personnel and the kinds of nursing in which they engage may be classified in several ways: by legal designation—registered nurses, licensed practical nurses, and, in addition, unlicensed nurses' aides; by the kind of educational preparation—vocational, technical, and professional, including graduate education—that they have received; by the kind of work they do—institutional, community, educational, research, or journalistic; by the level of responsibility that they assume—that of staff nurse, teacher, supervisor, administrator, or consultant; and by the place of employment—hospital, physician's office, public health agency, school (school health nurses), industry, a school of nursing, including undergraduate and graduate programs. Nurses may also be classified as generalists or specialists. Most descriptions of nursing, including the following, mix these classifications.

Hospital nursing. Hospital nursing occupies approximately two-thirds of the total nursing force in most countries. Nurses giving direct care to patients in hospitals of the United States include registered nurses, some of whom may be specialists; licensed practical nurses; and nurses' aides. Aides and sometimes practical nurses, if there are any, are called auxiliaries in many other countries. In large, highly organized hospitals, administrators of nursing services need special preparation and occupy demanding executive positions.

Certain areas of hospitals are highly specialized—surgical suites (operating rooms), recovery and intensive-care

units, and coronary-care units. In these, nurses giving direct patient care and their supervisors must have the additional preparation afforded by graduate study or by continuing education or staff-development (in-service-education) programs.

Almost all hospitals except the smallest have established staff-development programs; teaching and directing such programs for nursing personnel are relatively new roles for nurses, and the number of such nurses grows rapidly. These programs are usually responsible for the teaching of nurses' aides, who are often hired with no pre-employment preparation, as well as for general and special staff development in nursing throughout the hospital.

The rapid development of hospitals in response to advances in science and new community responsibilities requires new kinds of nursing. Some hospitals are using nurse-midwives in maternity departments. This practice is new in the United States. Nurse specialists are utilized in many other areas as well—medical, surgical, obstetric, pediatric, psychiatric, and rehabilitative; these nurses often cross over to other fields as consultants. For example, public-health nurses are employed to ease patients' adjustment to returning home and to assure planned continuity of health care. A few nurses occupy ombudsman, or patient-advocate, positions. Some perform studies and researches aimed at solving problems and at devising improved ways of caring for patients.

New health problems bring new responsibilities to nurses in hospitals. Drug abuse, alcoholism, and family planning, for example, are areas that call for new knowledge on the part of nurses, as well as for decisions about the organization of the new services in hospitals.

Outpatient and home care

Outpatient and home care are sometimes organized by hospitals to reduce costly hospitalization for patients. Some clinics in outpatient departments can be operated almost exclusively by nurses—clinics for hypertension (high blood pressure), chronic heart disease, and for poststroke and maternity patients are examples. Satellite clinics bring services close to patients' homes. All this means expanded learning of assessment methods by nurses—simple physical examination and history taking, for example, and new understanding of community problems and organization of nursing services.

Nursing homes provide services for a large number of patients throughout the United States. In such establishments nurses play a central role. A high proportion of the staff is composed of nurses' aides, who need to be taught and supervised by registered nurses in such a way that rehabilitative care rather than mere custodial care is given to patients. In the 1970s there were many excellent U.S. nursing homes, but many were in need of reform, a process in which nurses are expected to participate.

Mental hospitals accommodate almost half the hospitalized patients in the United States. Efforts to return many of these patients to their communities are meeting with some success. In some of these hospitals, as well as in the new community mental-health centres, nurses with graduate degrees share in the treatment of patients and conduct group therapy, another role that has recently developed.

Public-health nursing. Public-health nursing, more recently called community-health nursing, is given by two kinds of agencies in the United States, governmental and private. School nurses are included in reports of this field in the United States. Approximately 50,000 registered nurses were employed in the 1970s by almost 10,000 agencies, 6,000 local school boards, 600 visiting-nurse services, and more than 3,000 state and local health departments. These agencies also employed about 4,000 licensed practical nurses. Of the registered nurses employed by these agencies, 41 percent possessed a bachelor's or a higher degree.

Visiting-nurse services care for the sick at home and carry on individual, family, and community programs of prevention and health teaching. These services are usually supported by civic and other forms of private philanthropy and by fees from patients who can afford to pay. Some of the visiting-nurse services have contracts with local government health and welfare agencies and with industrial health units to provide services. Many have deep, traditional roots in their communities, which provide volunteer board members.

In most countries public-health nursing is carried on primarily by government agencies. The numbers reached by government services are highest in such countries as Great Britain, Canada, Australia, New Zealand, the Scandinavian countries, Finland, Yugoslavia, and the United States. In new countries, community nursing is important for the development of health programs. In rural and isolated communities, a public-health nurse may be the only person available for any kind of health service.

Nurses as educators. Nursing education is a field that combines nursing with the teaching of students of nursing and, for some nurses, with the administration of educational programs. A higher proportion of the teachers in each of the kinds of nursing-education programs teach in clinical situations, in which students learn to care for patients and families in hospitals, at home, and in other situations. Teachers must have knowledge of how learning takes place, be expert in their own field, and know how that field relates to the total scheme of health promotion. The full-time teachers of nursing in the United States are distributed among the various kinds of educational programs approximately as follows:

Teaching programs

Programs leading to bachelor's and higher degrees	4,500
Associate-degree programs	1,800
Diploma programs (hospital schools)	9,000
Licensed-practical-nurse programs	3,700
Total	19,000

In addition, there are about 2,500 part-time teachers with approximately the same distribution.

Teaching in staff-development programs of hospitals and other health agencies and in continuing-education programs offered usually by universities or associations is a growing field of employment.

Private, office, and industrial nursing. Private, office, and industrial (recently called occupational-health) nursing as fields of employment account for about 15 percent of the registered nurses in the United States. Licensed practical nurses are also employed in these fields.

Military nursing. Military nursing provides an essential part of the health care given to men and women of the armed services in most countries. Trained nurses were part of military forces of countries involved in World War I. In the United States many of these nurses were drawn from a reserve maintained by the Red Cross. By the close of World War II much progress had been made in assigning the nurses rank and responsibilities commensurate with their training and abilities.

Male nurses were not commissioned as officers in the nurse corps until after World War II. In the 1960s the flag grade (that of general) was accorded to leading nurses in the military nurse corps. Medical corpsmen save lives on battlefields and hospitals and are highly skilled. Several plans have been made to hold them in health work after they leave military service. The number of registered nurses in the early 1970s commissioned and on active duty in the United States Army, Navy, and Air Force Nurse corps was approximately 12,000.

Government nursing. This field includes military nursing and public-health nursing as carried on by federal, state, and local health departments, functions already touched upon.

The Veterans Administration operates the largest single hospital system in the United States, employing approximately 17,000 nurses and a large number of nursing assistants. And the Department of Health, Education, and Welfare provides health care, including both hospital and public-health nursing services, to approximately 400,000 American Indians on reservations and to Alaska natives.

Federal programs to further nursing

Various branches of the department, primarily the division of nursing, administer programs aimed to increase and improve nursing services and education throughout the United States. The programs take the following main forms: consultation service; conduct of studies and research and publication; collection of data on nurse supply and health needs as a basis for planning; mobilization of numerous advisory committees on plans and evaluations;

granting of federal funds for scholarships, operation of schools, construction of facilities, and for research and demonstrations. The sum of the funds granted to institutions and students and for projects (excluding research) in the period 1964–72 was $325,944,000.

The office of the Department of Health, Education, and Welfare carried out a similar program for practical-nurse education and can be given credit for a large part of the rapid expansion of schools and increase in students. The number of nurses employed in the operation of these programs of federal assistance to schools and students is extremely small, but the impact on the development of nursing is great.

Education. Modern education of nurses began with the Nightingale era. It has a dual purpose: (1) to provide personnel to meet the needs of people for nursing services of all kinds and (2) to supply training that will enable young men and women to follow one of the varied careers in this health service and to progress from one type of career to another.

The basic educational program for nurses is scientific and humanistic in content. Many programs lead to the bachelor's degree. Nursing specialists, teachers, and other leaders in the field may need advanced training at the master's or doctoral level. All educational programs include experience with patients in hospitals, homes, or other settings. Table 1 shows the diversity and relative capacities of programs in the United States. In addition, 850,000 nurses' aides of various kinds employed in the 1970s were prepared through training programs operated usually by the employing institution.

Table 1: Nursing Programs in the United States

	locale	number of programs	approximate number students graduating
Practical nurse	vocational high schools, community colleges, and hospitals	1,350	37,000
Diploma	hospitals	641	23,000
Associate degree	community colleges	450	27,000
Bachelor's degree	colleges and universities	270	20,000
Master's degree (for nurses with bachelor's degree)	universities	73	2,000
Doctoral degree in nursing*	universities	8	30

*A number of nurses study in related doctoral programs.

For comparison, in 1925 in the United States, there were approximately 2,500 diploma programs, no associate-degree programs, only a handful of master's degree programs, and a very small number of bachelor's degree and practical-nurse programs.

In almost all other countries in which nursing education exists there are at least two kinds of programs—those leading to diplomas and those that train auxiliaries, though a large portion of auxiliaries in some countries are untrained. A growing number of countries have one or more bachelor's degree programs; some have several.

Countries with bachelor's degree programs

Among the latter are Canada, Colombia, Ecuador, Panama, Peru, Great Britain, Lebanon, Egypt, India, the Philippines, Taiwan, and Thailand. Bachelor's degree programs are evolving in other countries, and master's degree programs have been established in Canada, Colombia, India, and elsewhere.

Postbasic programs for nurses with diplomas have been established in many countries. But not all of these are within universities. Some programs offer courses in general education, as well as nursing courses, and some, in universities, may become programs leading to a bachelor's degree. The purposes of the postbasic programs vary and include the preparation of teachers, supervisors, or administrators and of nurse specialists in various fields including midwifery, public health, and teaching of auxiliaries. Some augment the education received in other programs. Enrollment is generally small in relation to the need for their graduates.

A number of centres were established for the presentation of centralized postbasic programs for nurses from several countries. Among these are centres for African nurses in Senegal, at the University of Ghana and at the University of Ibadan, in Nigeria. In combination these provide opportunity for both French- and English-speaking nurses. Nurses from many countries study in postbasic programs in Wellington, New Zealand. The University of Alexandria serves in this role, as well as in other roles.

The effort to establish new bachelor's degree programs and postbasic programs in many countries and to improve existing diploma and midwifery courses and courses for auxiliaries was stimulated greatly by the World Health Organization and its six regional offices.

The progress of nursing education in any country is affected by the developments in general education. In the United States and some other countries, high school graduation or its equivalent has for many years been a requirement for admission to schools preparing registered nurses. In the United States this is also a requirement for admission to practical-nurse programs. Of the 136 schools preparing registered nurses in Latin America, 58 percent call for nine to ten years of previous schooling, the remainder having requirements roughly equivalent to those prevailing in the United States. In some other countries some schools may require only six years of previous education. This situation is expected to improve along with the programs for general education.

The rapid development of community colleges in the United States brought with it an increase in the number of associate-degree (two-year) programs in nursing, and the placement of professional education in colleges and universities helped nursing education. The first (*i.e.*, lowest) professional degree in nursing in the United States in the early 1970s was the bachelor's; in nursing the master's degree program was usually considered the education for specialization, whereas in many other professions the first degree is the master's.

The movement in general education to attract and retain youth with disadvantaged educational backgrounds is paralleled in nursing education.

Qualifications of teachers affect the quality of education. In the United States about 45 percent of teachers in all schools of nursing possessed the master's degree. The percentage of teachers in bachelor's and higher-degree programs with the master's was above 80, while about 10 percent of the administrators of bachelor's or higher-degree programs in nursing possessed doctoral degrees.

Texts and reference books in nursing were not available in the language of many countries in the early 1970s. A few texts in English were translated into Spanish and Japanese, but translations from other languages are not always appropriate to the culture and health problems of a particular country. Hence, international organizations, including the World Health Organization, were working on these problems and on the needed expansion of all libraries. The literature of nursing was growing rapidly, but in the early 1970s the nursing literature of the United States still represented a large portion of the total.

Licensing and registration. After the number of nurses has become substantial and the essential nature of nursing has become established in a country, the need to regulate the practice of nursing under law becomes evident. These laws are aimed at the protection of the public.

Laws define nursing and establish titles under which nurses practice. Most laws establish boards empowered to give examinations and maintain registers of qualified nurses. Boards also are usually empowered to maintain lists of schools the graduates of which are eligible to take examinations. Further, the laws that regulate nursing usually provide ways to recognize licenses acquired by nurses from other states and countries.

The first state to pass a nurse practice act in the United States was North Carolina, which did so in 1903, and within the next two decades all states had adopted laws regulating nursing. The title established by these laws and their revisions is that of registered nurse; graduates of three kinds of educational programs—bachelor's degree

and associate-degree programs and diploma programs—are eligible to seek this title. All states also have laws giving another title—licensed practical nurse, or, in two states, licensed vocational nurse. Educational programs for this title are typically one year in length.

Establishment of nurses' register in England and Wales

The Nurses Act of 1919 established the General Nursing Council for England and Wales to maintain a register of nurses; similar acts were passed in 1919 for Scotland and in 1922 for Northern Ireland. The advent of the National Health Service (1946) called for enlargement of the council and its functions, and in 1949 an act was passed that consolidated all previous acts and established regional committees to work with schools of nursing for their improvement and for ways to meet new challenges. The title used in England for the grade called registered nurse in the United States is that of state-registered nurse, with a few variations, and the required education is three years or more. Legislation in 1960 changed the title of a less qualified grade from state-enrolled assistant nurse to state-enrolled nurse. Educational preparation for this title is 12 to 18 months. These steps illustrate the progress of licensure and registration in a country and its responsiveness to change.

National and state nursing associations are prime movers in urging legislation for licensure and registration and in securing amendments. The American Nurses' Association gives leadership to state associations and state boards regarding changes in definitions, the desirable provisions of the law, and the legislative process.

The International Council of Nurses exerts leadership among national nursing associations in their efforts to pass and amend laws, and a guide was published recently. The staff of the World Health Organization also has assisted new countries in this important process.

Organizations. Near the beginning of the 20th century, nurses began to organize national associations. The purposes of such organizations usually include: promotion of nursing care of high quality, promotion of desirable legislation in nursing and health, formulation of nursing and educational standards, professional development and welfare of nurses, and representation of nursing with other professional associations and government agencies. Many nursing associations publish professional journals. In large countries the national associations have state or provincial and district constituent associations.

An example of a thriving and vigorous national organization is the American Nurses' Association, founded in 1896, which continues to operate aggressive programs in line with the purposes mentioned above. It affects legislation related to health and nursing in the Congress through its educational and lobbying activities and exerts leadership in revision of state laws governing registration and the practice of nursing. A strong program for economic security has elevated nurses' salaries and the conditions of patient care, partly through collective bargaining. The official statements of the association set the course of action. Its official publication is the *American Journal of Nursing*.

The International Council of Nurses

In 1899 the International Council of Nurses was founded as a federation of autonomous national nursing associations. By 1970 there were member associations from 74 countries, and the council was working in another 50 countries, many of them new countries, to assist in the development of national associations and ultimate membership in the council and to encourage national legislation on registration and nursing practice. An international congress is held every four years. The council's journal, *International Nursing Review*, makes a substantial contribution to worldwide literature of nursing. The Florence Nightingale International Foundation, which is related to the council, has conducted an international conference on research in nursing, as well as an international exchange of nurse scholars.

Several national colleges of nurse-midwifery and an international college work for improvement of maternal and infant health. Nurses participate in international and national associations for public health, mental health, industrial health, and school health and in such areas as heart-disease, cancer, and respiratory-disease control. The Nursing Section of the American Public Health Association is its second largest section. Practical nurses and auxiliaries are usually organized, if at all, in separate organizations. There are two such organizations in the United States, the National Association for Practical Nurse Education and Service and the National Federation of Licensed Practical Nurses. Various groups from special organizations in several countries—deans of schools of nursing and operation-room nurses are examples. Nursing councils are parts of regional commissions on higher education in the United States.

A unique organization—the National League for Nursing—combines nurses, related professionals, and other public-spirited citizens in the United States for meeting the nursing needs of people throughout the country. This organization also carries on a program of national accreditation of all the various kinds of educational programs. Its journal is *Nursing Outlook*. With the American Nurses' Association it sponsors *Nursing Research*.

The diversity of participation of nurses demonstrates the interrelationships with nurses needed for achievement in health, social, and educational missions.

Studies and surveys of nursing. Studies of nursing have been made in many countries. In the words of the Expert Committee on Nursing (1966) of the World Health Organization,

> Minor modifications of existing nursing systems will be inadequate to meet new situations and demands in a rapidly changing society. . . . Nursing must break with some of its traditions as well as alter existing stereotypes.

Studies often collect the facts and recommend and push needed changes.

In the United States in 1923, *Nursing and Nursing Education in the United States*, a classic, drew attention to the inadequacies of the hospital-nursing-school system to prepare nurses for community-health and new-hospital demands. The Grading Committee Reports (1928 and later) were comprehensive and showed the magnitude of the needed reform. Fiscal studies (1945 and following) related cost of education to quality and stimulated decisions on who should pay costs. *Nursing Schools at the Mid-Century* served as a basis for a new, unified system of accreditation. A group of studies by Esther Lucile Brown resulted in a series of publications (1948–71) dealing with professionalism, humanization of nursing, and new roles for nurses. The surgeon general's consultant group on nursing (1963) studied and recommended changes in federal aid to nursing education. The National Commission for the Study of Nursing and Nursing Education (United States, 1971) produced *An Abstract for Action*—a blueprint for new developments in nursing.

In almost every country where nursing is well established, similar studies have been produced. Early studies in each country express the beginnings of awareness of need for modernization of nursing. Studies are also done when marked changes in health programs indicate that changes are needed in nursing—for example, at the time of inception of the National Health Service in England. Efforts of nurses and others to start nursing-education programs in colleges and universities usually involve studies of health needs and nursing personnel.

WHO publications

A recent trend of great value has been the collaboration of several disciplines and interests, so that resultant actions in health are unified. The inclusion of *Planning and Programming for Nursing Services*, published in 1971 in the World Health Organization series of public-health papers in English, French, Spanish, and Russian, can serve as an example. The inclusion of nurses among the official groups that study health needs in many countries is another effective and unifying influence.

The current nursing situation: the worldwide nurse shortage. The world has more nurses than ever before, yet all but a few countries claim marked shortages of nursing personnel.

In countries with highly developed health services, rapid advances in the medical sciences have increased what can be done to aid a patient. But to implement the new knowledge requires a marked increase in nursing personnel.

Table 2: Ratio of Nurses to Population in the Americas

	nurses per 10,000 population	auxiliaries per 10,000 population
North America	34.8	53.0
Central America	3.7	8.4
South America	2.3	9.2

Further, an increase in the educational status of large populations in mass communication increases demands for services, and the acceptance of health as a human right leads to the acknowledgement of the need to provide services to all persons in society. Still, it is doubtful that these demands will be met in the near future.

In the United States in the early 1970s there were approximately 750,000 registered nurses employed—a ratio of about 360 per 100,000 people. The need for registered nurses was estimated at 1,000,000 for 1975; and for 1980, 1,100,000. The goal for licensed practical nurses for 1975 was 550,000, some 150,000 more than at the beginning of the decade. Aides, orderlies, and attendants in the early 1970s were estimated at 850,000.

Canada reported a need for at least 120,000 registered nurses, 20,000 more than at the beginning of the 1970s.

In new countries that only recently initiated education systems for health personnel, shortages were severe. Older countries that were still economically underdeveloped also exhibited acute shortages. The ratio of nurses to population in the Americas in the late 1960s is shown in Table 2.

Worldwide variation in availability of nursing personnel is shown in Table 3, which perforce uses many rough estimates of both nurses and population.

Distribution of nurses

The Americas and Europe (including the U.S.S.R.) have almost three-fourths of the world's nurses, leaving about one-fourth for all of Asia, Africa, and Oceania.

Shortages are claimed even in countries with the highest ratios of nurses to population. It is asked whether these shortages could be alleviated, partially, at least, by better distribution of the supply and by making better use of the nurses available. Shortages are most damaging in the quality and quantity of teachers, since teachers are critical in any effort to increase and improve nursing.

Table 3: World Ratio of Nurses to Population

area	ratio of nurses to population	
	in country with highest ratio	in country with lowest ratio
Europe	1 to 330	1 to 1,200
Americas	1 to 230	1 to 6,000
Asia	1 to 430	1 to 8,200
Africa	1 to 800	1 to 12,500

ROLES OF THE INTERNATIONAL RED CROSS AND THE WHO

The International Red Cross. The International Red Cross plays two major roles in nursing: it affords educational opportunities for nurses, and it affords nurses opportunity to serve in programs embodying the Red Cross principles of humanity, impartiality, and neutrality. Through participation in Red Cross activities, public-spirited citizens acquire valuable knowledge of health needs of people and how nursing helps to meet these needs. Also, nurses broaden their concepts of community and worldwide service.

Twenty-five national Red Cross societies operate 240 schools of nursing; six operate postbasic education programs; and 60 operate training programs for auxiliaries. In addition, many societies carry on programs preparing nurses to teach health in the home; this instruction reaches many thousands of people, who thus learn individual responsibility for health and the care of mothers and children and of patients suffering from simple illnesses. Many societies also train nurses to care for people in disasters and emergencies and deploy nurses to sites of emergencies when needed; some nurses volunteer to serve in emergencies that occur in other countries.

Red Cross societies have established community nursing services. In the United States these services were forerunners of some local health-department nursing programs, and they serve similarly in other countries in the 1970s.

National Red Cross societies are placing new emphasis on community development and youth programs. Societies are urged to be constantly alert to social changes, such as those resulting from urbanization, migration, or drug abuse, and to coordinate their efforts with other voluntary and official agencies. Nurses with knowledge of community conditions can and do contribute to planning and action through board and community memberships, as well as in their participation in the carrying out of Red Cross programs.

The World Health Organization. The World Health Organization has included nursing in its activities from its beginning in 1948. Nurses are included on teams, such as those concerned with maternal and child health, malaria, and tuberculosis. The development of nursing itself was recognized as a health force, and member nations request assistance in developing educational programs for nurses, auxiliaries, and midwives and to organize public-health programs and hospitals.

Nursing activities of WHO

More than 130 nations are members of the World Health Organization; in 1971 alone, more than 100 of these were helped in their nursing needs by WHO. The many activities of nursing programs of the World Health Organization headquarters and the six regional offices may be illustrated by the following few examples.

Nurses were recruited from many countries and provided for health teams and nursing projects requested by member nations. In 1971, 370 nurses and nurse-midwives from various countries served on 200 projects, of which 33 concerned more than one country.

Help was given countries to establish nursing in the organization of their national health departments to assure the planning and implementing of the nursing portions of programs. A travelling seminar held in four cities of the U.S.S.R. was attended by nurses from 17 countries who studied the use of information about people's health needs in planning programs.

Governments were aided in the establishment of nursing and nursing-education systems, including those in midwifery. Other programs in which assistance was given included upgrading diploma schools; development of basic and postbasic programs in universities; revision of entry requirements; coordination of classroom teaching with clinical practice of students; adaptation of hospitals and health agencies for students' experience; preparation of public-health nurses, administrators, midwives, and teachers, including teachers of auxiliaries and of indigenous midwives. (In many countries more than half of all the births are attended by untrained midwives.)

Fellowships were granted to nurses for overseas study. Nearly 400 fellowships were granted in 1970, most of which were used to help students prepare for teaching, administration, public-health nursing, midwifery, maternal and child health, and for learning to plan health services. When consultants from the World Health Organization work in countries, they strive to leave national counterpart personnel to continue the work, and study outside the country may be needed to develop such personnel.

BIBLIOGRAPHY. AMERICAN NURSES' ASSOCIATION, *Facts About Nursing* (compiled and published every one or two years), a comprehensive compilation of statistical information about all fields of nursing and nursing education; JEROME P. LYSAUGHT, *An Abstract for Action* (1970), a condensed report of the National Commission for the Study of Nursing and Nursing Education on the characteristics and concepts of nursing practice in the United States; J.C. GOLDMARK, *Nursing and Nursing Education in the United States* (1923), a classic report on the status of nursing in the United States with recommendations for actions that are still pertinent; FRED DAVIS (ed.), *The Nursing Profession: Five Sociological Essays* (1966), a review in depth of these areas covered: nursing leadership (cross national); structure and organization of nursing practice; problems in collegiate education in nursing; and nurse-patient relationships; INTERNATIONAL COUNCIL OF NURSES, *Principles of Legislation for Nursing*

Education and Practices (1969), a guide (available in English, French, and Spanish) to nursing associations that desire to initiate or revise nurse practice acts; WORLD HEALTH ORGANIZATION (WHO), *The Work of WHO* (annual), report of the Director General, a comprehensive report of WHO program activities including nursing; PAN AMERICAN HEALTH ORGANIZATION AND REGIONAL OFFICE OF WHO (annual), report of the Director, a comprehensive report of program activities including nursing; UNITED STATES DEPARTMENT OF HEALTH, EDUCATION, AND WELFARE, PUBLIC HEALTH SERVICE, *Toward Quality in Nursing, Needs and Goals: Report of the Surgeon General's Consultant Group on Nursing* (1963), survey of nursing in the United States with recommendations for federal and other actions to expand and improve nurse-power; Y.L. and B. BULLOUGH, *The Emergence of Modern Nursing*, 2nd ed. (1969), a survey of prospects for nursing; UNITED STATES DEPARTMENT OF HEALTH, EDUCATION, AND WELFARE, DIVISION OF NURSING, *Health Manpower Service Book*, sect. 2, "Nursing Personnel" (1969), a compilation of statistical information about nursing—most of this and more is found in *Facts About Nursing* (above); ESTHER LUCILE BROWN, *Nursing Reconsidered*, 2 vol. (1970–71), report and interpretation of new functions performed by nurses and a review of developments in the last decade; I.M. STEWART and A.L. AUSTEN, *A History of Nursing*, 5th ed. (1962), a concise and well-documented history of nursing from earliest time to present.

(L.P.Le.)

Nutrition

Nutrition may be defined as the study of all of the processes by which micro-organisms, plants, and animals absorb and utilize food substances. Nutrition involves the identification of individual nutrients essential for growth and for the maintenance of individual organisms; it includes the determination of interrelationships among nutrients within individual organisms, as well as the evaluation of the quantitative requirements of organisms for specific nutrients under various environmental conditions. Treated in this article are the identity of nutrients, the ways in which they vary from one organism to another, and the ways in which requirements for certain of them arose. Nutrition in man is dealt with in a separate article, NUTRITION AND DIET, HUMAN.

GENERAL FEATURES

Functions of food. Food serves three functions in most living organisms. It provides materials that are metabolized either by oxidative or by fermentative processes to supply the energy required for the absorption and translocation of nutrients, for the synthesis of cell materials, for motility and locomotion, for excretion of waste products, and for all other activities of the organism. Second, food must supply the electron donors (reducing agents) required for the formation of the reduced coenzymes (enzyme components) normally necessary for the synthetic processes that occur within the cell. The functions of food as energy sources and electron donors are shown diagrammatically in the bottom half of Figure 1 labelled catabolism. In the third instance, food provides the materials from which all of the structural and catalytic components of the living cell can be assembled by processes sometimes called anabolism (Figure 1). But the three roles of food are not mutually exclusive; energy-yielding substances in many organisms may function in all three ways, and essential nutrients, if present in excess, may frequently be metabolized to supply energy.

By courtesy of World Health Organization

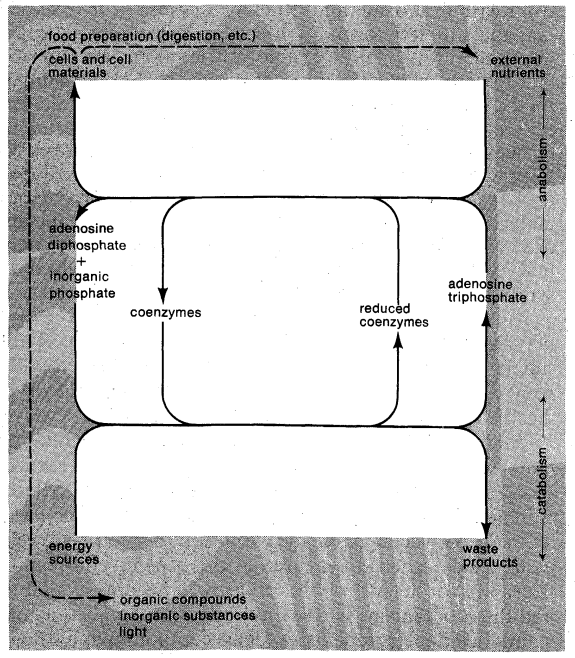


Figure 1: Relation between the functions of food (as energy source, ultimate electron donor, and source of essential nutrients for synthesis of cell materials) and the catabolic or anabolic phases of metabolism (see text).

The essential precursors (*i.e.*, the substances from which other substances are formed) of cell materials can be divided into two groups—nonessential nutrients, which can be synthesized by the cell from other materials, and essential nutrients, which, because they cannot be synthesized by the cell, must be supplied in foods. All of the inorganic materials required for growth, together with an assortment of organic compounds whose number may vary from one to 30 or more, depending on the organism, fall in the latter category. Although organisms are able to synthesize nonessential nutrients, such nutrients are nevertheless frequently utilized directly if present in food, thereby saving the organism the need to expend the energy required to synthesize them.

One method of classifying living organisms on a nutritional basis is based on the way in which the functions of food are carried out. Thus, organisms such as green plants and some bacteria that need only inorganic compounds for growth are called autotrophic organisms; and organisms, including all animals, fungi, and most bacteria, that require both inorganic and organic compounds for growth are called heterotrophic. Although these general terms are widely employed, as knowledge of nutritional behaviour of widely varied groups of organisms has increased, the classification has been extended to describe various nutritional patterns that have been observed. In one scheme, organisms are classified according to the energy source they utilize. Phototrophic, or photosynthetic, organisms trap light energy and convert it to chemical energy in the form of an energy-rich compound, adenosine triphosphate (ATP; see Figure 1); chemotrophic, or chemosynthetic, organisms utilize inorganic or organic compounds to supply their ATP requirements. An additional method of differentiation is based on the type of electron-donor material utilized to form the reduced coenzymes required for synthesis of cell constituents. If the electron-donor material consists of inorganic compounds, the organism is said to be lithotrophic; if organic, the organism is organotrophic. Combinations of these patterns may also be used to describe organisms; higher plants, for example, are photolithotrophic; *i.e.*, they utilize light energy, with the inorganic compound water serving as the ultimate electron donor. Certain photosynthetic bacteria, which cannot utilize water as the electron donor and require organic compounds for this purpose, are called photoorganotrophs. Animals, by this classification, are chemoorganotrophs, *i.e.*, they utilize chemical compounds to supply ATP and organic compounds as electron donors.

Nutritional patterns

Table 1 is a listing of several different types of organisms grouped according to these classifications. Such classification schemes are based on arbitrary criteria and, thus, provide a convenient description of the predominant life-style of an organism rather than an inflexible categorization of it. To illustrate, higher plants grown in sunlight are classic examples of autotrophic (or, more exactly, of photolithotrophic) organisms; during germination and growth of seedlings in the dark, however, higher plants have a heterotrophic (more specifically, a chemoorganotrophic) pattern of nutrition, utilizing the

Table 1: Some Representative Organisms of Various Nutritional Types

organism	energy source or substrate oxidized	oxidant	electron donor	waste product
Photolithotrophs				
Higher plants	light	—	water	oxygen
Green algae	light	—	water	oxygen
Photosynthetic bacteria				
Chlorobium species	light	—	hydrogen sulfide	sulfur
Chromatium species*	light		hydrogen sulfide or sulfur	sulfur or sulfate ion
Photoorganotrophic bacteria				
Athiorhodaceae†	light		organic compounds (*e.g.*, isopropanol)	oxidized compounds (*e.g.*, acetone)
Chemolithotrophic bacteria				
Nitrobacter species	nitrite ion	oxygen	same as substrate	nitrate ion
Nitrosomonas species	ammonia	oxygen	same as substrate	nitrite ion
Thiobacillus thiooxidans	sulfur	oxygen	same as substrate	sulfate ion
Hydrogenomonas species	hydrogen	oxygen	same as substrate	water
Carboxydomonas species	carbon monoxide	oxygen	same as substrate	carbon dioxide
Gallionella species	ferrous ion	oxygen	same as substrate	ferric ion
Chemoorganotrophs				
Animals	carbohydrates, fatty acids, amino acids, etc.	oxygen	same as substrate	carbon dioxide, water, ammonia or urea, etc.
Most bacteria	carbohydrates, fatty acids, amino acids, etc.	oxygen‡	same as substrate	‡
Most yeasts	carbohydrates	oxygen‡	same as substrate	carbon dioxide, water‡

*Some species also grow as photoorganotrophs. †Some species also grow as photolithotrophs. ‡Some species of bacteria and yeasts also grow without oxygen (*i.e.*, anaerobically); others grow only in the absence of oxygen (*i.e.*, they are obligate anaerobes). In such cases oxygen is not used; the substrate undergoes oxidoreduction reactions to yield ATP, together with a mixture of partially oxidized and reduced compounds as waste products.

organic compounds stored within the seed for growth. In contrast, certain chemolithotrophs—*e.g.*, *Thiobacillus thiooxidans*, a bacterium that obtains energy by oxidizing sulfur or hydrogen sulfide gas to sulfate and utilizes carbon dioxide as a sole source of carbon—cannot become organotrophic or heterotrophic under any conditions yet tested. These nutritional designations are useful, even though some predominantly chemolithotrophic organisms may have lost the ability to synthesize one or more organic compounds and, thus, require them for growth.

Despite wide variations in the nature of the external energy source utilized by various organisms, all organisms form from it the immediate intracellular source of energy, ATP, which is common to all cells. Through its conversion during metabolism to a less energy rich compound, adenosine diphosphate (ADP), ATP provides the energy for the chemical and mechanical work required if an organism is to maintain itself. This similarity in metabolic organization of even the most diverse organisms is shown diagrammatically in Figure 1, which emphasizes food as energy sources and as cell-precursor materials. The energy requirements of the organism can be measured in units called calories; the need for cell-precursor materials cannot be. In a related article, METABOLISM, bioenergetics and intermediary metabolism are considered.

All animals, fungi, and most bacteria are chemoorganotrophs (*i.e.*, organic compounds supply both ATP and serve as electron donors). In these organisms the multiple roles played by foodstuffs are partially obscured because the carbon-containing compounds that serve as energy sources also serve as electron donors and provide intermediate compounds from which many cell materials are synthesized. Similarly, essential organic nutrients, when present in excess, may be utilized in part to provide the energy needs of the organisms.

Nutritional evolution of organisms. Little factual knowledge concerning the nutritional evolution of living organisms is available. Nucleic acids, proteins, carbohydrates, and fats, which are present in all living cells, are formed by specific reaction sequences from a limited number of smaller compounds, most of which are common to all living organisms and, according to current theories, were available on earth before life arose. Since less complex metabolic organization and less energy are required, for example, to synthesize cellular proteins from the preformed amino acids that comprise them than from carbon dioxide and other precursors, it is assumed that the simplest forms of life were heterotrophic organisms requiring many organic nutrients for growth and that they selected such nutrients from their surroundings.

Evolution of autotrophy

As the supply of these preformed substances was exhausted (*e.g.*, through uptake by heterotrophic organisms or through fermentative and oxidative processes in organotrophic organisms), the organisms presumably developed the capacity to synthesize them from simpler (precursor) materials present in the environment; in some organisms, this synthesizing capacity eventually evolved to the extent that carbon from carbon dioxide could be utilized to synthesize organic compounds. At this point, autotrophy as it now is known became possible; autotrophy, in fact, may have evolved as a result of the exhaustion of the supply of preformed organic materials in the environment and the consequent necessity of organisms to synthesize the requirements themselves in order to survive. Implicit in this theory is the demonstrable assumption that autotrophic cells contain the most complex biosynthetic organization found in living things and that heterotrophic cells are simpler in that certain biosynthetic pathways do not occur. After the evolution of photosynthesis, a constantly renewable source of the organic compounds necessary for heterotrophic cell growth became available, and those organisms whose environments supplied a constantly available supply of a given compound could lose, through changes in their genetic material (mutations), the ability to synthesize that compound and still survive. Entire biosynthetic pathways may have been lost in this

way; as long as such mutant organisms remained in an environment that supplied the necessary compound, the simplification in cellular organization and the energy saved by using preformed cell components would have given them a competitive advantage over the more complex parents from which they were derived and permitted stabilization of the mutation within the cell type. A theory that the requirements of present-day organisms for essential organic nutrients arose through the loss of synthetic abilities present in more complex parent organisms was expressed as early as 1913, elaborated in classical papers published between 1936 and 1940, and confirmed by the discovery during the period from 1940 to 1950 that artificially produced mutant offspring of micro-organisms can be readily obtained and may require the presence of one or more preformed organic compounds, which the parent micro-organisms could synthesize (see also LIFE).

Methods of ingestion or penetration of nutrients. Foodstuffs of chemoorganotrophs are composed chiefly of large molecules, such as nucleic acids, proteins, lipids, and carbohydrates, which are derived from other organisms. When these materials are eaten as food, they usually must be broken down into the low-molecular-weight substances from which they were formed (*e.g.*, purine and pyrimidine bases from nucleic acids, amino acids from proteins, fatty acids from fats, and simple sugars from carbohydrates) before they can be absorbed and utilized. The breakdown process is generally accomplished outside the cells (extracellularly) of both single-celled and multicellular organisms through hydrolysis (break-down involving water), catalyzed by specific enzymes secreted for this purpose. (See DIGESTION AND DIGESTIVE SYSTEMS.) Some single-celled organisms (*e.g.*, the protozoan genus *Amoeba*), however, engulf particulate matter and hydrolyze it intracellularly.

The organic products of digestion produced extracellularly, together with essential inorganic nutrients, must cross one or more cell membranes before they can be utilized for growth. The absorption process is not yet understood and may vary according to the nature and the concentration of the nutrient involved. When nutrient concentrations are high, some materials (*e.g.*, water) may diffuse through the cell membrane to enter the cell. Efficient absorption of some materials, however, can occur from solutions containing very small quantities of nutrients, frequently against a concentration gradient; *i.e.*, from an extracellular solution containing very small quantities of the nutrients to an intracellular solution containing larger quantities. Such transfer processes, referred to as active transport, are energy dependent and are catalyzed within the cell membrane by a variety of systems, each of which is specific for the compound or class of compounds being transported. The nature of such catalysts has not yet been established with certainty; those best characterized involve either specific proteins with high affinity for the nutrient substance being transferred, enzymes that chemically modify the nutrient during its passage through the membrane or both.

In multicellular animals, absorbed nutrients are transferred into the circulating blood and lymph, from which they are reabsorbed by individual cells. In higher plants, translocation of nutrients occurs through transport elements found in the plant tissues (phloem and xylem). (See also MEMBRANE, BIOLOGICAL; PLANT INTERNAL TRANSPORT.)

The determination of essential nutrient requirements. It has been estimated that the approximate quantities of chemical elements in the human body (in percent of wet weight) are oxygen, 65; carbon, 18; hydrogen, 10; nitrogen, 3; calcium, 2; phosphorus, 1.1; potassium, 0.35; sulfur, 0.25; sodium, 0.15; chlorine, 0.15; magnesium, 0.05; iron, 0.004; manganese, 0.00013; copper, 0.00015; and iodine, 0.00004. Also present are traces of about 20 other elements—*e.g.*, zinc, cobalt, aluminum, arsenic, barium, boron, bromine, cadmium, chromium. Analytical figures such as those for man given above serve to distinguish between chemical elements that occur in relatively large amounts—oxygen, carbon, nitrogen, sulfur, and phosphorus—and the so-called trace elements, an obsolescent term applied to elements that are present in small amounts and that at one time could be detected in tissues but not analyzed accurately.

Trace elements

Although of interest, analyses such as the above are not instructive from the nutritional standpoint since only a few of the elements (*e.g.*, oxygen) are utilized in uncombined form in metabolic processes, and the mere presence in tissues of a trace element (the term is used loosely from this point to include all combined forms of the element that occur naturally; see Table 2) does not reveal whether its presence results from incidental contamination or whether it plays an essential metabolic role and therefore qualifies as an essential inorganic nutrient. In succeeding sections, inorganic and organic compounds that serve as essential nutrients (*i.e.*, cannot be synthesized by a cell and must be supplied in food) for one or another of the organisms studied thus far are dealt with.

The methods used to determine the individual pure substances necessary for growth are similar in principle for all organisms. Occasionally, crude diets lacking only one or a few nutrients result in some naturally occurring deficiency disease; the substance that cures such a condition can be determined directly by adding the nutrients to the diet. Certain vitamins were recognized in this way. In the absence of a deficiency condition, a crude nutrient medium (or diet) adequate for growth is selected for the organism to be studied, and attempts are then made to replace the various ingredients of the medium with purified materials and, finally, with substances of known composition. If replacement with substances of known composition permits growth, the added materials necessary for growth can be determined with relative ease; if the pure substances do not allow growth, attempts are next made to isolate and identify by chemical procedures the substances present in crude materials that permit growth.

Table 2: Essential Inorganic Nutrients for Living Organisms*

element	utilizable ionic form	representative organisms exhibiting the requirement
Boron	$B_4O_7^{2-}$	certain vascular plants and algae; no evidence of an animal requirement
Calcium	Ca^{2+}	plants, animals, most micro-organisms
Chlorine	Cl^-	higher animals; no evidence for requirement in plants
Chromium	Cr^{3+}	probably essential in higher animals
Cobalt	Co^{2+}	essential in ruminants; probably functions chiefly through microbial incorporation into vitamin B_{12}
Copper	Cu^{2+}	plants, animals, most micro-organisms
Fluorine	F^-	highly beneficial to bone and tooth formation in animals, including man
Iodine	I^-	higher animals; no evidence for requirement in plants or micro-organisms
Iron	Fe^{2+}	animals, higher plants, most micro-organisms
Magnesium	Mg^{2+}	animals, plants, micro-organisms
Manganese	Mn^{2+}	animals, plants, micro-organisms
Molybdenum	MoO_4^{2-}	animals, plants, nitrogen-fixing bacteria
Nitrogen	NO_3^-, NH_4^+	plants, micro-organisms (animals derive nitrogen mostly from organic sources, and utilize limited amounts of NH_4^+, but not NO_3^-)
Phosphorus	PO_4^{3-}	animals, plants, micro-organisms
Potassium	K^+	animals, plants, micro-organisms
Selenium	SeO_4^{2-}	higher animals
Silicon	SiO_4^{4-}	certain Protozoa and Porifera
Sodium	Na^+	animals, some plants, some marine bacteria
Sulfur	SO_4^{2-}	plants, many bacteria (animals derive sulfur mostly from organic sources)
Vanadium	VO_4^{2-}	various tunicates and holothurian echinoderms; some algae
Zinc	Zn^{2+}	all animals, plants, most micro-organisms

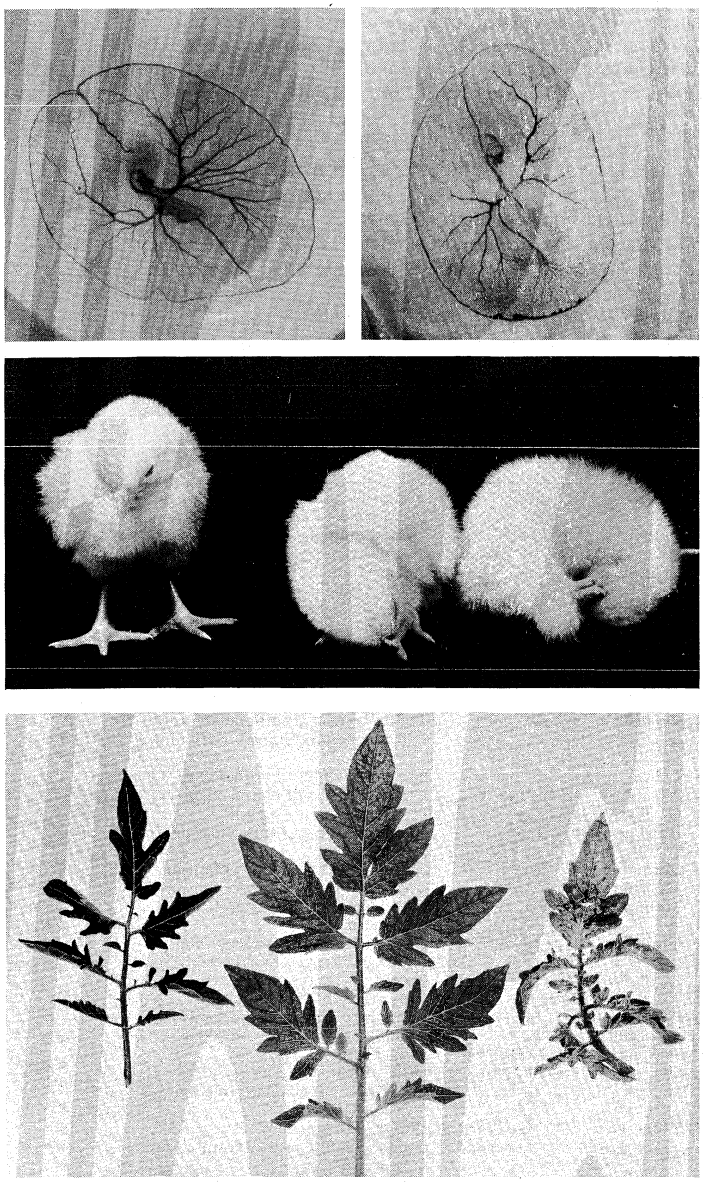
*Carbon dioxide (CO_2) and water (H_2O), although not formally tabulated, are important nutrients for all organisms; external sources of carbon dioxide, however, are required for only a few heterotrophic organisms since it is a prominent waste product of energy metabolism in these organisms.

The process was initially slow, because little was known of the detailed composition of tissues, and very few pure nutrients were available. With the identification of each new nutrient, however, the job of identifying additional ones became easier. Now that detailed nutritional requirements of many organisms are known, and essentially all of the organic nutrients are available as pure compounds, determination of the nutrient requirements of a previously unstudied organism is a comparatively easy task. Elucidation of trace-element requirements remains troublesome, however, because these substances occur commonly as contaminants in other materials.

ESSENTIAL INORGANIC NUTRIENTS

Table 2 is a list of inorganic elements (minerals) that are essential for the growth of living things, together with one or more examples of organisms that exhibit a requirement for them. The list excludes inorganic substances that serve only as energy sources for certain lithotrophic organisms. The effects of certain mineral deficiencies on the development of higher plants and animals are illustrated by several photographs. Naturally occurring deficiencies of several different trace elements occur throughout the world in special types of soils and are therefore of tremendous economic and agricultural importance. Several excellent accounts of these deficiencies have been published (see below *Bibliography* under Underwood, Hoagland, Altman).

By courtesy of (bottom) D. Arnon; from (top left, top right, centre) J.E. Savage, *Federation Proceedings*, vol. 27, p. 927 (1968)



Effects of mineral deficiencies on the development of higher plants and animals.
(Top left) Normal 72-hour embryo from hen fed ten parts per million of supplementary copper. (Top right) Anemic 72-hour embryo from hen receiving copper-deficient diet. (Centre) A normal chick, left, from hen receiving an adequate amount of zinc in its diet and two chicks from hens given zinc-deficient diets. (Bottom) Tomato leaves from plants grown in (left to right) copper-deficient soil, complete nutrient solution, and zinc-deficient soil.

Several examples of specialized mineral requirements are present in Table 2. Boron, for example, has been demonstrated to be required for the growth of many—perhaps all—higher plants but has not been implicated as an essential element in the nutrition of either micro-organisms or animals. Trace amounts of fluorine (as fluoride) are certainly beneficial and perhaps essential, for proper tooth formation in higher animals, but no essential role for fluorine in other organisms has been found. Similarly, iodine (as iodide) is required in animals for formation of thyroxine, the active component of an important regulatory hormone; it plays no known role in other organisms. Silicon (as silicate) is a prominent component of the outer skeletons of diatomaceous protozoans and similar organisms and is required in them for normal growth. Although silicon is present in higher animals, no essential role for it has been found thus far, and no nutritional requirement has been demonstrated. A less obvious example of a specialized mineral requirement is provided by calcium, which is required by higher animals in comparatively large amounts because it plays a role in the formation of bone and eggshells (in birds); for other organisms, calcium is an essential nutrient but only as a trace element. Several additional examples of specialized requirements are evident from Table 2.

Specialized mineral requirements

Mineral elements in wide variety are present in trace amounts in almost all foodstuffs. For this reason, diets adequate for growth but completely free of a given mineral element are extremely difficult to prepare. As a consequence it is not yet certain that all of the essential mineral nutrients are known. Furthermore, it cannot be assumed that, because a given mineral element is nonessential, it plays no useful role in metabolism. For example, the potassium requirement of certain bacterial species, called lactic-acid bacteria, which produce lactic acid as a byproduct of metabolism, can be completely replaced by rubidium. But, in other species, small amounts of potassium will still be required, even though rubidium is present. Thus, rubidium, although it is not an essential nutrient, can apparently fulfill in certain species many or all of the essential functions normally performed by potassium; in a specialized environment low in potassium, rubidium could partially replace the essential nutrient. (See below *Bibliography* under MacLeod and Snell.) Strontium has a somewhat similar relationship to calcium in these and other organisms.

Important antagonistic relationships between mineral nutrients also are known. A large excess of rubidium, for example, interferes with potassium utilization in some lactic-acid bacteria; zinc can also interfere with manganese utilization in the same organism. In animal nutrition, excessive molybdenum or zinc (both essential minerals) interferes with the utilization of copper, another essential mineral, and, in higher plants, excessive zinc can lead to a disorder known as iron chlorosis. Proper nutrient growth media for micro-organisms and plants or diets for animals, therefore, require not only that the essential mineral elements be provided in sufficient amounts but also that they be used in the proper ratios to each other.

Antagonistic relationships between mineral nutrients

ESSENTIAL ORGANIC NUTRIENTS

The principal energy sources of chemoorganotrophic organisms (carbohydrates, proteins, and fats) and the products of their breakdown are derived from other living organisms. In addition, bacteria from soil can be grown using as energy sources either hydrocarbons (organic compounds containing carbon and hydrogen) or any of a large variety of synthetic organic compounds that do not occur naturally and are not considered here.

Considered in this section is the nature of the essential organic nutrients that certain organisms can no longer synthesize and that, since they are essential building blocks of various cell components, must be supplied to them preformed. These essential compounds include certain amino acids, which are the precursors of protein; purine and pyrimidine bases or their derivatives, which are the precursors of nucleic acids; fatty acids, which are precursors of fats and phospholipids; and, occasionally (in mutant organisms), simple sugars (*e.g.*, glucose). Other essential organic nutrients include the vitamins, which are required in very small amounts (relative to, for example, the amino acids), because of either the catalytic role or the regulatory role they play in metabolism.

Amino-acid requirements. Proteins usually contain about 20 different amino acids (see Table 3) or their derivatives, and there are organisms known that require none, several, or most of them for growth. Many yeasts and bacteria (*e.g.*, *Escherichia coli*) are able to synthesize all of their amino acids from other compounds (glucose and ammonium salts). Animals may require up to nine or ten different amino acids in their diets (Table 3). Species of lactic-acid bacteria have been identified as requiring none (*Streptococcus bovis*), a few (*Lactobacillus plantarum*), or as many as 17 amino acids; *Streptococcus equinus* requires all of the amino acids listed in Table 3 except alanine.

Table 3: Amino Acids Required for Growth of Various Organisms

amino acid*	essential nutrient for
L-Alanine	*Pediococcus cerevisiae*
L-Arginine	chicks, rats, *Lactobacillus casei*
L-Aspartic acid	*Streptococcus equinus*
L-Cysteine	*S. equinus*
L-Glutamic acid	*Lactobacillus plantarum*
Glycine	*S. equinus*
L-Histidine	rats, chicks, man, *Streptococcus faecalis*
L-Isoleucine	rats, chicks, man, *S. faecalis*
L-Leucine	rats, chicks, man, *S. faecalis*
L-Lysine	rats, chicks, man, *S. equinus*
L-Methionine	rats, chicks, man, *S. faecalis*
L-Phenylalanine	rats, chicks, man, *S. equinus*
L-Proline	*S. equinus*
L-Serine	*Lactobacillus delbrueckii*
L-Threonine	rats, chicks, man *L. plantarum*
L-Tryptophan	rats, chicks, man, *L. plantarum*
L-Tyrosine	*L. delbrueckii*
L-Valine	*L. plantarum*, rats, chicks, man

*For structures of the amino acids, see PROTEIN.

Randomness of amino-acid requirements

The apparently random distribution of amino-acid requirements is compatible with the theory that the loss of synthesizing abilities occurred as the result of a random mutation process followed by stabilization of the mutations in different species (see above *Nutritional evolution of living organisms*). Indeed, mutant forms of the bacterium *Escherichia coli* that require any of the amino acids of Table 3 are readily obtained by appropriate procedures; the parent type *E. coli*, however, does not require any amino acids as essential nutrients. It is curious, therefore, that the nutritionally demanding lactic-acid bacteria and the much less demanding *E. coli* exist side by side and are the most prominent inhabitants of the intestinal tract of animals. Two amino-acid derivatives not listed in Table 3, glutamine and asparagine, are products of enzymatic digestion of proteins and are required for growth of some micro-organisms. An apparent requirement of some organisms for peptides, which are compounds containing two or more amino acids, is considered below.

Purine and pyrimidine bases and related compounds. A large number of species of micro-organisms require one or more of these substances as precursors for the synthesis of nucleic acids. Examples are given in Table 4; all higher animals examined have retained the capacity to synthesize purine and pyrimidine bases. Certain compounds (putrescine, spermidine, spermine) that are closely associated with nucleic acids in living cells, although synthesized by most organisms, are required preformed by certain bacteria; *e.g.*, *Haemophilus parainfluenzae, Neisseria perflava.*

Table 4: Purines, Pyrimidines and Their Derivatives Required as Essential Nutrients by Certain Micro-organisms*

compound	example of dependent micro-organism
Pyrimidine bases	
Uracil	*Staphylococcus aureus*
Orotic acid	*Lactobacillus bulgaricus* 09
Purine bases	
Adenine	*Shigella boydii* 9329
Guanine	*Streptococcus equinus*
Hypoxanthine	*Neisseria gonorrhoeae*
Xanthine	*Streptococcus equinus*
Pyrimidine deoxyribonucleosides	
Thymidine	*Lactobacillus delbrueckii* 730
Purine ribonucleotides	
Adenosine	*Gaffkya homari*
Guanosine	*Gaffkya homari*
Cytidine	*Tetrahymena geleii*

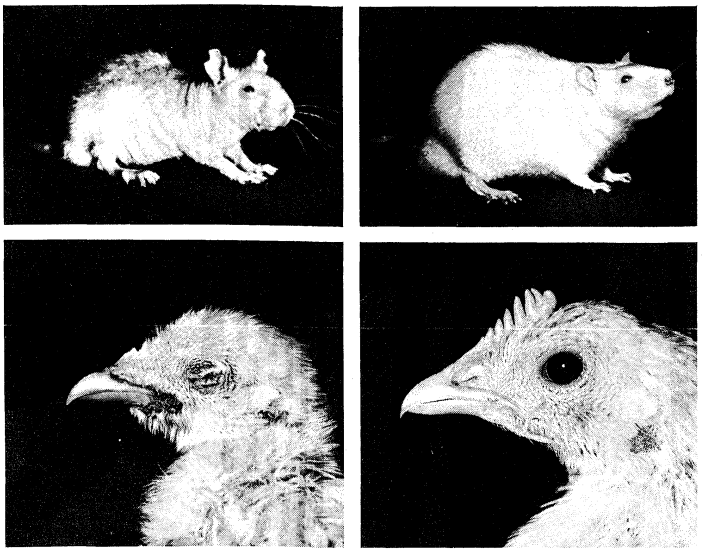
*Adapted from Guirard and Snell (1962) from which references may be obtained.

Lipids and related compounds. Phospholipids are essential components of all cell membranes, and it is not surprising, therefore, that some organisms, having lost the ability to synthesize certain lipids (since they are supplied in their diets), require them as essential nutrients. Essential fatty acids for rats (and most vertebrates) include linoleic and arachidonic acids and, perhaps, linolenic acid, as well. For many bacteria (*e.g.*, *Erysipelothrix rhusiopathiae*, several species of lactic-acid bacteria), oleic acid is essential for growth, although it can be replaced in most cases by linoleic acid and by an unusual fatty acid called lactobacillic acid. (See also LIPID.)

Choline, inositol, and cholesterol

Choline, a nitrogen-containing compound found in the phospholipid lecithin, is usually listed among the nutrients required by rats, although it can be synthesized by them if sufficient vitamin B_{12} is present. Choline is also required for growth of certain pneumococcal bacteria. Inositol, another structural component of certain lipids, is an essential nutrient for many yeasts but appears to be synthesized by most animals. Finally, the steroid cholesterol is an essential nutrient for a wide variety of protozoans and for the bacteria-like organisms known as mycoplasmas. Higher animals are able to synthesize this important cell constituent; it is not present in most bacteria.

Vitamins. One of the most dramatic continuing chapters in the history of the study of nutrition has been the discovery, isolation, and characterization of the many trace organic components required both in the diets of animals and in the growth media of micro-organisms. Vitamins may be defined as nutritionally essential organic substances that play a catalytic role within the cell, usually as components of coenzymes or other groups associated with enzymes. It is not generally realized that over half of the water-soluble vitamins required by man were discovered when they were found to be growth factors for micro-organisms. In Table 5 are listed the vitamins found thus far to be required by animals, by micro-organisms or by both; photographs are included to show typical effects of vitamin deficiency in animals. Vitamin deficiencies in young animals usually result in growth failure, various symptoms whose nature depends on the vitamin, and eventual death. Certain vitamins (retinol, calciferol, tocopherol and ascorbic acid) appear to play no essential role in single-celled organisms. Para-aminoben-



Effects of vitamin deficiencies in animals.
(Top) Rat fed biotin-deficient diet. Same rat after three months on a diet with an adequate amount of biotin. (Bottom) Chick fed a diet deficient in pantothenic acid. Same chick after three weeks on a diet sufficient in pantothenic acid.
By courtesy of The Upjohn Company, Kalamazoo, Michigan

zoic acid is a vitamin for several bacteria only because it is an essential precursor of the vitamin folic acid, which, since it is unable to cross the bacterial cell membrane, must be synthesized from para-aminobenzoic acid. Heme and lipoic acid are typical vitamins for organisms that cannot synthesize them, but they have a catalytic role in all organisms; they are synthesized by higher animals. A few bacteria have lost the ability to convert certain vitamins to their functional forms, the coenzymes; the preformed coenzymes (or intermediates in their synthesis) thus are vitamins for such organisms. *Haemophilus parainfluenzae*, for example, requires nicotinamide adenine dinucleotide (a coenzyme derived from nicotinic acid), certain strains of *Neisseria gonorrhoeae* required thiamine pyrophosphate (a coenzyme derived from thiamine), and certain lactic-acid bacteria require either pyridoxamine phosphate (a coenzyme derived from vitamin B_6) or pantetheine (a coenzyme derived from pantothenic acid) for growth. Further details of this important group of nutrients are dealt with in the article VITAMIN, and in detailed accounts of bacterial or animal nutrition (see below *Bibliography*).

Table 5: Vitamins Required by Various Organisms

vitamin*	examples of dependent organisms
Fat-soluble vitamins	
Retinol (Vitamin A)	all vertebrates; many invertebrates
Calciferol (Vitamin D)	most vertebrates
Tocopherol (Vitamin E)	many vertebrates
Vitamin K	many vertebrates; *Mycobacterium paratuberculosis*
Water-soluble vitamins	
Ascorbic acid (Vitamin C)	man, primate apes, guinea pigs
Biotin	all vertebrates studied; many microorganisms
Folic acid (Pteroylglutamic acid)	all vertebrates studied; many bacteria (*e.g., Lactobacillus casei, Streptococcus faecalis*)
Nicotinic acid	all vertebrates investigated; many yeasts and bacteria
Pantothenic acid	all vertebrates; many yeasts, bacteria, and protozoans
Riboflavin (Vitamin B_2 or G)	all vertebrates; many bacteria, protozoans
Thiamine (Vitamin B_1)	all vertebrates; many fungi, yeasts, bacteria, and protozoans
Vitamin B_6 (Pyridoxine, pyridoxal, pyridoxamine)	all vertebrates; many yeasts, bacteria, and protozoans
Vitamin B_{12} (Cobalamine)	all vertebrates studied; some bacteria
Para-aminobenzoic acid	*Acetobacter suboxydans, Clostridium acetobutylicum*
Heme	*Haemophilus influenzae*
Lipoic acid	*Streptococcus lactis, Tetrahymena geleii*

*For chemical formulas and references, see related article VITAMIN and bibliographic citations.

Interdependency of nutritional requirements. The effects of one mineral nutrient in reducing or increasing the requirement for another have been mentioned previously (see above *Essential inorganic nutrients*). Similar relationships occur among organic nutrients and originate for several reasons, the most common of which are discussed briefly below.

Competition for sites of absorption by the cell. Since absorption of nutrients frequently occurs by way of active transport processes within cell membranes, an excess of one nutrient (A) may inhibit absorption of a second nutrient (B), if they share the same absorption pathway. In such cases, the apparent requirement for nutrient B increases; B, however, can sometimes be supplied in an alternate form that is able to enter the cell by a different route. Many examples of amino-acid antagonism, in which inhibition of growth by one amino acid is counteracted by another amino acid, are best explained by this mechanism. Under some conditions, for example, *Lactobacillus casei* requires both D- and L-alanine, which differ from each other only in the position of the amino, or NH_2, group in the molecule, and the two forms of this amino acid share the same absorption pathway. Excess D-alanine inhibits growth of this species, but the inhibition can be alleviated either by supplying additional L-alanine or, more effectively, by supplying peptides of L-alanine. The peptides enter the cell by a pathway different from that of the two forms of alanine and, after they are in the cell, can be broken down to form L-alanine. Relationships of this type provide one explanation for the fact that peptides are frequently more effective than amino acids in promoting growth of bacteria (see below *Bibliography* under Leach and Snell).

Amino-acid antagonism

Competition for sites of utilization within the cell. This phenomenon is similar to that regarding competition for absorption sites, but it occurs inside the cell and only between structurally similar nutrients (*e.g.*, leucine and valine; serine and threonine).

Precursor–product relationships. The requirement of rats for the essential amino acids phenylalanine and methionine is substantially reduced if tyrosine, which is formed from phenylalanine, or cysteine, which is formed from methionine, is added to the diet. These relationships are explained by the fact that tyrosine and cysteine are synthesized in animals from phenylalanine and methionine, respectively. When the former amino acids are supplied preformed, the latter are required in smaller amounts. Several instances of the sparing of one nutrient by another because they have similar precursor–product relationships have been identified in other organisms.

Changes in metabolic pathways within the cell. It has been well established that rats fed diets containing large amounts of fat require substantially less thiamine (vitamin B_1) than do those fed diets high in carbohydrate. The utilization of carbohydrate as an energy source (*i.e.*, for ATP formation) is known to involve an important thiamine-dependent step, which is bypassed when fat is used as an energy source, and it is assumed that the lessened requirement for thiamine results from the change in metabolic pathways. *Streptococcus faecalis* provides additional examples, in that either the vitamin folic acid or purine bases plus a pyrimidine base (thymine) and either the vitamin biotin or the fatty acid oleic acid are required for growth of this organism on an otherwise-adequate nutrient medium. These relationships occur because the synthesis of purine and pyrimidine bases requires a folic-acid-dependent enzyme, and synthesis of oleic acid requires a biotin-dependent enzyme. The products, thymine and oleic acid, must therefore be available to the cell if growth is to occur in the absence of folic acid or biotin; growth occurs, on the other hand, if either the products or the vitamins that permit their synthesis are supplied to the cells in the nutrient medium. Many known similar examples emphasize the fact that the nutritional requirements of a given organism can vary within certain limits, depending upon the composition of the nutrient medium.

COMPARATIVE MAGNITUDES OF THE REQUIREMENTS FOR DIFFERENT TYPES OF NUTRIENTS

Trace elements and vitamins are distinguished from other nutrients on the basis of the relatively smaller amounts required. An idea of the magnitude of these differences is provided in Figure 2. Potassium, a mineral element present in comparatively large amounts in all organisms, is required by *Lactobacillus casei* in amounts over 200 times greater than the requirement for manganese (Figure 2). The magnitude of the requirement for potassium is similar to that for the amino acid leucine (Figure 2); that for manganese is similar to that for the vitamin pantothenic acid but substantially larger than that for another vitamin, folic acid. Nutrients such as manganese and the vitamins, which are required in small amounts, are sometimes referred to as micronutrients because of these quantitative relationships. Potassium and leucine are macronutrients. For *Lactobacillus*, glucose is the energy source, and, to permit the amount of growth described, approximately 1,000 times more glucose is required than is needed of the macronutrients potassium and leucine. These ratios are fairly representative of organoheterotrophs in general. Lactic-acid bacteria, however, form lactic acid from glucose; other organisms, which oxidize this substance to carbon dioxide and water, require only about 5 percent of the amount of glucose required by lactic-acid bacteria for equivalent ATP production and, hence, for equivalent growth.

From E.E. Snell, *Wallerstein Laboratory Communications* (1948)

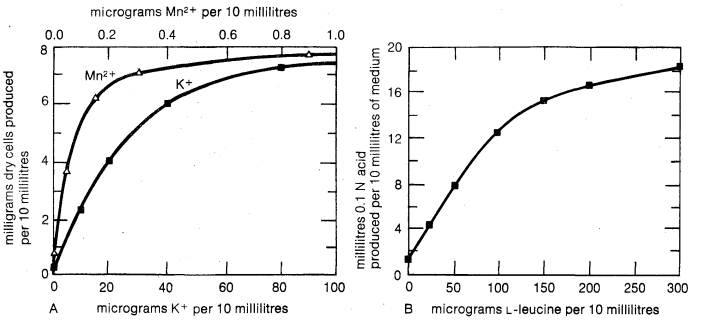


Figure 2: *Amounts of nutrients needed for growth of lactic acid bacteria.*
(A) Potassium (K^+) and manganese (Mn^{2+}) required for *Lactobacillus casei*. (B) Leucine required for *L. plantarum*.

Distinction between micronutrients and macronutrients

Graphic plots of data, called dose–response curves (Figure 2), if obtained in nutrient media that contain an excess of nutrients other than the one being examined, provide the basis for a convenient method (called microbiological assay) whereby all of the nutrients required by a specific organism may be determined. It is necessary only to compare the growth response elicited in a nutrient medium containing an unknown quantity of a nutrient to that produced in nutrient media containing various known quantities of the pure nutrient in order to determine the amount of the nutrient present in the unknown. Microbiological procedures employing lactic-acid bacteria have been widely used to determine quantities of the soluble vitamins in foods since they were first used, in 1939, for the determination of riboflavin. Similar methods, employing organisms that grow on soil, frequently have been used to determine deficiencies of inorganic nutrients in soil samples. The methodology is much more convenient and more accurate than are animal assays; the principle of all such assays, however, is identical, and any organisms with known specific nutritional requirements for a given substance can be used for its determination in this way.

SYNTROPHISM

Since the nutritional requirements and metabolic activities of organisms differ, it is clear that two or more organisms growing together may produce quite different overall changes in the environment than would the same organisms growing separately. A rough example is provided by a balanced aquarium, in which aquatic plants utilize light and the waste products of animals—*e.g.*, carbon dioxide, water, ammonia—to synthesize cell materials and generate oxygen, which in turn provide the materials necessary for animal growth. Such relationships are common among micro-organisms; *i.e.*, intermediate or end products of metabolism of one organism may provide essential nutrients for another. The mixed populations that result in nature provide examples of this phenomenon, which is called syntrophism; in some instances, the relationship may be so close as to constitute nutritional symbiosis, or mutualism. Several examples of this phenomenon have been found among thiamine-requiring yeasts and fungi, certain of which (group A) synthesized the thiazole component of thiamine molecule but require the pyrimidine portion preformed; for a second group (group B), the relationship is reversed. When group A and group B are grown together in a thiamine-free medium, both types of organisms survive, since each organism synthesizes the growth factor required by its partner; neither organism grows alone under these same conditions. Thus, two or more types of micro-organisms frequently grow in situations in which only one species would not. Such nutritional interrelationships may explain the previously noted fact that the nutritionally demanding lactic-acid bacteria happily coexist with the nutritionally nondemanding coliform bacteria in the intestinal tracts of animals. It is known that the bacterial flora of the intestinal tract synthesize sufficient amounts of certain vitamins (*e.g.*, vitamin K, folic acid) so that detection of deficiency symptoms in rats requires special measures, and the role of rumen bacteria in ruminant animals (*e.g.*, cows , sheep) in rendering otherwise-indigestible cellulose and other materials available to the host animal is well known. These few examples indicate that syntrophic interrelationships are widespread in nature and may contribute substantially to the nutrition of a wide variety of species. Parasitism, in which the host organism supplies all of the requirements for growth of the associated parasite without deriving any apparent advantage in return, represents a less desirable type of nutritional dependency.

Influence of populations on each other

BIBLIOGRAPHY. A very large and widely scattered literature exists on nutrition of micro-organisms, plants, and animals. Only a sampling is given below. Reasonably current lists of nutrients—both organic and inorganic—required by plants and animals, with a brief description of deficiency symptoms, may be found in the handbook by P.L. ALTMAN and D.S. DITTMER (eds.), *Metabolism* (1968). Similar compilations for bacteria and yeasts are presented by B.M. GUIRARD and E.E. SNELL, "Nutritional Requirements of Microorganisms," in I.C. GUNSALUS and R.Y. STANIER (eds.), *The Bacteria*, vol. 4, pp. 33–93 (1962); and by S.A. KOSER, *Vitamin Requirements of Bacteria and Yeasts* (1968). The latter two references also give an account of the discovery and inter-relationships of the organic nutrients for micro-organisms. Extended accounts of the vitamin requirements of animals, including their discovery and assay, are presented by W.H. SEBRELL and R.S. HARRIS (eds.), *The Vitamins*, 2nd ed., 7 vol. (1967–). N. JOLLIFFE describes the major vitamin deficiencies in man in *Clinical Nutrition*, 2nd ed. (1962). Amino-acid nutrition of man has been discussed fully in A. ALBANESE (ed.), *Protein and Amino Acid Nutrition* (1959). More specific accounts of the chemistry of the vitamins and the reactions by which they are degraded in living organisms are presented in M. FLORKIN and E.H. STOTZ (eds.), *Comprehensive Biochemistry*, vol. 11, *Water-Soluble Vitamins, Hormones, Antibiotics*, and vol. 21, *Metabolism of Vitamins and Trace Elements* (1963, 1970). Classical and still very useful accounts of the inorganic nutrition of plants and animals (including man) are, respectively, the books by D.S. HOAGLAND, *Inorganic Elements in Plant Nutrition* (1945), and by E.J. UNDERWOOD, *Trace Elements in Human and Animal Nutrition*, 3rd ed. (1971). R.A. MACLEOD and E.E. SNELL discuss cases of ion antagonism in bacteria and their significance in the article, "Ion Antagonism in Bacteria as Related to Antimetabolites," *Ann. N.Y. Acad. Sci.*, 52: 1249–1259 (1950). A useful account of the photosynthetic bacteria is given by S.R. ELSDEN, "Photosynthesis and Lithotrophic Carbon Dioxide Fixation," *The Bacteria*, vol. 3 (1962); periodic reviews of progress in our understanding of autotropic bacteria appear in the serial publication, *Annual Review of Microbiology*.

(E.E.Sn.)

Nutritional Diseases and Disorders

Nutritional diseases and disorders include (1) the effects of either an inadequate or an excessive consumption of food or of some particular type of food and (2) the inability to assimilate food or some particular type of food because of a disease process, a structural defect, or a chemical disorder. Good nutrition cannot be achieved without a satisfactory diet; the condition that develops in its absence is called dietary malnutrition. Malnutrition may also develop in a patient offered and consuming a satisfactory diet if he has certain diseases or disorders that interfere with the proper usage of his food. This type of malnutrition is referred to as malnutrition conditioned by a particular disease or disorder; *e.g.*, malnutrition conditioned by a gastrointestinal disorder or by a metabolic disorder. Most foods have to be digested, broken down to their component parts, before they can pass from the gastrointestinal tract into the substance of the body where they are used. Some of these components are used for the production of energy. Others are used for tissue growth and repair. The components of starches and fats—*i.e.*, sugars, glycerol, and fatty acids—are the chief energy sources. The principal components of proteins are amino acids, on which the first call is for tissue building or repair. If energy-yielding nutrients are in short supply, amino acids are diverted from their primary purpose and are used as energy sources, but the process is an economically wasteful one, since high-protein foods are expensive. Other products of the digestion of foods, not significant as energy sources, are required for metabolism; this group includes vitamins, minerals, and trace elements. (Trace elements are also minerals, but derive their name from the fact that only slight amounts of them are needed in the diet. They include iron, copper, manganese, and zinc.) All of these products of the digestion of foods may be referred to as nutrients. Nutrients are described as nonessential or essential according to whether the body is able or unable to synthesize them from other nutrients. Essential nutrients include many amino acids, some fatty acids, many vitamins, and some minerals and trace elements. Of some of these essential nutrients, the body carries large stores, which are exhausted only after a considerable length of time; *e.g.*, calcium from the skeleton and vitamin A from the liver. Water, obtained from both foods and drink, is one of the most essential nutrients, as is oxygen from the atmosphere. The latter, because it comes through the respiratory tract, is not usually called a nutrient (see NUTRITION).

Types of food and of nutrients

Nutrients are used by the body (1) to produce energy to maintain bodily temperature and effect internal and external work and (2) to build and maintain the bodily tissues. Ordinarily, energy is derived mainly from carbohydrates and fats. Proteins are ordinarily used mainly for tissue building and repair. Protective nutrients—*e.g.*, vitamins—are contained in most varieties of energy and protein sources. If energy-yielding foods are devoid of protective nutrients—*e.g.*, refined sugar—they are referred to as empty calories. Allowances of foods and nutrients have been recommended by various authorities for different ages, degrees of activity, and climates. These recommended allowances usually carry a substantial margin over minimum requirements to provide a safety factor for stresses of various types, including disease. In a fortunate environment, individuals may live in apparently good health on as much as a third less than the recommended allowances of many nutrients. Below that level health and efficiency must certainly be impaired, even if the effects are not immediately apparent. Customary dietary patterns in apparently healthy people in different parts of the world are vastly different; this indicates the great adaptability of the body, which possesses a number of alternate metabolic pathways.

Undernutrition, usually characterized by thinness, is due to an imbalance between calories available and those expended as energy. Overnutrition, characterized by fatness, is due to overconsumption of calories in relation to energy expenditure. Much malnutrition, as in scurvy and pellagra, is neither undernutrition nor overnutrition but, rather, a deficiency of one or more essential noncalorie nutrients. This type of malnutrition is sometimes called qualitative malnutrition, to distinguish it from quantitative malnutrition represented by undernutrition and overnutrition. Increasingly the term malnutrition is used to cover all the results of unsatisfactory provision or utilization of food, including overnutrition and obesity.

Undernutrition and qualitative malnutrition are responsible over the world as a whole for a vast amount of disease and disordered health. The problem, in the 20th century, is being enlarged by rapid increases in population. In tropical countries it is easier to produce relatively empty calories than good protective foods, and malnutrition complicates the problems of tropical parasitism, tropical disease, and the many other adverse environmental factors that may undermine growth and development during infancy, childhood, and adolescence, leading to heavy infant and child mortality and shortened life expectancy. Malnutrition represents one of the major problems of the 20th century. In theory, food production can be greatly increased, but it is questionable whether man has yet achieved the necessary international social organization either to match food production to the rate of population increase or to control the rate of that increase.

Throughout history there have been privileged groups whose resources have permitted the development of obesity. In the 20th century, western affluence has extended to the greater part of certain populations, bringing with it certain advantages (faster growth, increase in stature from generation to generation, and greater life expectancy), on the one hand and disadvantages on the other. Habitual indulgence in some foods is beginning to threaten the rising curve of life expectancy, especially through obesity or through degenerative disease of the arteries and the resultant coronary heart disease. Almost all diseases have multiple causes, some of which, especially the dietary causes, may operate over the greater part of a lifetime before manifesting their effects in middle or old age.

DEFICIENCY DISEASES

Deficiency diseases may best be considered under five main headings, according to the principal deficiency: (1) water, (2) calories, (3) vitamins, (4) proteins, and (5) minerals. There is considerable overlap between these groups.

Water. Some 65% of a normal body consists of water distributed between three compartments: (1) within the cells, (2) between the cells, and (3) within the circulatory system. A great variety of diseases disturb the water balance, but in healthy people only dehydration and overhydration result from wrong intake of water. The prominent symptom of dehydration is thirst, which develops quickly when water is unavailable. Dehydration becomes rapidly worse, and death may ensue within a day or two. Water is lost to the body by evaporation from the lungs and skin, through the urine, and through the stools, which loss can attain to serious amounts in states of diarrhea. Lung and skin losses are increased by work that causes sweating, particularly in a hot, dry climate. The kidneys usually excrete between one and two litres of urine per day (roughly one to two quarts), but when dehydration threatens they can conserve water through concentration of the urine; the specific gravity of the urine then increases and its colour darkens through concentration of pigments.

Overhydration from overdrinking is uncommon in healthy people, because the combined effects of excretion by the kidney and satiation, the opposite of thirst, are usually adequate even for the most enthusiastic drinker.

The distribution of water between the three compartments is largely determined by the electrolytes sodium and potassium. The important positively charged ion, or cation, in the interstitial fluid is sodium. Mammals, which have acquired an energy-dependent mechanism for removing sodium from the individual cells, the sodium pump, have potassium as their principal cation within the cells. (The cell membrane is permeable to sodium ions, and the sodium concentration outside the cells is stronger than within, so that one would expect sodium to pass

from the tissue fluids into the cells; its movement in the opposite direction requires energy. The process, whatever it may be, that involves this expenditure of energy is called the sodium pump.) Electrolyte and fluid balance are both maintained by the regulatory power of the sodium pump and of the kidney tubules. Upon this bare factual skeleton is built a complicated structure for medical regulation of fluid and electrolyte balance in disease; upset of the delicate electrolyte and fluid balance may be aggravated by deficient or excess intake of water.

In addition to sodium and potassium, sea water (the original source of animal life) contains the elements calcium, magnesium, sulfur, nitrogen, carbon, iodine, iron, and phosphorus. All of these can be detected in the human body, and many play important parts in metabolism (see below). There are also a number of elements present in the surface of the earth that are found in the human body in minimal quantities. Some of these have important known functions; others may be inert contaminants derived from the food consumed.

Calories. Energy, ultimately dependent upon solar energy, is taken into the body through food. It is expended in mechanical work, in the work of the internal organs, in maintaining body temperature, and in promoting growth. In adults, excess energy intake over energy expenditure leads to overnutrition and eventually to obesity, and the reverse leads to undernutrition with loss of weight. In children, moderate undernutrition, instead of leading to weight loss, may simply impair the rate of growth. Energy intake is measured in terms of kilocalories (abbreviated kcal) or simply as calories. In the near future, by international convention, the term calorie will be replaced by the term joule with an agreed conversion factor. The principal food sources of energy are starches and fats.

If the quantity of a properly balanced diet is insufficient, the result is simple undernutrition. In practice this balanced quantitative deficiency seldom occurs, except (1) during infancy, either with inadequate breast feeding or with insufficient although properly balanced bottle formulas; and (2) with voluntary limitation of intake among adolescents and adults in well-nourished communities. In adolescents and in adults who are voluntarily restricting their intake of a well-balanced diet, often for the cure or prevention of obesity, loss of weight occurs and energy and vitality may be impaired. In some cases neurotic behaviour develops, and in anorexia nervosa (see below), obsessions or delusions may be present.

Apart from these exceptions, undernutrition is almost invariably associated with an unbalanced diet. According to the particular pattern of imbalance, a variety of manifestations of qualitative malnutrition will be associated with evident undernutrition. Extreme examples may be seen in prolonged famines, during serious civil disorder, and in concentration camps. Of more common occurrence, except in situations of extreme crisis, is chronic malnutrition due to deficiency of one or more body-building or protective nutrients in persons who are not obviously undernourished. This type of malnutrition is discussed in the following sections.

Vitamins, antivitamins, and provitamins

Vitamins. For normal nutrition and metabolism the body requires certain organic substances, called vitamins, that it cannot make for itself, at least in sufficient quantity. Because they are required in such small quantity, the vitamins do not contribute significantly to the energy needs of the body. In their absence, the carbohydrates, fats, and proteins required for energy production, tissue building, and tissue repair cannot be properly metabolized.

The relation between the quantity of a vitamin obtained in the diet and its utilization and requirement by the body is not direct. Some vitamins, such as nicotinic acid, may occur partly in nonavailable form and may not be absorbed. Substances called antivitamins, present in some natural foods, may block the vitamins at their site of action. Some substances, called provitamins, are not themselves effective but may be converted into vitamins in the body; *e.g.*, beta-carotene is converted to vitamin A, and the amino acid tryptophan can be converted to niacin. Some vitamins, such as vitamin K, are synthesized in limited amount by bacteria present in the intestines.

Each of the vitamin-deficiency diseases is commonly attributed to a deficiency of a single vitamin, and many have been produced in animals and in human volunteers by diets apparently complete apart from the vitamin under trial. In human experience, these supposed single-vitamin diseases usually arise from unhealthy and unbalanced diets that are also deficient in other vitamins or in nonvitamin nutrients. In these complex nutrient-deficiency states, the nutrient whose deficiency plays the largest part in the causation of symptoms is often referred to as the most limiting nutrient; *e.g.*, niacin is the most limiting nutrient in pellagra, but there are often deficiencies of other B vitamins, and protein deficiency is commonly present.

Scurvy

A good example of a single-nutrient disease is scurvy, the scourge of mariners in the days of long ocean voyages. Its devastating effects have been immortalized in Samuel Coleridge's "The Rime of the Ancient Mariner." In 1497, the Portuguese navigator Vasco da Gama lost 100 of his 160 men in rounding the Cape of Good Hope. Early explorers suspected that the juice of the leaves of various plants had antiscorbutic properties. In 1753, a Scottish naval surgeon, James Lind, established the concept of deficiency diseases by showing that scurvy can be both cured and prevented by the use of orange and lemon juice. Captain James Cook used this knowledge in his exploration of Australia and New Zealand. Scurvy is an eminently preventable disease, and it is now seen mainly at the extremes of life and in institutions where there is willful disregard of the principles of nutrition. Infants reared on reconstituted dried milks suffer from scurvy unless ascorbic acid is added to the milk powder or orange juice is administered daily. Elderly widows and widowers not infrequently develop scurvy from negligent attempts to simplify their own catering and cooking. At an early stage scurvy results in swelling and bleeding of the gums and bleeding into the skin. Bleeding into the muscles and under the periosteum, the membrane that covers the bones, in small children, may also occur. Each of these manifestations is fairly easily recognized by those who are nutritionally aware. Anemia develops and resistance to infection is gravely lowered. It is probable that, in the classical epidemics among mariners and in prisons, the diet was unsatisfactory in many ways and that other deficiency states contributed to the more serious complications and to death; but experiments on human volunteers fed diets lacking in ascorbic acid but otherwise satisfactory have shown that scurvy can be and often is a single-nutrient-deficiency disease. Apart from scurvy and iron-deficiency anemia, described below, few single-nutrient-deficiency states are encountered naturally.

Fat-soluble vitamins

Vitamins A, D, E, and K are grouped together as the fat-soluble vitamins, although their chemical structure and their actions are quite distinct. They have certain features in common: (1) animal products in Western diets furnish a substantial part of the ordinary human requirement; (2) because of their fat solubility, they are not easily excreted and can accumulate to a dangerous degree if supplied too richly in the diet. Moreover, these vitamins tend to be stored in the liver of fish and marine mammals; thus vitamin A intoxication has occurred from the consumption by human beings of the liver of polar bears, and cod-liver oil is used in the treatment of rickets, a disease due to vitamin D deficiency.

The earliest and most widely prevalent manifestation of vitamin A deficiency is night blindness, because the retinal pigment rhodopsin (visual purple) requires retinol (vitamin A) for its functioning. Retinol deficiency manifests itself as inability of the eyes to adapt in dim light. Moreover, epithelial cells throughout the body—the various cells covering or lining the body and its organs—degenerate in vitamin A deficiency. In the eyes, the conjunctiva and cornea undergo changes that lead eventually to breakdown and complete blindness. Other epithelial cells including the skin and the mucous lining of the respiratory and urinary tracts degenerate, with characteristic results such as hardening and drying of the skin,

respiratory infections, and kidney stones. (All of these changes can be simulated to more or less degree by diseases that have nothing to do with vitamin A deficiency.) Vitamin A deficiency is an important cause of blindness, particularly in the first five years of life.

The vitamin in the diet is obtained from milk, butter, cheese, egg yolk, liver, and some of the fatty fish. The body can also make vitamin A from beta-carotene and some other plant pigments furnished by yellow and red fruits and vegetables and many green leaves. Absorption and conversion of carotene to retinol is variable but is averaged at one-sixth for dietetic calculations. It is sparsely supplied in starchy diets, which, after weaning, lead to widespread protein-calorie malnutrition. There is not much loss of retinol or carotene in ordinary cooking, but exposure of curative fish oils (*e.g.*, cod-liver oil) to bright sunlight and of vegetables and fruits to drying in the sunlight leads to considerable losses.

Vitamin D deficiencies

Rickets and osteomalacia, the adult form of rickets, are the important manifestations of vitamin D deficiency. Their differences are due largely to the fact that the skeleton reacts in one way to vitamin D deficiency in the growing phase and in another way in adult life.

Vitamin D, like vitamin A, is fat soluble; it is obtained in the diet principally from dairy products and from fatty fish and their oils and livers. Fish obtain the vitamin from plankton that has been exposed to sunlight at the surface of the sea. The human body is not entirely dependent upon its consumption of vitamin D because certain substances in the skin can be converted to vitamin D by exposure to sunlight. The dietary requirement for vitamin D is therefore dependent in part upon exposure to sunlight. Much of the pioneer work on rickets was done in Scotland, where the smog of the industrial revolution clouded the limited sunlight in tenement districts. The condition was largely abolished by routine administration of cod-liver oil to children. In tropical and subtropical climates, sunlight plays an important role in the availability of vitamin D, although exposure to sunlight in these areas may be limited by social customs such as swaddling of infants and the purdah system of veiling and covering women of childbearing age.

Vitamin D is concerned with the absorption of calcium and, secondarily, of phosphorus from the gastrointestinal tract. This process is also affected by the ratio of calcium to phorphorus and the amount of phytate in the diet. (Phytate, present in cereals, interferes with the availability of calcium.) Internally, vitamin D in part controls the metabolism of calcium and phosphorus in relation to growth, development, and maintenance of the bony skeleton. Relationships particularly with the parathyroid hormone are complex, and many aspects are still obscure. The most evident effect of vitamin D deficiency is interference with bone growth and bone maintenance. In children who have learned to walk, the growing ends of bones are enlarged and misshapen, with many characteristic skeletal deformities, including knock-knee, bossing of the frontal bone of the skull, and a line of beadlike nodules (the "rachitic rosary") at the junctures of the ribs and the breastbone. Softening of the skull bones (craniotabes), which may result from a number of causes, is commonly attributable to vitamin D deficiency when it occurs in the first year of life; the frequency is still controversial. The bony deformities are always slow to resolve, and some of them may remain permanent. A common permanent adult deformity in the past was contracted pelvis leading to difficulties in childbirth.

Rickets has been largely abolished in certain countries by routine prophylactic administration of vitamin D to infants and young children. During World War II there was compulsory fortification of infant formulas, margarine, and certain other foods with vitamin D. Subsequently, food manufacturers began to add vitamin D to other foods for infants and children, and suspicions about widespread overdosage were aroused. This aspect of food fortification has to be carefully supervised. In developing countries, particularly within the tropics, prevention of vitamin D deficiency depends upon health education, which discourages swaddling of children and purdah in women. If there is heavy cloud overcast, the prophylactic administration of vitamin D may be required in some form suitable to local conditions.

Vitamin E

Vitamin E covers a mixture of vitamins, called tocopherols, of which the important alpha variety was isolated from wheat-germ oil and has been synthesized. Early experiments on vitamin-E-deficient rats led to abortion in the female and sterility in the male. The anti-oxidant activity of the vitamin is related in some way to the health of muscles, skin, and perhaps all tissues. The tocopherols have been found in virtually every foodstuff used by man. The richest sources are eggs and oils expressed from some cereals, beans, and peas. Western diets are estimated to contain some 20 milligrams of vitamin E, which is thought to be an abundance.

Vitamin K

Vitamin K is necessary for the normal formation of prothrombin, one of the coagulation factors, by the liver. Vitamin K deficiency has never been clearly demonstrated in adults, but since the anticoagulants used in treatment of coronary thrombosis, and in other disorders involving blood clotting, act by inhibiting the production of prothrombin by the liver, vitamin K is used to counteract overdosage. A fat-soluble vitamin, it is often used prophylactically if there is disordered absorption of fats, as in obstruction of the bile ducts. It is thought that the adult freedom from vitamin K deficiency is achieved by synthesis of vitamin K by the bacteria of the large bowel.

The newborn infant, who has a sterile intestinal tract and consumes mostly milk, a poor source of the vitamin, suffers easily from prothrombin deficiency and may show a tendency to bleed in the first week of life. This deficiency is treated with vitamin K, but it is almost certainly complicated by other deficiencies that interfere with the early production of prothrombin by the liver of the newborn babe.

Vitamin B complex; vitamin B_{12}, folic acid, and anemia

Vitamins B_1 and B_{12} are well-identified members of a group of water-soluble vitamins that tend to have the same food sources. Some of the members have been eliminated with advancing knowledge (*e.g.*, vitamins B_3 and B_4 are probably the same as B_1). Vitamin B_1 is now identified as thiamine; nicotinic acid (niacin) was previously known as vitamin B_2 or the pellagra-preventing factor; riboflavin was separated from the same complex; vitamin B_6 is now identified as pantothenic acid and biotin; vitamin B_{12}, identified as cyanocobalamin, has important relationships with folic acid.

Deficiencies of the vitamin B complex lead to many disorders of function and structure. Deficiency of one member is likely to be complicated by deficiency of other members of the complex. These complex deficiencies tend, in general, to occur when diets are excessively dependent upon starches as staple sources of calories for energy production. This dependence is often magnified by the consumption of refined cereals (*e.g.*, white flour, white rice), in which the vitamin-carrying germ and husk are eliminated, and by growing consumption of refined sugar.

The discovery that the previously fatal disease Addisonian pernicious anemia could be cured by the consumption of large amounts of semicooked liver led ultimately to the identification of the effective factor in liver as cyanocobalamin (vitamin B_{12}). Pernicious anemia is due to deficiency of cyanocobalamin, the failure resulting not from a deficiency in the diet but from a familial or acquired deficiency of the small intestine's ability to absorb the vitamin. Dietary deficiency of cyanocobalamin is rare except in those who consume no animal flesh or products at all, but a similar type of anemia is widely encountered as a manifestation of dietary deficiency of folate coenzymes (substances that enhance or make possible the action of enzymes), usually referred to as folic acid, folacin, or folates. Both cyanocobalamin and folacin are needed for formation of healthy red blood cells, but their actions are in other respects different.

In some parts of the developing world, folate-deficiency anemia is a serious public health problem. Even in the Western world, there is evidence that many pregnant women suffer from undetected folate deficiency. Programs for routine administration of folic acid as well as

of iron are under consideration for improving the health of women during the childbearing period. During pregnancy the fetus has to be supplied with these and other nutrients; stores are often drawn from a woman already marginally depleted. The whole subject of dietary requirement and sources in respect of cyanocobalamin and folic acid, and the contribution of the intestinal bacterial flora to folate uptake, is still somewhat obscure.

Beriberi

Beriberi and pellagra are well-known diseases, usually regarded as single-nutrient-deficiency diseases, that result from too great reliance on certain cereals as sources of energy. Pellagra occurs widely in populations depending mainly on maize, while beriberi is endemic among populations that depend upon excessively milled rice. Both diseases can be prevented by addition of small quantities of protein-rich foods that contain the missing principles. Pellagra is due to deficiency of the vitamin niacin in available form and of the amino acid tryptophan, from which the body can make niacin. Maize is deficient in both nutrients. Natural rice is a good source of energy, but when it is highly milled, the vitamin thiamine (B_1) is removed with the husk and the germ of the grain. The milling of rice improves its keeping qualities, and thiamine can be restored in synthetic form or can be provided in sauces or gravy. The keeping quality of rice can be improved by parboiling, without removal of the husk and germ.

In the 1870s, after beriberi had become a devastating disease of sailors in the Japanese navy, the Japanese naval diet was revised; part of the polished rice was replaced by barley, and evaporated milk and meat were introduced. The success of these changes in eradicating the disease was attributed to the beneficial effect of a more liberal protein allowance, and the important part played by polished rice in the unrevised diet was missed.

In the absence of the vitamin thiamine, carbohydrates are incompletely metabolized, and poisonous lactic and pyruvic acids accumulate in the tissues. This deficiency state affects the nerves, heart, and circulation. According to the effect on these different systems, beriberi may be "dry" or "wet." The former is due to degeneration of the long nerves to the extremities, and the latter to edema, the presence of abnormal amounts of fluid in the tissues, resulting from failure of the heart and circulation. Mixed types are found. Experimentally, a pigeon paralyzed from a diet of polished rice can be restored to activity in a few minutes by an injection of thiamine. In moderate cases of human beriberi, thiamine, taken orally or injected, causes rapid improvement in a few days. In longstanding cases, possibly because of associated deficiencies, cure is much slower; indeed, in severe cases, part of the degeneration may be irreversible. Beriberi has been described in infants suckling from mothers suffering from thiamine deficiency. In Western countries, thiamine deficiency is now encountered almost solely in chronic alcoholics.

Pellagra

Pellagra, from a word meaning "rough" or "sour" skin, was first described in Spain in the 18th century, after the introduction of maize from the Americas. Preventable, it still occurs widely in the maize belts of the world, except that of the southern United States, which has been freed of pellagra by the economic advancement that occurred during World War II. Investigation of the relationship between pellagra and maize provides a fascinating story in the history of human nutrition research, but there is still an element of uncertainty. The three classical theories of a toxin in maize, of deficiency of protein, and of vitamin deficiency have been partly resolved by the knowledge that the amino acid tryptophan, relatively lacking in the protein of maize, is the precursor of the vitamin niacin and that the niacin that does exist in maize and in some other cereals is held in a bound or unavailable form. A more recent theory, that an excess of the amino acid leucine leads to amino acid imbalance, has still to be finally evaluated. The dermatitis of pellagra results from sensitization of the skin to sunlight and hence has a characteristic distribution. In some of the maize belts of the world today, "rough skin" with light-sensitive distribution is the only sign of widespread pellagra. At a more severe stage, smooth sore tongue, gastrointestinal upset, and diarrhea occur in epidemics. In severe cases mental disturbance and psychosis appear. The more severe features of the disease are easily curable with the administration of niacin. Complete restoration to health almost certainly requires liberalization of the diet with more varied sources of protein. This enhanced diet is apparently what has abolished the condition in the southern United States. As in the case of beriberi, pellagra in developed regions is seen particularly in chronic alcoholics. Hartnup disease, a rare disease resembling pellagra, is due to an inborn error of metabolism.

The foregoing discussions of beriberi and pellagra concern deficiencies in two important staple cereals upon which a large part of the world is dependent for its energy production. Equivalent deficiencies exist, or can be induced by modern processing, in other important staple cereals such as wheat, barley, and sorghum. Undue dependence upon these processed cereals or upon root staples such as potatoes, yams, and cassava may produce not only vitamin deficiences but also deficiencies of minerals and of amino acids, the building blocks of proteins.

Protein deficiencies; kwashiorkor

Proteins. Proteins are required in children for tissue building and in adults for tissue repair. One of the principal reasons for the widespread occurrence of protein deficiency in the world is excessive dependence upon starchy foodstuffs derived from cereals and root staples. If, on the other hand, calories from carbohydrates and fats are in short supply, protein is wastefully used to supply energy. Protein requirement and recommended allowances are therefore dependent in large degree upon the adequacy of supply of energy from carbohydrates and fats (see PROTEIN).

All proteins consist of a mixture of amino acids. There are more than twenty such acids, slightly under half of which cannot be synthesized by man and are known as essential amino acids. Every protein is unique in pattern, and its quality is almost more important than its quantity. Most good sources of protein are also carriers of important vitamins and minerals.

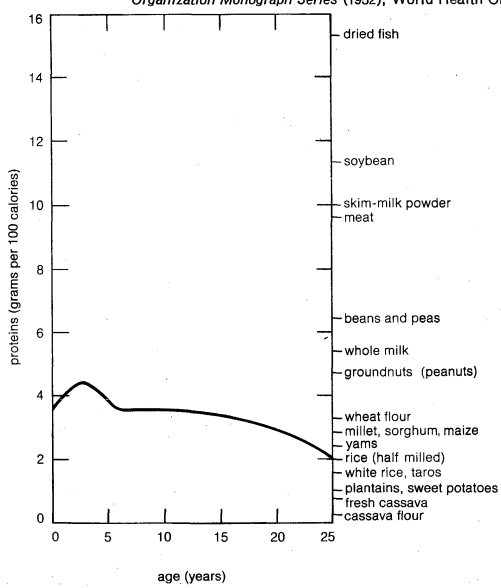
Up to the 1940s the attention of nutritional scientists had largely been applied to vitamin-deficiency diseases. It was not generally recognized that protein deficiency might be severe and widespread without gross undernutrition. An English physician, Cecily Delphine Williams, in 1932 described, in what is now Ghana, a disease in young children that she suspected to be due to deficiency of some amino acids and that she called kwashiorkor from a local tribal word. To some extent before and to a considerable extent after her description workers in underprivileged countries had described similar types of disease under a variety of names. As a result of inquiries instituted by the World Health Organization (WHO), it was soon evident that all of these diseases were due, in children recently weaned from the breast, to reliance on starchy foods that were relatively low in proteins, often also defective in their amino acid composition. The name kwashiorkor became internationally accepted. In the typical case, the child may appear not to be much undernourished, but the false appearance is contributed to by edema. It was later realized that every gradation exists between classical (edematous) kwashiorkor and the more severe undernutrition called marasmus, with a corresponding decrease in weight for age. The term protein-calorie malnutrition was coined to cover the full spectrum. The variable common characteristics of this spectrum are growth failure, peevish apathy, edema, fatty liver, changes in hair and skin, diarrhea, and death. Protein-calorie malnutrition is always due to failure to include adequate quantity or quality of amino acids, in the form of good protein, in the diet of children recently weaned from the breast or still in the preschool period, at which age the protein requirement is relatively high. It has been observed that cure of kwashiorkor can be initiated by the use of synthetic formulas containing essential amino acids, glucose for calories, an appropriate salt mixture, and water. Such initiation of cure can be brought about even in the absence of all vitamins from the formula. Since the protein deficiency of protein-cal-

orie malnutrition is always complicated by varying patterns of calorie, vitamin, and mineral deficiency, a full diet is naturally required for consolidation of cure and for return to full health. As the supply of calories, usually from cereal or root carbohydrates, including plantains, becomes less and less satisfactory for the needs of energy and growth, the spectrum of protein-calorie malnutrition passes from classical kwashiorkor through marasmic kwashiorkor to infantile marasmus. The same gradation occurs as breast feeding becomes less satisfactory in quantity and duration; when classical kwashiorkor appears, it is at an interval of three to six months after the late weaning of a well-developed baby. Marasmus is often seen at three months in the baby of a mother who has gone back early to work and left her child in the care of another.

Equivalent syndromes may be present at all ages; it is now recognized that severe undernutrition almost always includes some protein deficiency. When starved of calories, the body wastefully uses the amino acids of proteins for energy production instead of for tissue building or repair. The need for protein is aggravated in tropical countries by parasitic infestation. In all regions gastrointestinal infection, which promotes severe diarrhea in an intestine already wasted by protein-calorie deficiency, often leads ultimately to death.

The role of protein quantity is well illustrated in the Figure, which shows the amounts of protein provided by the common foodstuffs in quantities yielding 100 kilocalories of energy. Protein quality is more complex and depends in part on the pattern of constituent amino acids

From J. Brock and M. Autret, "Kwashiorkor in Africa," *World Health Organization Monograph Series* (1952); World Health Organization



Daily protein requirement per 100 calories, according to age, and yields of protein provided by common foods, in quantities yielding 100 calories.

and on the vitamins also present in the protein-containing foodstuff. Throughout the developing and underprivileged areas of the world, protein-calorie malnutrition is by far the most severe and widespread human dietary deficiency. In many of these regions almost every child, unless he happens to come from a privileged home, passes through a phase of retarded growth and development during the second and third years of life; *i.e.*, after weaning. When dietary protein, better in quality or in quantity, is consumed as the child competes with the rest of the members around the family cooking pot, growth and general improvement in health occur. The lost ground is at least partially made up by this growth spurt. In recent years, there has been conjecture about the possibility that the body is permanently damaged during the period of retardation of growth and development. It has been thought, for example, that this early deprivation might be partially responsible for widespread adult cirrhosis of the liver in areas where kwashiorkor is endemic. Recent protein-deprivation studies in a variety of animals strongly suggest that there is a vulnerable period between birth and weaning when protein deprivation may permanently impair brain development. Such a possibility could be of vast economic and social importance. It appears that this vulnerable time, if it occurs in human beings, is between birth and weaning—*i.e.*, during the first six months of life —rather than after weaning. On this premise, permanent mental retardation would be more likely to occur in infantile marasmus than in classical kwashiorkor. Since both clinical conditions are part of the spectrum of protein-calorie malnutrition, and since man is a long-lived animal, it is difficult to draw hard and fast conclusions.

Possible permanent after-effects of protein-calorie malnutrition

Fats and carbohydrates. Most human diets contain between 10 and 15 percent of protein. The balance is carbohydrates and fats. In general, carbohydrates may vary from as high as 90 percent among poor persons to as low as 40 percent among the wealthy; the balance is made up from fats, which may amount to 45 percent among the wealthy. Emphasis has been placed upon the great importance of protein quality because of the selective need for essential amino acids in growth and tissue repair. The quality of carbohydrates and fats as sources of energy is less important, and the body is adaptable to a relatively high or low ratio of carbohydrates to fat as sources of energy and to great differences in the types of carbohydrate and fatty foods.

In the discussion of protein-calorie malnutrition above, it was pointed out that adequate supplies of energy allow economy of protein for tissue building or repair and prevent the wasteful use of protein as a source of energy. This process of economy has a limit, set by the quality of the protein. If energy sources are largely provided by carbohydrates from root vegetables and cereals, protein-calorie malnutrition easily occurs if the quality of protein is limited by relative deficiency of one or more essential amino acids; *e.g.*, lysine and tryptophan in maize. Another important factor with respect to carbohydrate foods is the tendency in the last few decades toward milling and refining, which eliminate important vitamins and produce so-called empty calories, which yield energy but contain little of the protective nutrients present in unrefined food. White bread, white rice, and white table sugar are examples of refined foods. Modern agricultural techniques using selective breeding are able to produce cereals and other carbohydrate foods with high quantities of protein and better amino acid patterns.

Fats also vary greatly in their composition, and some of them contain more than others of certain unsaturated essential fatty acids such as linoleic, which the body cannot synthesize. (Fats are called saturated if they cannot absorb more hydrogen.) Fats of animal origin and oils that have been "hardened," as in the pastry fats, are saturated—they contain less of the unsaturated essential fatty acids. In some way (see below), the relative consumption of unsaturated and of saturated fats has a bearing on the problem of the human disease atheroma and particularly of its special presentation as coronary heart disease.

Fats are also important sources of the fat-soluble vitamins. The view of nutritionists that proteins should make up 10–15 percent of the diet, and carbohydrates 55–65 percent, does not take adequate cognizance of the importance of quality in all three of these main components of the diet's energy sources. Moreover, it does not give sufficient credit to the remarkable adaptability of the human body to varying diets (see CARBOHYDRATE).

Minerals. A normal adult has 1.2 kilograms (about 2.6 pounds) of calcium in his body; of this, all but a fraction is in the bony skeleton, where, mostly in combination with phosphate, it provides the hard structure to a protein cellular matrix. In childhood and adolescence, adequate dietary intake of calcium is essential for healthy growth; in adult life the skeleton acts as a reserve store of calcium. That the skeleton is not inert but is potentially in active exchange, even in the adult, is shown by the fact that lead accumulating slowly in the skeleton of lead workers can be mobilized rapidly by induced metabolic changes and subsequently removed. That this exchange is

Calcium

incomplete is shown by the damaging effect of radioactive strontium from atmospheric fallout; this becomes relatively fixed with ultimate damaging effect on bone. Disturbances of the absorption, metabolism, and excretion of calcium may result from a variety of endocrine disturbances and other diseases.

Extra quantities of calcium are needed throughout the period of growth and during pregnancy and lactation, when the mother has to supply the calcium needs of her fetus or infant. An important source of calcium for growing children is milk, whole or skimmed or as cheese. The rest of the calcium is derived from a great variety of foods. The calcium in a half-pint of milk is enough to meet the needs of all ages, except at puberty and during pregnancy and lactation, when one pint is an adequate supplement to a good diet. There has been considerable dispute about recommended allowances of calcium, and undoubtedly the tendency has been to set these too high. Allowance has not been made by nutritionists for the versatility of the body's processes of adaptation. Excessive recommendations of calcium lead inevitably to overconsumption of cows' milk, which may contribute to excessive intakes of saturated fats.

Magnesium; iron

Magnesium, like calcium, is stored in the skeleton; it appears to have some overlap with calcium in metabolism. It is also, next to potassium, the predominant metallic cation in living cells. It is an essential part of many enzyme systems. It is also an essential component of chlorophyll, the respiratory pigment of plants. Because of its reserve in the skeleton and its widespread presence in diets containing vegetables, there is probably never a dietary deficiency of magnesium.

Human disease resulting from dietary deficiency of phosphorus, sulfur, nitrogen, or carbon is unknown, presumably because these substances are found in a wide range of foods. Phosphorus deficiency has been reported in cattle on certain pastures deficient in phosphate.

Iron is a nutrient of great interest and importance. It is an essential component of hemoglobin and myoglobin, the oxygen-transporting proteins of the blood and of the muscles, respectively; deficiency might be expected to have serious effects. Its quantity in most daily diets is rather small (10–20 milligrams). The intestinal tract absorbs iron from the diet by an active process that is controlled in such a way that when iron stores are satisfied, most of the dietary iron is excreted in the feces, unless the dietary iron intake is so excessive that the control mechanism is inadequate. The normal daily wear-and-tear destruction of circulating red cells releases hemoglobin to the extent of 20 milligrams of iron per day. This is recycled, and most of it is incorporated in new hemoglobin; otherwise the body would rapidly run short. By the same token, the body's iron reserves are seriously taxed by increased demands on the blood supply in pregnancy and by loss of blood as by menstruation or from bleeding ulcers. If acute and severe, blood losses are usually corrected by blood transfusion. The long-continued loss of blood often leads to chronic iron deficiency, of which anemia of the type called hypochromic is the commonest manifestation. Iron deficiency, although easily corrected by iron preparations administered orally, is still probably the commonest single-nutrient disease in Western culture. It is probably responsible for considerable minor ill health among women who bear many children.

Trace element deficiencies

Early work in the mid-19th century on plant and animal requirements of macromineral elements—the elements such as sodium, potassium, and calcium that are needed in substantial amounts—was handicapped by the absence of chemically pure preparations. What are now called the micromineral or trace mineral elements often contaminated experimental preparations of the macromineral elements with which they are frequently associated in the soil. Since they have been chemically separated and identified, a number of these trace mineral elements have proved to be essential constituents of vitamins and hormones. The term trace element is now applied to elements that occur as not more than 1 part in 20,000 in the body. For centuries it has been known that goitre (enlargement of the thyroid gland) is endemic throughout the world in regions such as Switzerland that are heavily mountainous. Burnt sponge had been used as a cure for goitre since the 13th century, and it was suggested in Switzerland in 1820 that the then recently discovered element iodine might be the curative agent contained in burnt sponge. In 1896 it was demonstrated that the thyroid gland is rich in iodine. Thyroxine, an iodine-containing amino acid, was isolated in 1915 and synthesized in 1927 as the hormone of the thyroid gland. In the goitre belts of the world, deficiency of iodine prevents the formation of thyroxine, and the thyroid gland becomes enlarged and nodular in its abortive attempts to produce its hormone. The geographical distribution of endemic goitre is probably explained by leaching of the very soluble iodine out of the soil of mountainous areas into the sea where, *inter alia*, it is incorporated in sponges. Endemic goitre can now be prevented by the addition of iodine, as iodide or iodate, to table salt.

Fluorine

The nutritional significance of fluorine is a more recent discovery. Nineteenth-century chemists noted its invariable presence in bones and teeth. In the last quarter of the 19th century, fluorine was recommended for administration to pregnant women in the interest of sound teeth, and in 1892 it was suggested that the high incidence of caries (tooth decay) in England was due to deficiency of fluoride in the diet. In the present century, abundant evidence has established that in geographical regions where the water supply contains less than 0.5 part per million of fluorine, the addition of small quantities to the water or the local application of fluorine to the teeth of growing children greatly reduces the incidence of caries. In other parts of the world where the level of fluorine in the natural water is extremely high, teeth become affected by a mottling, and parts of the skeleton become dense and are affected by rheumatic and degenerative changes.

Scientific enquiry suggests most strongly that the easiest way of preventing fluorine deficiency tooth decay is to add fluorine to the drinking water to bring its fluorine content up to 0.5–1.0 part per million, at which level there is no risk to the body. Exposure to levels above ten parts per million for several years is required before mottling of the teeth appears; fluorosis of the skeleton is seen only after decades of exposure. In the absence of fluoridation of public water supplies, physicians and dentists can successfully prevent early caries by administering appropriate quantities of fluorine to small children or by ensuring application of fluorine to the teeth at regular intervals.

Copper and other metallic elements

Copper appears to be necessary to aid the incorporation of iron into hemoglobin, the oxygen-carrying protein of the red blood cells. Cobalt is a component of vitamin B_{12} (cyanocobalomin), essential for the formation of healthy red blood cells. Zinc deficiency is suspected of contributing to a syndrome (encountered in Iran) of retarded growth and sexual development. Zinc deficiency in animals may produce dermatitis and gastrointestinal upset. Manganese deficiency has not been recognized in man; in cattle grazing on manganese-deficient peat pastures, poor growth and reproduction and sometimes anemia, bone changes, and disturbances of the central nervous system may be observed. Manganese poisoning, affecting the central nervous system, occurs in workers in manganese mines. Selenium deficiency produces disease in animals but has not yet been described in man. The possible relations of deficiency of molybdenum, nickel, aluminum, and silicon are being investigated.

OVERCONSUMPTION

Obesity. By far the commonest effect of continued overconsumption of calories is obesity, a state of excess accumulation of fat in and on the body. It is often wrongly defined as a state in which weight is in excess of a 10 percent margin allowed over standard weight tables for age, height, sex, and race. Fat on the body is easily seen and measured. It is assumed that if there is excess fat in the subcutaneous tissue, there is probably also excess fat inside the body. More accurately, the amount of fat in the body can be calculated from the specific gravity of the total body, which in turn is measurable by comparison of weight underwater with weight in air. The more fat there

is, the lower is the body specific gravity. To describe weight in excess of 10 percent above standard weight as obesity is misleading, because some persons who have inherited a large frame and bulky muscles may be 20 percent or more in excess of standard weight without being obese. Nevertheless, weight is a valuable screening measure for obesity. After middle age, the specific gravity of the body may fall with little or no change in weight or skin-fold thickness. This results from disuse atrophy of the muscles and from replacement of muscle with fat and does not amount to obesity. This age change can be delayed by exercise but is not necessarily unhealthy.

Theories concerning causes of obesity

Obesity is becoming more and more of a problem in developed countries and among privileged communities. In simplest terms it is caused by consistent consumption of more calories than are required to meet the energy expenditure of that particular person, but this simple explanation conceals a great deal of difficult theory about causation. It is an impression, and probably a fact, that some persons put on fat more readily than others. This may have a simple double explanation: namely, that these people are more interested in food or that they expend less energy—the guzzling and lazy theory. This latter theory probably accounts for obesity in only a small minority. Overeating may be a habit culturally imposed by family custom, may be an occupational hazard (*i.e.*, of the pastry cook), may be induced by institutional dietetics (as in schools and tuberculosis sanatoriums), or may be an expression of boredom or emotional frustration. Underexpenditure of energy may be determined by cultural family environment, city office life, or temperament. It has been shown that obese adolescents at summer holiday camps expend less energy than their leaner friends even when engaged in active games such as tennis and swimming. Some persons may consume too many calories through ignorance of the high calorie content of alcoholic liquors. There is evidence that nibbling may contribute to obesity, even when the total calories consumed in the day are the same as when eating is confined to regular meals.

In summary, many aspects of the causation of obesity can be explained in terms of disproportion between calorie consumption and energy expenditure. Other factors may also play a role: it is frequently suggested that obese subjects may have a metabolic trend toward easy storage of fat, and a large number of differences have been demonstrated between obese and nonobese subjects in respect of hormone secretion and metabolism. The reasons for most of the metabolic and hormonal differences are obscure, since the basal metabolic rate is not lower than the normal range, and obesity due to identifiable endocrine diseases (*e.g.*, diseases of the thyroid and pituitary) constitutes a small minority of cases. Endocrine changes at puberty, during pregnancy, and at the menopause may contribute to obesity at these stages of life, but their effects can be obviated by appropriate adjustment of calorie consumption to energy expenditure. One theory for obesity, that some defect in nervous control may lead to abnormal deposition of fat in certain fat depots, called "greedy fat depots," originated from observation of certain rare diseases (lipodystrophies) in which fat is laid down in one part and disappears from another part of the body. This suggests control of fat deposition and fat mobilization by segmental nerves. There is abundant evidence that depot fat is in a continuous state of flux, or turnover, and is not an inert store.

Effects of obesity

If, as is evident from the discussion thus far, a complete and certain explanation of the causes of obesity cannot be given in the majority of instances, the effects of obesity can be stated with some certainty. Obese persons have a reduced life expectancy, and their life insurance premiums are greatly increased. They suffer disproportionately from a number of diseases and disabilities. The tendency toward obesity and susceptibility to these diseases may both be inherited, or either may cause the other. Diabetes of the middle-aged is strongly associated with obesity, and there is association of obesity in females with gallbladder disease. Contrary to general impression, coronary heart disease and obesity are not causally related, but obesity should be firmly treated to reduce the load on damaged hearts. Obesity is a mechanical load on the lower spine and the major weight-bearing joints, the hips, the knees, and the ankles. Reduction of the weight of obese persons may mitigate the middle-aged tendency to degenerative diseases of the joints and muscular rheumatism. Obesity of the trunk may interfere seriously with breathing and may contribute to pulmonary heart disease. Obesity is associated with proneness to accident. Finally, there is considerable association, both as cause and effect, between obesity and psycho-emotional disturbance. Many obese persons are jovial, but some suffer a considerable sense of inferiority. This is particularly evident among girls and young women in Western culture, where slimness is the vogue. Misdirected attempts at weight reduction may lead to deficiency disease and debility or, in a few persons, may go on to self-imposed starvation (anorexia nervosa). All in all, improved expectation of life, health, and happiness can result from the prevention of obesity or from effective and maintained weight reduction in obesity.

For obese people, the solution is simple but difficult: a reduction in calorie intake and an increase in expenditure of energy. To achieve reduction of calorie intake and increased energy expenditure is comparatively easy over a month or two; to maintain this altered balance over many years is extremely difficult. For many obese persons the craving to eat more is sometimes as strong as that for alcohol in the alcoholic and for a cigarette in the chain smoker. Vigorous games such as squash may lead to reduction of weight by several pounds, but this reduction is due simply to sweat, and weight is quickly restored when any liquid, even water, is consumed. Effective attempts to increase energy expenditure require many hours of muscular work or exercise on the feet.

In almost no field are there more fads and theories than in the dietary management of obesity. The many programs advocated for correction of obesity include complete starvation for a variable period, avoidance of either carbohydrates or protein at any one meal, avoidance of liquids at mealtimes, avoidance of nibbling and between-meal snacks, avoidance of refined sugar, high-fat diets, citrus diets, and grape diets. Few of these principles have any scientific basis except in so far as they achieve reduction of total calorie intake. Reduction of calorie intake is the only sound principle, provided that it is associated with the regular consumption of enough of those foods that provide essential nutrients: good proteins, vitamins, and minerals.

Surfeits of sugars and fats; overconsumption of vitamins and of other individual nutrients

Short-term overconsumption. Apart from the long-term effects of overconsumption of calories, short-term effects may occur from surfeit. The most obvious effect is acute alcoholic intoxication, if alcohol, in reasonable quantities, be regarded as a legitimate source of calories. Acute excess both of fats and of sugars may cause nausea and vomiting, perhaps through overloading of some metabolic functions of the liver. The same effect may come from surfeit of meat and other protein-rich foodstuffs. Gastrointestinal distension is largely mechanical and is related to the bulk of food consumed. Acute intoxication with either vitamin A or vitamin D has been reported after eating the liver of polar bears, which is extraordinarily rich in these two vitamins.

Long-term overconsumption of individual nutrients. Long-term overconsumption of individual nutrients is uncommon except in the case of fat-soluble vitamins, which are not easily excreted. Chronic intoxication both with vitamin A and with vitamin D is well known. Excess of the more common water-soluble nutrients is usually cleared by the body's excretory mechanisms. Nutritional siderosis (iron-overload) from overconsumption of iron is common throughout the African continent. This results principally from the use of iron containers for the brewing of alcohol and the processing of foods by fermentation. The daily intake of iron may be increased to ten times the normal average by such procedures. Siderosis is almost certainly harmful to the body, although the mechanisms by which it causes damage are not fully understood.

DISEASE AS A CAUSE OF MALNUTRITION

In the previous sections diseases and disorders resulting from wrong eating or malnutrition have been considered. In reverse, many diseases of themselves cause malnutrition even when a perfect diet is available. The principal mechanisms include (1) inborn errors of metabolism (see METABOLISM, DISEASES OF); (2) diseases of the gastrointestinal tract that impair digestion and absorption (see DIGESTIVE SYSTEM DISEASES); (3) kidney disease that causes retention of toxic substances that should have been removed and causes loss of nutrients such as proteins (see EXCRETORY SYSTEM DISEASES); (4) acquired disorders of metabolism that prevent absorbed nutrients from being properly handled inside the body; this group includes liver disease, disturbances of the endocrine glands, a great variety of infections, and the effects of cancer; (5) mental disease; (6) chains of abnormalities—vicious circles.

Inborn errors of metabolism. In the early 1900s a group of diseases was identified that were due to inherited defects, called inborn errors of metabolism. At present between 500 and 600 inborn errors of metabolism have been described, and the number is likely to grow. The inherited defect in these diseases is incurable, but the body does have some limited powers of adaptation by switching to metabolic pathways that are dependent upon other enzymes. The general principle of treatment is to eliminate from the diet those foods or nutrients that require for their digestion, absorption, or metabolism an enzyme that the body lacks. Alternative pathways can be encouraged, provided that essential nutrients are not involved. In the latter case little can be done. Some classical disorders of this type, all named from abnormal substances in the urine, are alkaptonuria, pentosuria, phenylketonuria, porphyria, and cystinuria. Many similar diseases are now known to be due to inborn enzyme defects in tubular reabsorption in the kidney.

Abnormal reactions to certain foods—food idiosyncrasies—often have a familial pattern. Some produce allergic disturbances, especially urticaria (hives). Unusual foods such as shellfish and exotic fruits are often offenders. Many of the reactions that remain unexplained are expected to be explained eventually in terms of minor enzyme deficiencies that are inborn errors of metabolism. These include the taste of sulfur after consuming eggs and "acidity" and "hives" after consuming certain fruits, such as tomatoes and oranges. Other reactions may be of mental origin; these often form the basis of fad diets.

Forms of sprue

Gastrointestinal tract. There are several varieties of sprue, all of which involve failure of absorption of dietary fat. The absorption of many other nutrients is primarily or secondarily impaired. The general term malabsorption syndrome is commonly used. Tropical sprue affects Europeans living and working for extended periods in the tropics. Another variety, occurring outside of tropical areas, is due to reaction against a protein called gluten that occurs in wheat and in other cereals. Another group of diseases impairs primarily the digestion, assimilation, and excretion of protein (protein-losing enteropathy).

Kidney diseases. In one group of kidney diseases (nephrotic syndromes), protein leaks out of the kidney into the urine in quantities sufficient to contribute to states of protein deficiency. In other groups, amino acids, the building blocks of protein, or glucose and other sugars may be lost. Many cases in this group are due to inborn errors of metabolism. In some acquired kidney diseases, nutrients are lost in the urine or their metabolism in the body is impaired by the presence of toxic substances—*e.g.*, urea, which upsets metabolic cycles.

Secondary metabolic diseases. Most disturbances of the endocrine glands upset metabolism to some extent and in variable ways. Diabetes mellitus causes nutritional wasting in one form, while another form of diabetes is associated with and may in part be caused by obesity. Many metabolic processes are regulated by the liver, and therefore all disorders or disturbances of the liver may affect metabolism. All infections and intoxications may do likewise, as does cancer. The mechanisms are complex and only partially understood.

Mental disease or disorder. In mental diseases, bizarre disturbances of appetite, taste, and food preference may result. A well-known example is the condition known as anorexia nervosa. Obsessive reactions against one's own appearance may develop, particularly in girls, at the time of puberty. Pubertal puppy fat may lead to fear of obesity and may link with fantasies of a sylphlike figure. In late adolescence or early adult life, the same symptoms may develop after bereavement or after disappointment in love. These persons cut down their food intake to an inadequate level. They become extraordinarily cunning in concealing their self-imposed starvation. Food is surreptitiously disposed of in the toilet or is vomited after a meal. The physical results in women include cessation of menstruation, and a peculiar growth of downy hair develops on various parts of the body, especially over the upper cheeks.

Vicious circles. Particularly in the gastrointestinal tract, deficiencies of some vitamins or protein seriously impair the functional capacity of the mucous membranes and glands that produce enzymes required for digestion and absorption. These nutritional deficiencies impair the capacity of the gastrointestinal tract to digest and absorb the nutrients contained in the defective diet. In this way a chain of effects is set up whereby nutrient deficiency causes failure of digestion and absorption, and this in turn further aggravates the developing deficiency. A worldwide manifestation of such chains of effects is seen in protein-calorie malnutrition in infants and young children, which may manifest itself as diarrhea. This diarrhea in turn aggravates the malnutrition and finally causes collapse from deficiency of water and electrolytes. Of particular interest in this context is the fact that many of the dietary deficiency diseases described earlier may be closely mimicked by diseases of metabolism. Thus rickets and pellagra may develop in persons who live on a diet that has been recommended because of inborn errors of metabolism.

HABITUAL DIETARY PATTERN AND DISEASE

Interplay of heredity and environment

The interplay of heredity and environment is clearly to be seen in diseases arising from inborn errors of metabolism. Phenylketonuria, arising from inability to digest the amino acid phenylalanine, leads, if unchecked, to mental deficiency and ultimately to death. If a diet lacking phenylalanine is instituted early enough in infancy and maintained, normal intelligence can be safeguarded. Similarly, in the disorder arising from gluten intolerance, lifelong ill health can be warded off by a diet lacking gluten. In other words, the environment (in the form of a proper diet) even in inborn errors of metabolism may be more important than heredity in shaping the ultimate constitution of the sufferer.

The two diseases just mentioned have been selected because in them the important environmental determinants are dietary. It has become increasingly apparent that there are a large number of diseases in which the ultimate breakdown is the result of the interplay over years or decades between heredity and many aspects of the environment. In some of these there is considerable evidence to suggest that the habitual dietary pattern is one of several important causative agents in the environment. Two good examples are coronary heart disease, resulting from fatty deposits in the coronary arteries (atherosclerosis), and diabetes mellitus.

Coronary heart disease, or ischemic heart disease, is the modern great scourge of privileged people over the age of 40. There is abundant evidence that it is on the increase and the increase is related to the privileged living that results from social and economic success in Western civilization. The environmental causative factors may include tension and strain, cigarette smoking, lack of exercise, and wrong eating. None of these environmental factors can be considered alone when attempting to explain causation; inherited differences between families must be considered along with the environment. In regard to diet there are many differences between what is consumed by privileged Western communities and what is consumed by underprivileged developing rural and

urban people. These differences tend to disappear with privilege and socioeconomic "development," and the incidence and prevalence of coronary heart disease goes upward with the change. This has been exemplified in the case of Japanese migrating to Hawaii and California, of the Irish migrating to the United States, of Yemenites migrating to Israel, and of the rural Bantu in southern Africa migrating to cities. Among the most unfavourable dietary trends are the total quantity and degree of saturation of fats in the diet. Saturated fats of animal origin and vegetable oils artificially saturated by manufacturing processes increase at the expense of unsaturated oils of natural vegetable origin. Cane sugar (sucrose) increases at the expense of natural starchy carbohydrates. Both these dietary changes have demonstrably unfavourable effects on fats (lipids) in the blood, which are known to be linked with the tendency to atherosclerosis in coronary arteries. In brief, Western privileged diet may be one of several factors jointly producing a most devastating disease of modern civilization. There is evidence that atherosclerosis in the coronary arteries may start in adolescence and may have been operating continuously over three or more decades before the disease manifests itself.

Diabetes mellitus is another disease of multiple causation that undoubtedly has inherited roots and that must be precipitated by a variety of environmental factors operating over long periods. One of these environmental factors may be diet. Its prevalence varies widely from place to place in the world and varies significantly among different populations in the same geographical area. Much complicated research is summarized in the statement that latent or potential diabetes is a worldwide disorder that is precipitated into damaging diabetes by environmental factors that include diet. The dietary causes certainly include overconsumption of calories, with resultant obesity, and perhaps undue dependence upon sugar (sucrose) as a source of calories.

SOPHISTICATED DIETS

Refined foods; harmful additives

Since man graduated from being a food gatherer to being a food producer, there has been an increasing tendency to cater, in food production, storage, and processing, to taste at the expense of physiological need. The extraction of white flour from the more healthy whole wheat meal was one of the trend setters. In the 20th century this process has followed a steeply rising curve. The results include certain widely prevalent deficiency diseases such as beriberi, increasing dental caries, and orthodontic problems.

Another manifestation of 20th-century food technology is the increasing presence of potentially toxic additives and contaminants in foods processed for delivery and consumption in large urban communities. Foods produced and consumed on one's own farm have many psychological and gustatory advantages, but it is not possible to make such foods available to inhabitants of large cities. Good bottled, canned, or frozen foods purchased in a supermarket may be more healthful than their so-called fresh counterparts that have been exposed and soiled in city retail markets. There is nothing inherently unhealthy in properly preserved foods. The same must be said about foods that have been produced with the aid of artificial fertilizers, such as phosphates and nitrates. Natural humus and manure are used by producers if they yield better and healthier growth of plants; these plants in turn will be more acceptable both to dairy herds and to human consumer-purchasers.

Of more potential danger is the modern tendency to stimulate the growth of meat-yielding animals with antibiotics and hormones that remain in meat, milk, or eggs intended for human consumption. The same is true of a great variety of modern pesticides that are used because they produce better crop yields, also of the colouring matters that are included in sweets and some canned foods to catch the consumers' eye. These and other additives and contaminants have considerable potentiality for harm to their human consumers. Some of them, ordinarily in doses relatively much larger than are consumed by the public, may produce cancer in experimental animals. This whole subject is under intensive scrutiny by the World Health Organization (WHO) and the Food and Agriculture Organization (FAO) and will need more and more careful scrutiny by national health authorities. Decisions concerning these potential hazards cannot be made without taking the growing population pressure into consideration; this is becoming so heavy that alternatives to the use of artificial agricultural and veterinary aids that become potentially dangerous additives and contaminants may be world starvation or malnutrition.

Many authorities now believe that the increasing pace of food sophistication in the present technological age has led to a considerable body of back-to-nature food faddism, characterized by dogmatic statements based on little or no scientific evidence. It is unreasonable to consider the real and alleged deleterious effects of food sophistication without recognizing that some degree of sophistication is inevitable as a result of the displacement of peasant subsistence by urban industrialized living. The latter is necessary to modern technology, and in spite of its dangers technology has brought with it many advantages, including a great increase in life expectation, increased productivity, and leisure for culture. Technological sophistication has brought improved quantity and quality of foodstuffs to the ordinary man and held at bay the effects of population pressure. It undoubtedly entails some potentially dangerous trends that will need to be carefully watched.

HEALTHFUL DIETS

In the face of 20th-century population pressure, one of the greatest challenges to mankind is to produce and make available to every person in the world sufficient calories to meet his energy demands, and a sufficient pattern and balance of protective nutrients to ensure healthy growth and maintenance. In developed nations, for all but the submerged minority, availability is ensured, and only health education is needed to bring about discriminating selection, healthy preparation, and disciplined consumption of food. For the submerged sections, lack of education and of earning power must be temporarily met by subsidized distribution. For a very large part of the developing world, however, neither education, earning power, nor adequate food production are assured. In general, hunger is periodic or seasonal and the drive of hunger keeps calorie deficiency vocal. Far more dangerous is the silent apathy of the millions whose hunger is satisfied by starchy calories and whose deficiencies of protein, vitamins, and minerals impair development and health. This is particularly dangerous for the under-five age group. Visible undernutrition and malnutrition represent only the peak of the iceberg; the real cause of poor growth and failure to thrive in young children, and of apathy at all ages, is represented in that part of the iceberg which is below the level of ordinary recognition.

It is usually admitted that recommended allowances have too wide a margin over the genuine minimum requirement and that this makes them unrealistic in the developing world. Nevertheless, they are probably right as an objective to be aimed at, and the width of the margin protects well-nourished communities from unusual stresses and does no harm except where excessive calories replace undernutrition with obesity.

The mere production, distribution, and marketing of food in appropriate quantity and quality is not enough. Health education in the broadest sense is necessary to ensure that recommended allowances are translated into familiar foodstuffs in appropriate quantity and combination. It should also ensure intelligent marketing, hygienic storage, cooking, and serving of food, and relaxed, unhurried eating in the home.

THERAPEUTIC DIETS

Certain diets that were previously believed to be important in the cure of disease are now recognized to be more appropriately directed towards the relief of symptoms. For example, it is now believed that, apart from the need to fulfill recommended allowances, diets for the treatment of duodenal ulcer should be directed towards the

relief of symptoms rather than regarded as fundamental for the cure of the ulcer. Diets for the management of diabetes mellitus in obese middle-aged adults are now directed particularly toward the reduction of weight, rather than to the disturbed metabolism itself. In some fields the importance of therapeutic diets has been greatly increased. The introduction of long-term dialysis (the use of semipermeable membranes, as in the "artificial kidney," to remove waste products from the blood) in chronic kidney disease has led to the inevitable removal of some part of the patient's store of nutrients. The restoration, through diet, of these nutrients and the maintenance of the best nutrient and calorie balance has become a complex and expert dietetic science. The therapeutic dietician has great responsibility for ensuring that a person being treated is able to translate recommended allowances into foods that are available and acceptable to him and is informed on the proper preparation, cooking, and serving of these diets.

BIBLIOGRAPHY. S. DAVIDSON and R. PASSMORE, *Human Nutrition and Dietetics*, 4th ed. (1969), a standard text intended for students in medicine, nutrition, and public health; N. JOLLIFFE *et al.* (eds.), *Clinical Nutrition*, 2nd ed. (1962), a standard work on medcial aspects of nutrition; A. KEYS *et al.*, *The Biology of Human Starvation*, 2 vol. (1950), a report of the well-known Minnesota Experiment on human volunteers to study the effects of war-time dietary deprivation; E.A. MARTIN, *Nutrition in Action*, 2nd ed. (1965), a readable text that provides background information on American concepts of nutrition; M. PYKE, *Food and Society* (1968), a brief, lucid discussion of British concepts of nutrition; M.G. WOHL and R.S. GOODHART (eds.), *Modern Nutrition in Health and Disease*, 4th ed. (1968), a standard work on the medical aspects of nutrition.

(J.F.B.)

Nutrition and Diet, Human

Nutrition is the process of assimilating food. The study of nutrition is the study of foods and their use in diet and therapy. Nutrients are the chemical components of foods utilized by the body, either as a source of energy or for building and maintaining tissues. Nutrients are conveniently divided into proteins, carbohydrates, fats, minerals, and vitamins.

The scope of nutrition

The scope of human nutrition and diet has now extended far beyond the purely chemical and physiological interpretation of the processes involved. Modern nutritionists are professional people whose training may have been in medicine, biochemistry, physiology, agriculture, or other science, but who have become increasingly involved in a wide range of problems in the general area of public health nutrition (see also FOOD SUPPLY OF THE WORLD). Infants and young children have physiological needs for growth above those of the normal adult. Old people, because of physical or mental infirmity, may fail to ingest adequate amounts of the right kinds of foods. These are vulnerable groups among whom ill health may occur because of undernutrition. The effects of undernutrition on health are seen most clearly on a worldwide basis, among the poverty-stricken, regardless of their location. The prevention of undernutrition, though it requires social action and financial assistance, is dependent as well on professional nutritional advice. It is not only the poor who suffer from malnutrition. In wealthy communities the prevalence of obesity, diabetes, atherosclerosis, and other dietary disorders is increasing and, although the precise causes of these are not always known, faulty food intake appears to be an important factor. There are nutritional problems among the affluent as well as among the poor.

Dietetics is concerned with menu planning and the provision of detailed advice about food and special diets for both individuals and institutions. Dietitians have a professional training with emphasis on the dietary treatment of the sick. Many of them work in hospitals, but others are employed in teaching or advisory capacities.

FUNCTIONS OF FOOD

The human body is a heat engine that releases the energy present in the foods and utilizes it partly for the mechanical work performed by muscles and in secretory processes and partly for the work necessary to maintain its structure and functions. The performance of this work is associated with the production of heat; heat loss is controlled so as to keep body temperature within a narrow range (see also METABOLISM).

Unlike other engines, the human body is continually breaking down (catabolism) and building up (anabolism) its component parts. Certain foods supply nutrients essential to the manufacture of the new material and provide energy needed for the chemical reactions involved.

The supply of energy. The energy taken in food and that utilized in daily life can both be measured. The measuring unit has been the kilocalorie (kcal) which is 1,000 gram calories; a gram calorie is the amount of heat required to raise one gram of water from 14.5° to 15.5° Celsius at one atmosphere of pressure. Nutritionists are now coming into line with other branches of science and using joules as the unit of energy; one kilocalorie is approximately equivalent to 4.2 kilojoules.

The energy present in food can be determined directly by measuring the output of heat when the food is oxidized in a calorimeter. Heats of combustion of individual proteins, fats, and starches are about 5.4, 9.3, and 4.1 kilocalories per gram, respectively.

Not all of this energy is available to the body because some ingested material is not absorbed from the gut and is lost in the feces; further, the nitrogenous compounds are not completely combusted, and some of the energy in proteins is lost in the urine, mostly as urea. Corrections for these losses give physiological values for dietary protein, fat, and carbohydrate of approximately four, nine, and four kilocalories per gram, respectively. These are called the Atwater factors, after an American physiologist, who, between 1895 and 1905, calculated the quantitative aspects of energy exchanges. Tables of energy value and nutrient content of common foods provide background data for general dietetic advice (see Table 1).

Sources of energy

Proteins, fats, and carbohydrates can, within wide limits, be interchanged as sources of energy. Among the members of prosperous communities most diets provide 12 percent of the energy as protein, about 40 percent as fat, and 48 percent as carbohydrate. In many poor agricultural societies, for whom cereals provide most of the energy, the figures for the individuals' diets are 10 percent for protein, 10 percent for fat, and 80 percent for carbohydrate. Throughout most of the world, protein provides between 8 and 14 percent of the energy ingested. High fat diets are associated with a high incidence of atherosclerosis (deposits of lipid material in the larger arteries), and high carbohydrate diets are likely to lead to obesity. Both diets are likely to be deficient in minerals and vitamins. Wide variations in the proportions of fat and carbohydrate in the diet may, nevertheless, be compatible with good health. Ethyl alcohol is another source of energy to the body. Only a small part of the intake, usually less than 5 percent, is excreted in the urine and expired air; most is oxidized in the liver, where it serves as a source of energy with a value of seven kilocalories per gram (almost that of pure fat). Muscle cannot utilize alcohol, and alcohol intake reduces the use of fat and carbohydrate by the liver. In this way alcoholic drinks increase the energy value of a diet and often contribute to obesity and fat deposits in the liver.

Building and maintenance. The body of a young adult human male weighing 65 kilograms (143 pounds) consists of some 11 kilograms of protein, nine of fat, one of carbohydrate, four of minerals, and 40 of water. During the first 20 years of life, an average of about 1.5 grams of protein and 150 milligrams of calcium must be retained from the diet every day in order to build soft tissues and skeleton. Tissues, however, are not static, and their components are being continually catabolized and replaced at varying rates. The inner epithelial lining of the gut is replaced every three or four days and red blood cells have a lifespan of only 120 days. On the other hand, collagen, a protein constituent of tendons, is probably turned over at intervals of ten years or longer. Much of the material derived from the breakdown of tissues is re-utilized, but

Table 1: The Energy Value and Nutrient Content of Some Common Foods (values per 100 g)

	energy (kcal)	water (g)	protein (g)	fat (g)	carbohydrate (g)	alcohol (g)
Whole wheat flour	339	15.0	13.6	2.5	69.1	—
White bread	243	38.3	7.8	1.4	52.7	—
Rice, raw	361	11.7	6.2	1.0	86.8	—
Milk, fresh, whole	66	87.0	3.4	3.7	4.8	—
Butter	793	13.9	0.4	85.1	trace	—
Cheese, Cheddar	425	37.0	25.4	34.5	trace	—
Beef steak, fried	273	56.9	20.4	20.4	0	—
Haddock, fried	175	65.1	20.4	8.3	3.6	—
Potatoes, raw	70	80.0	2.5	trace	15.9	—
Peas, canned	86	72.7	5.9	trace	16.5	—
Cabbage, boiled	9	95.7	1.3	trace	1.1	—
Orange, with peel	27	64.8	0.6	trace	6.4	—
Apple	47	84.1	0.3	trace	12.2	—
White sugar	394	trace	trace	0	105.0	—
Beer,* bitter	31	96.7	0.2	trace	2.2	3.1
Spirits* (gin, whisky 70 proof)	222	63.5	trace	0	trace	31.5

*Values per 100 ml.
Source: R.A. McCance and E.M. Widdowson, *The Composition of Foods* (1960).

new material is also required from food. Probably some 400 grams of tissue protein is replenished daily, but the minimal dietary requirement of protein is only about 40 grams per day.

Essential nutrients. Experiments with animals on artificial diets have enabled nutritionists to establish a long list of dietary substances essential for normal growth and the maintenance of good health. Most of these are present in sufficient quantities in all human diets; deficiency of any of them, however, may lead to serious symptoms of ill health in man. They include protein, minerals, and vitamins.

Protein. Growing children need more protein per kilogram of body weight than do adults. Protein requirement at all ages is increased by infections not only because there is an increased utilization of protein but also because illness usually impairs the appetite and thus reduces dietary intake of all substances, including protein. In many countries of the world children are weaned on a diet of cereal paps with little or no supplement of milk or other protein-containing foods. Such a diet at the least retards growth and development, and if a child on such a diet suffers from an acute infection, notably measles or gastroenteritis, a severe illness, which may take several forms—known variously as protein-calorie malnutrition, kwashiorkor, or nutritional marasmus—may ensue. The death rate from protein malnutrition in many poverty-stricken communities is enormous. There are areas in which more than 50 percent of the children die before their fifth birthday because of protein and energy deficiencies and repeated infections. The availability of suitable protein-rich foods for all children under five is the most important nutritional problem in the world today. In some instances, even where such foods are available or could be made easily available, they are not used because of ignorance of methods of child feeding.

Protein deficiency

Primary protein deficiency is not common among adults, for whom cereals in general satisfy the protein requirement. It does occur, however, in Africa and in certain other parts of the world where the staple food is cassava, a root with a low protein content.

Secondary protein deficiency, in which there is ample protein in the diet, arises from: (1) gastrointestinal disorders which interfere with the intake or digestion of protein or the absorption of amino acids, (2) chronic infections and injury, which increase the need for and utilization of protein, and (3) large losses of protein from the body, *e.g.*, in the urine, in some forms of kidney disease, and from the skin surface in the serous exudate from extensive burn areas.

Minerals. Iron is required for the synthesis of hemoglobin, the pigment in red blood cells. Normally the iron liberated from old cells is retained and can be re-utilized. When, however, there is chronic bleeding from wounds or during severe and prolonged menstruation, the normal amount of dietary iron may be insufficient to replenish the body's supply. Losses of iron in the menses, the needs of a fetus, and the inevitable loss at labour and in the milk of a lactating woman increase the iron requirements of women during their reproductive life. Most dietary iron is in a form which cannot be absorbed from the gut. Many diets contain about 12 milligrams of iron, of which less than ten percent need be absorbed by a normal adult male. At best only about 25 percent (3 milligrams) can be absorbed and this is only marginally adequate to meet the needs of menstruating women. If menstrual losses are large, iron-deficiency anemia inevitably follows unless the diet is supplemented with absorbable iron compounds. In most countries, because the dietary iron is insufficient to replenish menstrual losses, some 20 percent of the women are anemic, fortunately, as a rule, only to a minor degree. Iron-deficiency anemia may follow bleeding from any part of the gastrointestinal tract and, in such cases, may be severe. This is likely to happen, for example, when there is a heavy infection with hookworms, as occurs in many areas of the tropics, or in cases of bleeding ulcers.

Iron, calcium, sodium, potassium, magnesium, iodine, and trace elements

Calcium is needed for the development, growth, and maintenance of bone tissue, including teeth. It is probable that all human diets contain an amount sufficient for this purpose. Calcium deficiency may, however, arise from a failure to absorb the mineral, from a lack of vitamin D—causing a failure of bone development in children (rickets) and softening of bone in adults (osteomalacia)—or from disturbances of the secretions of the parathyroid hormone and of estrogen. Osteomalacia may follow any long-standing disease of the small intestine interfering with absorption. Osteoporosis, despite adequate calcium intake, may arise after the menopause (see BONE DISEASES AND INJURIES).

Sodium is present only in small quantities in most natural foods, but salt is added, often in large amounts, by food processers and by cooks. Dietary sodium deficiency occurs only in poor tropical communities in which low intakes of salt are unable to meet large losses in the sweat. Sodium is the predominant ion in extracellular fluid; an excess can cause edema, an accumulation of such fluid, especially in conditions such as congestive heart failure. A low sodium intake leads to a lowering of the blood pressure and brings about diuresis, ridding the body of the excess extracellular fluid. There is now much evidence that excess dietary salt may contribute to high blood pressure in some individuals, but many other factors are also responsible.

Potassium, present in all natural foods, is the predominant intracellular ion. Deficiency of it does not occur as a result of a primary dietary lack, but it may arise when there is acute water loss following diuresis or chronic diarrhea and may result from persistent misuse of purgatives or diuretics. All wasting diseases are associated with loss of potassium from the tissues. Potassium deficiency disturbs the excitability of tissues and leads to paralysis of muscle, including cardiac muscle. Sodium and potassium

Table 2: Recommended Daily Intakes of Energy and Nutrients for the U.K.

age range	occupational category	body weight (kg)	energy (kcal)	protein* (g)	thiamine† (mg)	ribo-flavin (mg)	nicotinic‡ acid (mg equivalents)	ascorbic acid (mg)	vitamin A§ μg retinol equivalents	vitamin D‖ μg cholecal-ciferol	calcium (mg)	iron (mg)
Boys and Girls												
0 up to 1 year		7.3	800	20	0.3	0.4	5	15	450	10	600¶	0¶
1 up to 2 years		11.4	1200	30	0.5	0.6	7	20	300	10	300	7
2 up to 3 years		13.5	1400	35	0.6	0.7	8	20	300	10	500	7
3 up to 5 years		16.5	1600	40	0.6	0.8	9	20	300	10	500	8
5 up to 7 years		20.5	1800	45	0.7	0.9	10	20	300	2.5	500	8
7 up to 9 years		25.1	2100	53	0.8	1.0	11	20	400	2.5	500	10
Boys												
9 up to 12 years		31.9	2500	63	1.0	1.2	14	25	575	2.5	700	13
12 up to 15 years		45.5	2800	70	1.1	1.4	16	25	725	2.5	700	14
15 up to 18 years		61.0	3000	75	1.2	1.7	19	30	750	2.5	600	15
Girls												
9 up to 12 years		33.0	2300	58	0.9	1.2	13	25	575	2.5	700	13
12 up to 15 years		48.6	2300	58	0.9	1.4	16	25	725	2.5	700	14
15 up to 18 years		56.1	2300	58	0.9	1.4	16	30	750	2.5	600	15
Men												
18 up to 35 years	sedentary	65	2700	68	1.1	1.7	18	30	750	2.5	500	10
	moderately active		3000	75	1.2	1.7	18	30	750	2.5	500	10
	very active		3600	90	1.4	1.7	18	30	750	2.5	500	10
35 up to 65 years	sedentary	65	2600	65	1.0	1.7	18	30	750	2.5	500	10
	moderately active		2900	73	1.2	1.7	18	30	750	2.5	500	10
	very active		3600	90	1.4	1.7	18	30	750	2.5	500	10
65 up to 75 years	assuming a	63	2350	59	0.9	1.7	18	30	750	2.5	500	10
75 and over	sedentary life	63	2100	53	0.8	1.7	18	30	750	2.5	500	10
Women												
18 up to 55 years	most occupations	55	2200	55	0.9	1.3	15	30	750	2.5	500	12
	very active		2500	63	1.0	1.3	15	30	750	2.5	500	12
55 up to 75 years	assuming a	53	2050	51	0.8	1.3	15	30	750	2.5	500	10
75 and over	sedentary life	53	1900	48	0.7	1.3	15	30	750	2.5	500	10
Pregnancy, 2nd and 3rd trimester			2400	60	1.0	1.6	18	60	750	10	1200	15
Lactation			2700	68	1.1	1.8	21	60	1200	10	1200	15

*Recommended intakes calculated as providing 10 percent of energy requirements. †The figures are calculated from energy requirements and the recommended intake of thiamine of 0.4 mg/1000 kcal. ‡1 nicotinic acid equivalent = 1 mg available nicotinic acid or 60 mg tryptophan. §1 retinol equivalent = 1 μg retinol or 6 μg β-carotene or 12 μg other biologically active carotenoids. ‖No dietary source may be necessary for those adequately exposed to sunlight, but the requirement for the housebound may be greater than that recommended. ¶These figures apply to infants who are not breast-fed. Infants who are entirely breast-fed receive smaller quantities; these are adequate since absorption from breast milk is higher.

are two of the most important ions in maintaining the homeostatic equilibrium of the body fluids.

Like potassium, magnesium is present in most foods. Deficiency, however, may follow chronic diarrhea and be a cause of weakness, depression, and disturbances in muscle contraction.

Iodine is necessary for the synthesis of the hormones of the thyroid gland. In the absence of adequate intake, the gland enlarges into what is known as a goitre. Goitres can become so large as to interfere with speaking and breathing, but can be prevented by administering medicinal iodine or adding iodides to dietary salt. Iodine deficiency is most marked in isolated villages in the Himalayas, the Andes, and other mountainous districts, where it probably affects more than 1,000,000 people and is difficult to prevent. Iodized salt was not used before the 1930s, and the Midwestern part of the United States was known as the "goitre belt" because of the low iodide content of its drinking water.

Trace elements include copper, cobalt, zinc, manganese, selenium, and molybdenum, which are all present in the body and probably essential nutrients serving as catalysts or components of organic molecules. Diseases attributable to lack of these elements in animals are known to veterinarians; if similar diseases occur in man, they are rare, although undoubted cases of zinc deficiency have been reported in children in villages in the Middle East. Cobalt is a component of vitamin B_{12}, a deficiency of which may lead to pernicious anemia.

Vitamins. These are organic compounds that the body is unable to synthesize and which, therefore, must be supplied in the diet or manufactured in the gut by the intestinal flora. Many of them function as components of enzyme systems. Vitamins were originally identified by letters as each new one was discovered, but this method of nomenclature is being replaced as the vitamins become known by their chemical names. Five major diseases are associated with dietary lack of vitamins: (1) beriberi and thiamine (vitamin B_1), (2) pellagra and nicotinamide (niacinamide), (3) scurvy and ascorbic acid (vitamin C), (4) xerophthalmia, a type of dry conjunctivitis leading to keratomalacia, or softening of the cornea (carotene or vitamin A), (5) rickets or osteomalacia in adults (calciferol or vitamin D).

Vitamin deficiency diseases

These were major diseases in the 19th century and were all "man-made" in that they arose as a result of changes in food habits or the way of life. During the last 70 years most of them have ceased to be of major importance, although small outbreaks and isolated cases still occur.

Keratomalacia, however, caused by severe deficiency of vitamin A, leads to permanent blindness in children, and appears to be on the increase. It has been estimated that it causes blindness in 20,000 children every year. A source of vitamin A is milk, which the child receives first from the breast (or "formula") and then from dairy animals. The vitamin also occurs in fish-liver oils, butter, egg yolks, cheese, and certain vegetables. In poor tropical countries where the supplies of dairy milk and other sources of carotene are inadequate, children usually continue to receive a little breast milk for two to three years. This may not provide sufficient vitamin A to prevent xerophthalmia and the associated night blindness caused by lack of the vitamin's action in the retina of the eye, but it usually prevents irreversible keratomalacia. The latter may occur after an infant has been weaned early onto a diet totally devoid of vitamin A. This is occurring more frequently in the growing cities in the tropics where there are many pressures on migrant women from the countryside to adopt Western ways of life, including early weaning of infants. These tragic cases of blindness arise from lack of knowledge of mothercraft and the unavailability of suitable infant foods.

Vitamin B_{12} and folic acid are two vitamins necessary for the normal development of the red blood cells and deficiency of either or both of these vitamins leads to anemia. A primary dietary vitamin B_{12} deficiency is rare, but a secondary deficiency caused by the failure of the stomach to secrete a factor necessary for its absorption is responsible for the disease known as pernicious or Addison's anemia. Folic acid deficiency leads to a similar anemia and may be secondary to chronic gastrointestinal disorder. It occurs commonly as a primary dietary deficiency in many poor tropical countries. Pregnancy increases the requirements of folic acid.

Table 3: Recommended Daily Dietary Allowances, 1968 Revisions*

person	age† (years from–to)	weight kg	weight lb	height cm	height in.	calories (kcal)	protein (g)	minerals: calcium (g)	minerals: phosphorus (g)	minerals: iodine (μg)	minerals: iron (mg)	minerals: magnesium (mg)	fat-soluble vitamins: vitamin A activity (i.u.)	fat-soluble vitamins: vitamin D (i.u.)	fat-soluble vitamins: vitamin E activity (i.u.)
Infants	0–⅙	4	9	55	22	kg × 120	kg × 2.2‡	0.4	0.2	25	6	40	1,500	400	5
	⅙–½	7	15	63	25	kg × 110	kg × 2.0‡	0.5	0.4	40	10	60	1,500	400	5
	½–1	9	20	72	28	kg × 100	kg × 1.8‡	0.6	0.5	45	15	70	1,500	400	5
Children	1–2	12	26	81	32	1,100	25	0.7	0.7	55	15	100	2,000	400	10
	2–3	14	31	91	36	1,250	25	0.8	0.8	60	15	150	2,000	400	10
	3–4	16	35	100	39	1,400	30	0.8	0.8	70	10	200	2,500	400	10
	4–6	19	42	110	43	1,600	30	0.8	0.8	80	10	200	2,500	400	10
	6–8	23	51	121	48	2,000	35	0.9	0.9	100	10	250	3,500	400	15
	8–10	28	62	131	52	2,200	40	1.0	1.0	110	10	250	3,500	400	15
Males	10–12	35	77	140	55	2,500	45	1.2	1.2	125	10	300	4,500	400	20
	12–14	43	95	151	59	2,700	50	1.4	1.4	135	18	350	5,000	400	20
	14–18	59	130	170	67	3,000	60	1.4	1.4	150	18	400	5,000	400	25
	18–22	67	147	175	69	2,800	60	0.8	0.8	140	10	400	5,000	400	30
	22–35	70	154	175	69	2,800	65	0.8	0.8	140	10	350	5,000	...	30
	35–55	70	154	173	68	2,600	65	0.8	0.8	125	10	350	5,000	...	30
	55–75+	70	154	171	67	2,400	65	0.8	0.8	110	10	350	5,000	...	30
Females	10–12	35	77	142	56	2,250	50	1.2	1.2	110	18	300	4,500	400	20
	12–14	44	97	154	61	2,300	50	1.3	1.3	115	18	350	5,000	400	20
	14–16	52	114	157	62	2,400	55	1.3	1.3	120	18	350	5,000	400	25
	16–18	54	119	160	63	2,300	55	1.3	1.3	115	18	350	5,000	400	25
	18–22	58	128	163	64	2,000	55	0.8	0.8	100	18	350	5,000	400	25
	22–35	58	128	163	64	2,000	55	0.8	0.8	100	18	300	5,000	...	25
	35–55	58	128	160	63	1,850	55	0.8	0.8	90	18	300	5,000	...	25
	55–75+	58	128	157	62	1,700	55	0.8	0.8	80	10	300	5,000	...	25
Pregnant						+200	65	+0.4	+0.4	125	18	450	6,000	400	30
Lactating						+1,000	75	+0.5	+0.5	150	18	450	8,000	400	30

*The allowance levels are intended to cover individual variations among most normal persons as they live in the United States under usual environmental stresses. The recommended allowances can be attained with a variety of common foods, providing other nutrients for which human requirements have been less well defined. †Entries on lines for age range 22–35 years represent the reference man and woman at age 22. All other entries represent allowances for the midpoint of the specified age range. ‡Assumes protein equivalent to human milk. For proteins not 100% utilized factors should be increased proportionately.

Riboflavin is a vitamin frequently lacking in the diet of people on inadequate cereal diets, and deficiency is associated with degenerative conditions of the skin and mucous membranes.

Fat-soluble vitamins

Vitamin K is essential for the production in the liver of prothrombin, a substance necessary in the process of blood coagulation. Deficiency probably never arises as a primary dietary defect, but, because the vitamin is fat soluble, it occurs when there is a failure to absorb fats from the small intestine, *e.g.*, in obstructive jaundice or with frequent ingestion of mineral oil. Vitamin K deficiency may lead to frequent and massive hemorrhage.

Vitamin E (tocopherol) is required by man, but there are few records of dietary deficiency. The vitamin has been recommended as a panacea for many ailments, but claims have seldom been supported by reliable evidence.

The dietary intake of each vitamin required to prevent the appearance of a deficiency disease can be measured with some precision. These are minimum requirements and much less than the amounts present in a "good diet"; they do not allow a reserve to be built up in the tissues. It is possible by various blood and urine tests to assess these reserves; when they are found to be low, a state of subclinical vitamin deficiency is said to exist. It is also claimed that high levels of vitamin intake are associated with high levels of health and in particular with freedom from infectious diseases. The evidence that vitamins promote health in any way apart from preventing deficiency diseases has not been universally accepted. The recommended intakes or allowances of vitamins (Tables 2 and 3) carry a safety margin which allows for a wide range of individual needs. Intakes may be less than recommended without any obvious effect on health; high intakes are of no benefit and may be dangerous in the case of the fat-soluble vitamins A and D (see VITAMIN).

CLASSES OF FOODS

Foods can be divided into nine classes: (1) cereals and cereal products, (2) starchy roots, (3) pulses and legumes, (4) vegetables and fruits, (5) sugars, preserves, and syrups, (6) meat, fish, and eggs, (7) milk and milk products, (8) fats and oils, and (9) beverages. Each of these contributes some of the nutrients essential for the makeup of a good diet.

Cereals. Rice, wheat, maize, and various millets are the main foods of man. In many poor communities over 75 percent of the dietary energy (calories) comes from cereals; this proportion falls with increasing prosperity, though even among the affluent it is seldom below 25 percent. The main source of this energy is starch, but from 9 to 14 percent comes from protein. Cereals are thus a good source of protein. The mixture of amino acids in different cereal proteins, however, is not the same, and any diet in which the protein is derived mainly from a single cereal is likely to be deficient in one or more essential amino acids and thus unsatisfactory, especially for growing children. One protein, however, can supplement another, and thus a suitable mixture of cereals may provide a sufficient amount of all essential amino acids. Cereals contain little fat and hence diets with a high proportion of cereals are likely to be deficient in the fat-soluble vitamins A and D. Cereals likewise contain no vitamin C, but the outer layers of the grain and the germ are rich sources of the water-soluble B vitamins: thiamine, niacin, and riboflavin. Unfortunately, these are largely removed in modern milling processes. White rice and white wheat flour are nutritionally inadequate; essential B vitamins must be provided by other foods or as supplements to the cereal product (*i.e.*, "fortified" cereals). Cereals contain calcium and iron, but not always in forms readily absorbed from the gut.

Cereal protein

Starchy roots. Potatoes, cassava, and yams are easily cultivated and are valuable as cheap sources of energy. Their nutritive value, in general, resembles that of cereals, but their protein content is lower. Protein deficiency may be common in tropical communities in which the staple food is cassava or yams. The potato, however, provides some protein—less than cereal does—but also contains some vitamin C and the pigment beta-carotene, convertible in the body into vitamin A. In 1910 a Danish investigator lived for a whole year on a diet comprised solely of potatoes and kept fit and well. Were it not for the remarkable nutritive value of the potato, the poor natives of the Andes and of Ireland might not have survived during certain periods of their history.

Pulses and legumes. Peas and beans, the seeds of *Fabaceae*, resemble the cereals in nutritive value but have a slightly higher protein content and, since they are not subject to milling, are a good source of B vitamins. They are thus a valuable supplement to a cereal diet, especially in tropical or subtropical countries; moreover, they, particularly the soybean, are also valued for their taste.

Vegetables and fruits. Vegetables and fruits all have similar nutritive properties. Because 70 percent or more of their weight is water, they provide comparatively little energy or protein, but many contain vitamin C and carotene, two nutrients which cereals lack. Fresh fruits, particularly the citrus variety, and their juices, are usually rich in vitamin C, but vegetables are an uncertain source, especially as this vitamin is easily destroyed in cooking. Vegetables and fruits contain indigestible cellulose, which adds bulk to the intestinal content and is useful in preventing constipation. Vegetables also provide calcium and iron but often in a form which is poorly absorbed.

Sugars, preserves, and syrups. The average consumption of sugar in the United States, United Kingdom, and other affluent nations is more than two pounds per person per week, and provides more than 20 percent of the energy intake of most individuals in these countries. Sugar, however, contains no protein, no minerals, and no vitamins, and thus has been called a source of "empty calories." Diets containing a large proportion of sugar may be low in essential nutrients. High sugar intake is associated with dental caries, diabetes mellitus, and atherosclerosis. The evidence for the causal relationship is strong in the case of dental caries, much disputed with respect to diabetes and atherosclerosis.

Advantages and disadvantages of dietary sugar

Sugar is, of course, an excellent preservative, and jams contain from 30 to 60 percent. Honey and natural syrups (*e.g.*, maple) are composed of more than 75 percent sugar and only extremely small amounts of other nutrients. Toffees, candies, and boiled sweets may be 50 to 90 percent sugar.

Meat, fish, and eggs. Meats generally consist of about 20 percent protein, 20 percent fat, and 60 percent water; the amount of fat present in a particular portion of meat varies greatly, not only with the kind of meat—pork, beef, lamb, etc.—but also with its quality; the "energy value" varies in direct proportion with the fat content. Meat is valuable for its protein, which is of high biological value. Meat also is a good source of iron and contains nicotinic acid and riboflavin, but, except for liver, little vitamin C or vitamin A. Most meats have similar nutritive value. Although many peoples are particular about the meats they eat, and many have religious customs prohibiting their eating certain meats or may be revolted by the idea of eating offal, snails, guinea pigs, frogs, locusts, lizards, snakes, dogs, and monkeys, the fact is that all of these are nutritious.

Beef teas and beef extractives, on the other hand, may be to some pleasant drinks that stimulate appetite, but they provide few nutrients.

Fish and other seafoods are excellent sources of protein and many species contain oils rich in vitamins A and D. Small fish (*e.g.*, sardines), if ingested whole, are also a source of calcium.

Eggs have a deservedly high reputation as a food. The white is protein and the yolk is rich both in protein and vitamin A. Eggs also provide calcium and iron; egg yolk, however, has a high cholesterol content.

Milk and milk products. Milk serves as a complete food for the human infant during its early months. Table 4 compares the composition of human and cows' milk, and indicates how rich both are in nutrients, especially calcium. Milk, however, is a poor source of iron and infants kept solely on milk for many months may develop iron deficiency anemia.

The vitamin C present in milk is destroyed by heating (pasteurization), which in many countries is required to prevent the milk from spreading bacterial and other infections. Infants fed solely on boiled milk are likely to develop scurvy unless given fruit juice or other sources of ascorbic acid. Ergosterol in milk may be converted to vitamin D by irradiation of the milk.

There are a large number of preparations of cows' milk suitable for feeding infants. Those made by firms of repute are excellent; others, sold to the rapidly growing urban population in the tropics cannot always be recommended; if made from skimmed milk they contain no fat-soluble vitamins. If these are his sole diet, an infant may go blind from keratomalacia unless vitamin supplements are also given. Skimmed milk as a supplement for children on mixed diets is a valuable additional source of protein (casein) and calcium.

Table 4: The Composition of Human and Cows' Milk
(all values per 100 g)

	human	cows'
Calories	70	66
Carbohydrate (g)	7	5
Protein (g)	2	3.5
Fat (g)	4.0	3.5
Calcium (mg)	25	120
Phosphorus (mg)	16	95
Iron (mg)	0.1	0.1
Vitamin A (i.u.)	170	150
Vitamin D (i.u.)	1.0	1.5
Thiamine (μg)	17	40
Riboflavin (μg)	30	150
Nicotinic acid (μg)	170	80
Ascorbic acid (mg)	3.5	2.0

Cows' milk is not suitable for extremely young infants because its protein concentration is too high. It can be "humanized" by dilution and the addition of sugar. Preparations are also available in which the butterfat has been replaced by a preparation of vegetable oils that closely resembles the fat present in breast milk. Babies thrive on such milks, or "formulas," but there is no scientific evidence that the normal infant gains extra benefit from the fat substitution that inevitably makes the preparations more expensive.

Infant "formulas"

It would be difficult, but not impossible, to rear a healthy child without milk, after weaning. Experimental nutritional data shows how milk benefits toddlers and school children. For healthy adults milk may be a pleasant but not necessarily an essential food. If large quantities are ingested, the intake of fats containing unsaturated fatty acids becomes high and this could possibly predispose to arteriosclerosis.

Cheese making is an ancient art, formerly used by farmers' wives to dispose of surplus milk. There are many local varieties, all nutritious and especially rich in protein and calcium.

Sour milks are prepared by fermentation, using as a starter a preparation containing *Lactobacillus acidophilus*, which converts lactose to lactic acid and prevents the growth of pathogenic organisms. Sour milks are not only hygienic but also nutritious.

Fats and oils. The fats most used by man are butter, suet from beef, and lard from pork. Important vegetable oils include olive oil, groundnut oil, coconut oil, cottonseed oil, sesame or gingelly oil, mustard oil, red palm oil, and corn (maize) oil. All these are high in calories. Only butter (other than the previously mentioned fish-liver oils) contains any of the vitamins A and D, but red palm oil does contain carotene which is converted to vitamin A in the body. All natural fats and oils serve as sources of "empty calories" and a large intake may depress the appetite for more nutritious foods. Some fatty acids, however, are essential to the human body's metabolic reactions, and some intake of absorbable fat is necessary for the absorption of the fat-soluble vitamins.

Margarines, vanaspatis, and other artificial fats are prepared by hardening mixtures of vegetable oils and also whale oil. Fat-soluble vitamins are usually added and some margarines are nutritionally equivalent to butter.

Beverages. Although most adults drink one to two litres (about one to two quarts) of water a day, much of this is in the form of coffee, tea, fruit juices, "soft drinks," beer, wines or spirits, or other liquids. In general, these are appreciated more for their taste and the substances they contain, or for their effects, than for their nutritive value. Fruit juices are, of course, useful for their vitamin C content and, being rich in potassium and low in sodium, are valuable in such therapeutic diets. Coffee and tea by themselves are of no nutritive value, but may be a vehicle for large intakes of sugar, milk, or lemon. The alcohol in beer, wines, and spirits can serve as a source of energy. Beer contains two to six percent alcohol, natural

Ethyl alcohol and "empty calories"

wines ten percent, and most spirits just over 30 percent. Since ethyl alcohol has an energy value of seven kilocalories per gram, very significant amounts of energy can be obtained from alcoholic drinks. With one or two exceptions, they contain no nutrients and are only a source of "empty calories." One exception, so-called country beer, may be an important source of the B vitamin group for some South American and African communities. The only vitamin present in significant amounts in beer from a brewery is riboflavin. Wines are devoid of vitamins, but sometimes contain large amounts of iron, probably acquired from iron vessels used in preparation, especially of cheap wine. It is possible for excess iron to be absorbed and stored in the liver where it may contribute to toxic manifestations.

RECOMMENDED INTAKES OF NUTRIENTS

Public Health nutrition involves standards of physiological and metabolic requirements that can be used to plan, within variable limits, nutritional policies, to assess the significance of surveys of dietary intakes, and to draw up ration scales; *e.g.*, for the armed services and those who live in such institutions as boarding schools, hospitals, prisons, extended care facilities, and the like.

The Food and Nutrition Board of the United States National Academy of Sciences–National Research Council, the United Kingdom Department of Health and Social Security, the United Nations Food and Agricultural Organization, and other national bodies have drawn up tables of daily nutrient allowances. Tables 3 and 4 set out the British and U.S. standards. Other authorities differ, mainly in detail, and some of these differences are considered below following a statement of four general points:

1. The figures are for recommended intakes and are not necessarily requirements.
2. Individuals vary significantly in their need for nutrients, notably for vitamins. The figures in Tables 2 and 3 are intended for application to large groups in which individual differences are averaged out.

Allowances versus requirements

3. The recommendations in Tables 2 and 3 are higher than the minimal requirements for the prevention for deficiency diseases and provide safety margins intended to cover most individual variations and also to prevent ill health associated with subclinical deficiency states. Even if an individual had an intake 25 percent or more below the recommended figure for any nutrient, it need not necessarily follow that he would suffer from malnutrition. In the present state of knowledge the correct size for the safety margins can be little more than a guess and there are legitimate differences of opinion. Both environments—internal and external—exert a great effect on individual nutrient requirements.
4. The requirement for energy is in a category different from that of most other nutrients, since the body has no means of disposing of an excess other than by converting it into fat and storing it in depots. Excess protein can be converted into carbohydrate and keto-acids and utilized as fuel or converted to fat and stored as such. An excess of minerals is rejected by control mechanisms regulating absorption in the small intestine and passed out in the feces. Excess intake of any of the water-soluble vitamins is rapidly excreted in the urine. There is no way by which the body can dispose of excesses of the fat-soluble vitamins A and D, which are stored in the liver and can give rise to dangerous toxic effects. Long ago Eskimos learned that polar bear liver was poisonous to man, and it is now known that this is because of its high vitamin A content. Poisoning with vitamins A or D is extremely rare with most natural foods, but it can occur among children overdosed with concentrates.

Energy requirements. The body needs dietary energy in amounts equal to the energy expended in the external work of daily physical activities, in the internal work of tissue maintenance, for repair in the case of disease or injury, and, in the case of children, for growth.

The French chemist, Antoine-Laurent Lavoisier, initiated the quantitative study of energy exchange in animals. He constructed a small calorimeter in which he was able to place a guinea pig and calculate the animal's heat output from measurements of the rate at which ice in the calorimeter melted. Lavoisier also measured man's rate of oxygen consumption and showed that it rose with exercise. During the 19th century numerous calorimeters for small animals were constructed, as were respiration chambers for humans, which allowed the rates of oxygen utilization and carbon dioxide release to be measured. It was not until the end of the century, however, that a calorimeter chamber was constructed in which a man could live for three or four days. During this 72–96-hour period all the energy expenditure could be measured as heat and equated with the net energy intake (energy in food minus energy in feces and urine). In this way the law of conservation of energy in man was demonstrated. There is, of course, no doubt that the human body operates within the limits imposed by the law of conservation of energy. Energy output is related to the rate of oxygen utilization, with one litre of oxygen equivalent to approximately 4.8 kilocalories of energy.

Quantitative studies in energy exchange

Measurements of the heat output of man, direct calorimetry, are difficult. By contrast, measurements of oxygen consumption by indirect calorimetry are relatively easy and can be made on humans in their normal day-to-day environment, at home, and at work in offices, factories, fields, and mines, and during recreations, including most sports.

Basal metabolism

The energy expended for maintenance at rest, known as the basal metabolism, amounts to about 1.25 kilocalories per minute for a man weighing 65 kilograms (one pound = 2.2 kilograms) and 0.90 kilocalorie per minute for a woman weighing 55 kilograms. The basal metabolism, if expressed per unit of body weight or per unit of surface area, as has been the tradition for a long time, appears higher in men than in women. This is so because women have a large store of body fat that is metabolically inert. The basal metabolism per unit of lean body mass is the same in the two sexes.

The metabolic rate is raised by as much as 30 percent after a meal, if the meal is rich in protein. This is known as the specific dynamic action of protein and is due in part to the work of secretion of the digestive juices and in part to chemical processes, mainly in the liver, involved in the metabolism of amino acids and other absorbed products of digestion.

The metabolism of a seated person usually is not raised by more than 50 percent above the basal level. The arms are light, and little physical work is involved in writing or sewing. Although the metabolism of the brain is high and accounts for 20 percent of the resting energy expenditure, it is not raised significantly by mental activity.

Energy expenditure is increased fourfold, up to about five kilocalories per minute, by a brisk walk and is usually between two and four times the resting rate during the accomplishment of such light work as most assembly-line labour in industry, domestic activities, painting and carpentry, and in participation in recreations such as golf or bowling.

An increase of up to sixfold or 7.5 kilocalories per minute constitutes moderate work and is characteristic of most pick-and-shovel jobs, gardening, tennis, and bicycling. Work involving energy expenditures above this is graded as heavy. Although many jobs in coal mining, lumbering, and the steel industry involve periods of strenuous work, relatively few persons in a modern industrial society do significant amounts of continuous heavy work. Cross-country skiing is the recreation with one of the greatest demands for energy and involves rates of 15 kilocalories per minute, or even more, for long periods.

Most individuals in urban societies have their days divided into three eight-hour portions. Normally, one is spent in bed and asleep, one at work, and one in recreation or in other nonoccupational activities. The period in bed at approximately basal rates involves about 500 kilocalories of energy. A rough guide for occupational work is as follows: sedentary (about 900 kilocalories per eight-hour period), including office workers, drivers, pilots, teachers, journalists, clergy, doctors, lawyers, architects, and shop workers; and moderately active (about 1,200 kilo-

calories per eight-hour period), including virtually all engaged in light industry and assembly plants, railway workers, postmen, joiners, most farm labourers, and builders' labourers.

A man's energy expenditure, however, is determined in an urban society more by how he uses his spare time than by the nature of his job. Nonoccupational activities may range from 800–1,800 kilocalories per eight-hour period. Energy requirements from food may vary from 2,200 kilocalories per day for sedentary activities to more than 4,000 for the very active. These figures are generalities and cannot be applied to peasant farming, still the way of life for the majority of the world's male population. Here there may be marked seasonal variations with men "very active" at seed and harvest time, and "sedentary" during some periods in between. Surveys in Africa and Asia indicate that a yearly average shows most peasants to be moderately active.

Corresponding figures for women show a total of about 2,200 kilocalories per day. The range of expenditure is not as wide as for men, and the great majority of women need food providing from 1,800 to 2,400 kilocalories per 24-hour period.

Much of the work done by humans consists in contracting muscles to move the body or its parts. Energy expenditure during walking, for example, is directly related to body weight, and the basal metabolism to body size. So, in theory, big people should require more food than small people. The theory does not always hold true, and some large men and women are less physically active than smaller ones and thus have somewhat lower energy requirements. On the other hand, athletes, large and small, have required food (energy) intakes far in excess of their size counterparts in less active occupations. It should be borne in mind, however, that exercise alone is not, by far, the most important factor in weight loss. The few glasses of beer consumed at the end of a golf game more than replace the calories used in playing the game.

As age advances, physical activity is curtailed and the need for dietary energy is reduced. Tables 2 and 3 indicate what may be considered normal requirements at different ages, including childhood, and also the additional needs for pregnancy and lactation, but, again, there is great individual variation.

Protein requirements. As long ago as 1909 it was shown that adult men on year-long experimental diets providing only 40 grams of protein daily remained in good health. The minimal protein requirement, as estimated by the amount of dietary protein required to maintain nitrogen balance (*i.e.*, replacement of body protein constantly broken down) is of the same order. Most Western diets provide at least twice this amount, and there is no doubt that such diets contain a great surplus of protein. The recommended intakes of protein given in Table 2 provide 10 percent of the energy input. Several authorities show lower figures, which are physiologically sound but appear to be impractical since it is difficult to prescribe low-protein diets that are nutritionally acceptable and provide as well as the necessary amounts of other nutrients.

Comparative requirements for calcium, iron, phosphorus, magnesium, and iodine

Mineral requirements. The United Kingdom recommended intake of calcium is 500 milligrams daily, which is in line with the United Nations recommendation, but lower than the United States figure of 800 milligrams. Good Western diets, on the average, provide 1,000 milligrams, over half of which usually comes from milk. Many Eastern diets provide less than 500 milligrams but are not associated with any signs of calcium deficiency, and all the evidence points to the fact that about 500 milligrams is ample.

The recommended iron intake for a woman is 15 milligrams daily in Table 2, whereas the United States statement of requirement is 18 milligrams and the United Nations ranges from 14 to 28 milligrams, depending on the protein content of the diet. The higher figures are physiologically sound in that these intakes are necessary to meet the large menstrual losses occurring in some 5 to 10 percent of women. It is neither easy nor inexpensive to provide diets containing 15 milligrams of iron, and higher figures may be unrealistic, unless foods are artificially enriched with iron—a technical problem not yet satisfactorily solved. It has been recommended that the best way to prevent iron-deficiency anemia in women with large menstrual losses is with medicinal iron.

In addition the United States daily recommendation covers phosphorus, 0.8 milligram; iodine, 140 micrograms; and magnesium, 350 milligrams. Few, if any, natural diets do not meet the requirements for phosphorus and magnesium. It has been pointed out that endemic goitre caused by iodine deficiency is widespread in some areas. Prevention is possible by adding iodides or iodates to table salt or by mass medication with a single injection of iodized oil.

Vitamin requirements. The recommended intakes in Table 2 for all of the vitamins, except vitamin C, are almost identical with those of the United States and United Nations. All provide a good margin of safety above the minimal requirements and are compatible with practical dietetics. In addition, United States authorities recommend vitamin E, 30 units; folic acid, 400 milligrams; pyridoxine (vitamin B_6), two milligrams; and cyanocobalamin (vitamin B_{12}), five micrograms.

There is little doubt that 10 milligrams daily of vitamin C is more than ample to prevent scurvy. Most British diets provide about 30 milligrams daily, and many much less. United States authorities have been slowly reducing their recommendation, which now stands at 60 milligrams daily. There is still disagreement on the matter of vitamin C intake, and in 1970 a Nobel Laureate, Linus Pauling, published a book on vitamin C in which he recommended a daily intake of relatively large amounts. His arguments were considered by some to be only theoretical; others appear to agree.

HUNGER AND FEEDING BEHAVIOUR

Man, like other animals, evolved in an environment in which food was scarce. Hunger is a sensation associated with various physiological changes that stimulate the drive to search for food. Feeding behavior depends on the activity of nerve centres in the hypothalamus, situated at the base of the brain. If, in a rat or other experimental animal, the medial hypothalamic nuclei are destroyed, the animal will begin to eat voraciously as soon as it has recovered from the anesthetic, and in a few weeks will become obese. If the lateral hypothalamic nuclei are destroyed, however, the animal may refuse to eat and, unless force-fed, may die of starvation. These effects are assumed to be due to the destruction of a "satiety centre" on the one hand and a "feeding centre" on the other.

Types of hunger stimuli

The stimuli activating the feeding centre may be speculated as being: (1) chemical; *e.g.*, a rise in the level of free fatty acids or a fall in the level of glucose in the blood; (2) nervous; *e.g.*, the hunger contractions that occur in an empty stomach; or (3) thermal; *e.g.*, the slight fall in the temperature of the blood that may occur in starvation. Satiety may be signalled by gross distension of the stomach and alimentary canal. (This, however, does not explain hunger in a gastrectomized individual.) The capacity of a hungry man for feeding is sometimes enormous. Soldiers after a period on half rations have been observed to eat 9000 kilocalories in a day, the equivalent of nine large meals. There are many tales of how quickly a group of Africans can dispose of the meat on a dead elephant. On the other hand, individuals deprived of food for long periods of time may require days or weeks of semisolid feeding before being able to eat a normal meal.

It is only relatively recently in human history that large numbers of people have been able to obtain their food with little or no physical effort. New sources of power do the work formerly carried out by human muscles, and now large numbers of people spend the greater part of their days sitting. Members of the "affluent society" are seldom really hungry, although they may have specific appetites; usually their feeding times are determined by social customs and convenience rather than by physiological need, and they readily become obese.

The nature of the control system that regulates the

Short- versus long-term calorie requirements

amount of energy reserve in the adipose tissue and, in turn, the body weight is not understood. Over long periods the system appears to be precise. An individual's weight, for example, may remain in the range 60–65 kilograms for 40 years, during which time he will have eaten about 20 tons of food. Yet, control over short periods is imprecise. In many field surveys the daily energy intake in food and the energy output in physical activity have been measured simultaneously. The results indicate that few people adjust their food intake to their needs on a daily basis, but nearly all do so over about a week. It is not unusual to accumulate a surplus or deficit of 2000 kilocalories, the equivalent of about three normal meals. Dietary habits are sometimes abruptly changed at the weekend. It is difficult to see how any of the physiological changes associated with hunger could be stimuli for coordinating the long term regulation of the reserves of energy. Presumably, the centres in the hypothalamus receive information about the size of the energy reserve in the fat stores, but how this information is conveyed is not known.

The regulatory mechanism is apparently altered in pregnancy. It is normal for a pregnant woman's weight to rise by about 12 kilograms in the 40 weeks of gestation; of this about four kilograms is fat in the adipose tissue—an extra energy reserve that presumably has survival value and is lost during lactation. Possibly the hormones responsible for the maintenance of the fetus in the uterus also modify the setting of the energy balance in pregnancy. Similar steroid hormones might have a regulatory function in the nonpregnant state (see also HORMONE).

There is no apparent marked difference, however, between functions of the endocrine glands in healthy people, both thin and fat. There is also no explanation of how some people remain thin while eating well and without restraining their natural appetite; others readily become obese, unless they diet continuously. Writers on this subject usually assume that the diet is adjusted to meet the demands of physical activity; the reverse possibility that physical activity is adjusted to meet the dietary intake is seldom discussed. Children obviously have a drive to play as well as a drive to eat. In contemporary urban society, opportunity for physical recreation is greatly restricted, and lack of exercise can contribute to the condition of obesity.

It is important to appreciate how little the regulating system has to be offset for obesity to follow. Thus a man who, without changing any of his other habits, began to eat each day an extra half slice of bread at breakfast or started to drive his car to work instead of walking briskly for five minutes each way, would put himself in energy surplus by about 50 kilocalories daily; after a year he theoretically would have gained over 2 kilograms in weight and in 10 years would be seriously obese.

THERAPEUTIC DIETS

Therapeutic diets may be restorative and designed to restore losses resulting from wasting disease. Alternatively, they may be designed to remove previous excess or to reduce the load on the functional capacity of an organ or on tissue damaged by, or susceptible to, disease.

Restorative diets. Any severe infectious disease or trauma from injury or major surgery results in wasting. Once the infection is overcome and the injury repaired, convalescence is aided by adequate nutrition. If infection persists, or if wounds heal slowly, appetite is often impaired, and the patient will not receive adequate nutrition without medical assistance. Nutritional problems are also present when patients are too ill to feed themselves, or are unconscious, or have diseases affecting the mouth or alimentary canal that interfere with the intake or retention of food.

Nutrition and the repair of injury

When normal feeding is possible but appetite is impaired, much can be accomplished by the use of food concentrates rich in protein and energy. Soups, souffles, jellies, cakes, and ice creams, for example, can be enriched with dried milk, casein, fat emulsions, or sugar. Supplements of vitamin preparations and minerals also may be supplied.

When normal feeding is difficult or impossible, nutrition can be maintained in a number of ways: by feeding through a tube passed into the stomach through the nose (nasogastric feeding), through a tube inserted surgically into the stomach or intestine, or into a vein. For intravenous feeding, mixtures of amino acids, fat emulsions, simple sugars, vitamins, and other substances may be used. A patient may be nourished in one or other of these ways for many weeks.

Restrictive diets. *Low-energy diets.* All reducing regimens owe what success they achieve to restricting the dietary energy below the level of energy expenditure. If a person is not more than 10 percent overweight, then weight loss may usually be achieved by reducing intake of carbohydrates, fats, and alcoholic beverages. If an individual is more than 10 percent overweight, it is best to get medical advice. Most obese people, if sufficiently motivated, can keep to a 1,000 kilocalories daily diet, and on this should lose a little over a kilogram (about two pounds) per week. Such a reducing diet should contain sufficient meat, fruit, and vegetables to ensure that needs for protein, minerals, and vitamins are met. This means restriction on other foods, particularly those rich in carbohydrates. A good hospital dietitian can advise an individual how best to achieve this in relation to his normal dietary habits. People need constant advice and encouragement and, if they are outpatients, should see their physician at regular intervals. Those who are seriously overweight (more than 15 kilograms in excess, even in someone young and otherwise healthy) should not undertake starvation diets without strict medical supervision.

A low-calorie diet greatly benefits overweight patients with diabetes or high blood pressures. Indeed, for many patients with these diseases, such a diet may give marked relief from symptoms and, in some cases, no other therapy may be needed.

Low-protein diets. Dietary protein restriction is indicated when disease of the kidney impairs its ability to excrete urea and other nitrogenous waste products. A diet providing 45 grams of protein daily is frequently recommended when there is evidence of renal failure. The reduction in protein intake substantially lowers the load on the kidneys but still provides sufficient amino acids to meet minimal requirements. Such a diet can be made appetizing and attractive; kidney patients can live on it for years. Further protein restriction is possible and a diet that provides only 20 grams of protein daily is sometimes prescribed for individuals with severe renal failure. Some manage well with this diet, and its use may lead to improvement in renal function, but for others it is not only unappealing but also provides insufficient dietary protein for maintenance of tissue integrity. Whether or not manipulation of the diet in treatment of kidney disease is useful depends to a great extent on the facilities for other methods; *e.g.*, hemodialysis, for treating the renal failure.

Low-salt diets. Wherever there is edema, or an accumulation of fluid in the tissues, an excess of salt in the body fluids must be suspected because excess sodium chloride "holds back" tissue fluid. In patients with cardiac failure, or kidney or liver disease, edema can often be markedly reduced by severely restricting salt intake. Salt-free, or extremely low-salt-content diets are most unpleasant. There are now available, however, salt substitutes the flavour of which closely resembles that of table salt, and low-salt diets, coupled with the use of diuretics to increase the flow of urine and consequently lower the body water content, are effective in promoting excretion of the excess fluid. All who are susceptible to edema are usually advised to restrict their salt intake by taking no table salt and avoiding salt-rich foods, such as cured meats, cheese, and most sauces, and to use canned goods processed without salt.

Restricting the salt intake has a tendency to lower blood pressure, even in health, and is part of the regimen for most patients with high blood pressure.

Low-fat diets. Before fat can be utilized by the body it first has to be emulsified and digested, for which bile and pancreatic juice are required (see DIGESTION, HUMAN).

Diseases of the pancreas, liver, and biliary passage may impair digestion, causing fat to appear in the feces, a condition known as steatorrhoea. In these circumstances a low-fat diet is needed. It is not difficult to reduce the fat intake from the normal figure of about 100 to 50 grams per day and to keep the diet appetizing. Under certain circumstances, more severe restriction may be needed, and it is sometimes difficult to prepare diets providing a daily content of only 25 grams of fat. When a patient is on a low-fat diet, a supplement of fat-soluble vitamins is usually given.

Low carbohydrate diets in diabetes mellitus. As already mentioned, a low-energy diet is of benefit to diabetics who are obese. All diabetic patients maintain better health if they restrict their diet so that their body weight is a little below rather than above their ideal weight. There is no firm agreement as to whether the carbohydrate content of a diabetic diet should be specifically reduced, but in most the proportion of calories derived from carbohydrates is kept low.

Low saturated fatty-acid diets. Reducing the intake of fats containing a large proportion of saturated fatty acids (*e.g.*, animal fats) or replacing them in the diet with fats high in unsaturated fatty acids (*e.g.*, vegetable oils) lowers the level of cholesterol in the blood plasma. As atherosclerosis (deposition of lipid materials within the arteries) is associated with a high level of plasma cholesterol, vegetable-oil diets have been tried to see whether they might prevent the appearance or reduce the progress of clinical manifestations of this disease. The results have been equivocal. Physicians now warn patients with evidence of atherosclerosis of the dangers of excess intakes of butter, cream, or fat bacon.

Gluten-free diets. Celiac disease, which usually begins in the first three years of life, is caused by a sensitivity of the small intestine to gluten, a mixture of proteins present in wheat and rye. The affected individual has chronic diarrhea, sore tongue, frothy, fatty stools, and a blood picture similar to that of pernicious anemia. These symptoms disappear when he is put on a diet from which gluten is completely excluded. Most celiac or sprue patients must remain on a gluten-free diet throughout life, although some affected children, as they grow up, acquire a tolerance for gluten and can then resume normal diets.

Special diets for inborn errors of metabolism. Phenylketonuria and galactosemia are two examples of these relatively rare diseases (see METABOLISM, DISEASES OF). In the former there is a failure to metabolize the amino acid phenylalanine, and in the latter, the sugar galactose, a constituent of lactose in milk. Children with either of these diseases can be reared successfully through the use of synthetic diets that contain little or no phenylalanine or galactose, respectively. Commercial preparations are available.

Many elaborate dietetic regimes, formerly often prescribed, are now known to have no scientific basis. There is no specific dietary treatment, for example, for either peptic ulcer or gout; *e.g.*, milk diets and low-purine diets as cures are now part of medical history rather than practice. This does not mean, however, that patients with these diseases need no dietary help. Their condition may have been aggravated (rather than caused) by faulty feeding habits and choices of foods "incompatible" with their ailment. The celebrated spas of Europe acquired their reputation because they imposed on patients a disciplined manner of life with restrictions on dietary excess. Many patients today might require similar regimens.

PUBLIC HEALTH NUTRITION

Historically, in times of famine government held the responsibility of feeding the people; during prosperity, each man had been expected to grow sufficient food for his family or to work for wages to purchase it; charitable persons fed the sick and the needy. Only within the 20th century have governments undertaken large-scale measures to prevent malnutrition. Wars, strangely enough, have played an important role in the evolution of this policy. In the 19th century, for example, malnutrition had become so rife that many industrial communities were not able to produce sufficient young men with the physique necessary for the armed services; moreover, the civilian population had not the working capacity needed to sustain wartime rates of production. Governments drew up wartime nutrition policies, many features of which have survived in peacetime with great benefit to the community. The enormous growth of the food industries has been beneficial to health, due principally to the cooperation of manufacturers and governments in ensuring that products are of acceptable standards. The rapid population growth, especially in many tropical countries, together with the increasing migrations from the countryside to the towns, has raised a number of public health problems concerned with nutrition. In many countries the importance and urgency of these problems are not yet sufficiently realized, and, even if they were, there are not enough people available with the technical skills to deal with them. The assessment and the mechanisms for handling nutritional problems on a national level are discussed in the paragraphs that follow.

Nutritional surveys. Statistical surveys may assess the nutritional status of a representative sample of a whole community or of a special group; *e.g.*, preschool children, elderly people, or the armed forces. There are three complementary components to such a survey: dietary, clinical, and biochemical.

Assessment of intake

The dietary intake of each individual or family is assessed over a specific period—usually seven days. If possible, every article of food consumed is weighed. If this is impractical, reliance must be placed on an accurate dietary history. Leftover and wasted food must be accounted for. Specially trained persons, preferably dietitians, collect the data. Using appropriate tables of food analyses, intakes of individual nutrients can then be calculated and be compared with tables of recommended allowances.

A clinical examination of persons taking part in the survey is made, in which physicians look for signs known to be associated with malnutrition. The heights and weights of all subjects are recorded and, whenever possible, other anthropometric measurements made; *e.g.*, skin-fold thickness, arm, head, or abdomen circumference.

Biochemical analyses may be made of blood or urine levels of vitamins and other substances that reflect the nutritional state. These levels are then compared with standards of "normality" for indigenous persons.

The execution and interpretation of nutritional surveys is difficult. First, there is the problem of selecting and contacting a statistically representative sample. Cooperation of the subjects is essential, and a sample more often than not may include "problem" families, or individuals who are more likely to suffer from malnutrition. Recommended allowances of nutrients and the biochemical standards of normality have relatively wide safety margins. When significant members of the sample surveyed fail to reach any of the standards, it may be difficult to draw up an unbiased report. In the past, nutritionists have sometimes exaggerated the importance of failure to reach these generous standards. Even experienced school doctors find it difficult to assess clinically the nutritional status of children who show no physical signs of deficiency disease. A single measurement of height and weight is also often misleading. Successive measurements over a period of months, showing whether or not growth is proceeding normally, are good criteria of whether a child or a group of children is receiving adequate nourishment.

In so-called advanced countries, in which many of the population are either well- or over-nourished, a survey of a random sample of the population designed to detect those persons suffering from malnutrition is misleading. Thus, in Britain, for example, it is well known and much publicized that some old people are seriously undernourished. A survey covering many parts of the country showed that over 95 percent of old people received a good diet and, indeed, many were obese. Probably no more than 2 percent of the old people in the country are

undernourished. This means, however, that there are 100,000 who need extra food and care—a sizeable social problem. In prosperous countries, medical authorities receive prompt reports concerning the incidence of infectious diseases such as smallpox, diphtheria, and typhoid, but there is no comparable mechanism for obtaining information about malnutrition.

Prevention of malnutrition. *Agricultural production.* In overpopulated countries a primary nutritional need is to increase production efficiency of the staple cereal. Countries with low yields (Table 5) could triple

Table 5: Yields of Cereals in Various Countries
(quintals per hectare)

wheat		rice	
The Netherlands	44.4	Japan	51.8
United Kingdom	40.5	Italy	48.9
United Arab Republic	26.8	United States	47.6
United States	17.4	Malaysia	25.8
Australia	12.0	Ceylon	19.3
Kenya	11.4	Hong Kong	18.3
India	8.3	Thailand	16.3
Pakistan	8.1	India and Pakistan	15.3
South Africa	7.0	Puerto Rico	6.8
Iraq	4.8		

Source: FAO.

Development of new varieties of staple cereals

or quadruple output by use of modern agricultural techniques. New varieties of wheat (Sonora 64 and Lerma Roja 64), and of rice (IH8), developed largely by Rockefeller and Ford foundations sponsored research, have produced a "Green Revolution," and large areas of India, Pakistan, and other countries are now sown with these new seeds. If plans materialize, the food supply in these countries should be sufficient for two or three decades, during which there might be time to give thought to checking the present alarming growth of population. High yields of staple cereal will free land for the production of other crops—*e.g.*, pulses, vegetables, and fruits—and also for improved animal husbandry. In this way, the quality of a national diet can be improved. Agricultural planning in these areas requires guidance more by human nutritional requirements than by market economy. This is possible only where there is firm governmental direction of agricultural activities and where the government has competent and influential nutritional advisors.

Food processing industries. It would be impossible to feed the populations of the large cities of the world if there were no industries capable of taking in bulk food from the farms and of preserving and processing it for distribution through wholesale and retail outlets. In all the technically developed countries the food industries employ chemists and nutritionists to see that the nutritive value of the raw food is retained as much as possible. In general, canned and frozen foods purchased in a supermarket may be of higher nutritive value than corresponding "fresh" foods in an open small-town market: the latter may have been on sale many days after harvesting. It would also be impossible to preserve and maintain the flavour of foods for the feeding of large populations without the addition, during processing, of chemicals which act as preservatives, emulsifiers, stabilizing agents, antioxidants, flour improvers, sweeteners, and flavouring agents. People in large cities may ingest small amounts daily of 100 such food additives, all of which could be potentially poisonous. The food industry and government agencies are well aware of this danger. Much effort is expended on research to determine safety margins for each additive and safety limits are set forth in legislation.

Food additives

The addition of vitamins, minerals, or amino acids can increase the nutritive value of foods. In many countries thiamine (vitamin B_1), and sometimes other nutrients are added to refined wheat flour to give it a nutritive value more closely resembling whole wheat flour. Margarines are enriched with vitamins A and D and, in general, have a nutritive value similar to that of good quality butter. Many infant foods are enriched with vitamin D and other nutrients. In some countries there is a statutory obligation to enrich some foods, while other foods may be enriched at the discretion of the manufacturers. Enrichment policies normally require legal sanctions in most countries.

Most nutritionists agree that in certain circumstances enrichment is beneficial; *e.g.*, in the manufacture of margarines. There is, however, considerable disagreement about the value of widespread enrichment.

Food rationing. Famine, war, and natural disasters, such as earthquakes and floods, may necessitate some form of food rationing. If rations, regardless of quantity, are provided regularly, the effect on morale may be excellent. A rationing scheme that does not work has a disastrous effect on morale. The supply of one pound of cereal per person per day provides about 1,600 kilocalories, approximately the basal metabolism of an adult man. With such a ration, the health and working capacity of a community does not deteriorate seriously over a period of weeks, during which the emergency may end. It may be necessary in such cases to commandeer as much of the milk supply as is possible for distribution to mothers, young children, and hospitals. Elaborate rationing schemes are possible only when the people have a strong sense of self-discipline and there are good administrative services.

Welfare foods. Young children and pregnant and lactating women form vulnerable groups, in that they are particularly susceptible to the ill effects of a poor diet. In many countries, free or cheap milk and vitamin supplements, sometimes in the form of fish-liver oils and fruit juices, are provided as welfare foods for them. Many well-controlled trials have demonstrated the benefit to young children of a small daily supplement of milk. The United Nations and other international organizations have given millions of tons of dried milk powder for distribution to children in underdeveloped countries.

Community feeding. The nutrition of an impoverished community can sometimes be improved by providing subsidized or free meals from communal kitchens. In times of famine, this is usually the most effective and practical method of giving relief. In the 19th century in Britain a few schoolteachers, realizing that it was impossible to teach effectively a hungry or ill-fed child, stimulated the formation of charitable organizations that provided school meals for needy children. Those organizations have been extended gradually to provide a national service for all children, and there was, early in the 1970s, a standard charge that approximately covers the cost of the food, but a government subsidy covers the cost of preparation and distribution. Children from poor families, about 8 percent of the total, received the meals free. School meals are a valuable preventive medicine measure in all countries. How to provide the most nutritious meals with the small amounts of money available is the enigma of nutritionists in poor countries.

School meals

Industrial workers frequently take their meals in the company canteen or cafeteria, usually run by the firm concerned but in some countries by the government. Management of most large industries appreciates the importance to employees of having nutritious meals readily available at a price well within their budgets.

Nutrition education. In many agricultural societies certain dietary patterns and food habits have been handed down from generation to generation, with traditional recipes and meals prepared today in a fashion similar to that of centuries ago. Some of these are well-balanced nutritionally, but, on the other hand, there are also national and religious taboos which, in times of other dietary restriction, could be dangerous.

Urbanization leads to changes in dietary habits, which inevitably constitute an initial retrogression. In a rapidly growing new town, for example, a fresh supply of traditional foods from the country is limited and costly; low wages can prohibit the family a free choice of food. Malnutrition became widespread in the industrial towns of Europe and North America in the 19th century, and it is becoming even more clearly manifest in the rapidly growing towns of Latin America, Africa, and Asia. Pov-

erty is, of course, largely responsible, but ignorance of food values is also an important factor. Even with a full purse and access to a well-stocked supermarket, it is possible to eat unbalanced meals. There is a degree of malnutrition among the rich as well as among the poor.

All young people, either at school or soon after leaving home and before setting up a home of their own, must know something of the general principles of nutrition and the relative nutritive values of various foods. This may be learned in many ways but perhaps most effectively in a course on health education or general biology.

During the last 50 years there has been a decline in the popularity of breast feeding of infants, particularly in Europe and North America. It is probably of no nutritional significance when there is an adequate supply of infant "formula" preparations, and the mothers know how to use them. A corresponding decline in breast feeding has begun in other parts of the world, notably in the larger cities in the tropics. There, however, suitable alternative preparations are not readily available and the women, for the most part, may be ignorant of methods of child feeding other than at the breast. As a result, early weaning often leads to malnutrition, with accompanying disastrous effects on the child. Clearly, in all countries and in all societies, instruction in nutritional mothercraft is of the utmost importance.

A community is well nourished only if it has proper nutritional services, with available advice of nutrition experts and supported by an informed public opinion. In all countries of the world there appears to be a shortage of nutritional experts. Nutrition is a broad, general subject with many technical and differing aspects. Not all universities are well equipped to teach it.

No governmental or other broad policies can survive without the backing of an informed public opinion. Food and nutrition are topics which attract faddists of all societies, and dietary "cranks" have skillfully propagated many false and dangerous ideas.

Nutritional administration. In developing countries in which malnutrition may be rife, technical resources scarce, and administrative experience limited, it may be valuable to have a central nondepartmental Nutritional Committee able to coordinate work in several departments. Where malnutrition is widespread, much of the effort must be organized and effected at local government or village level.

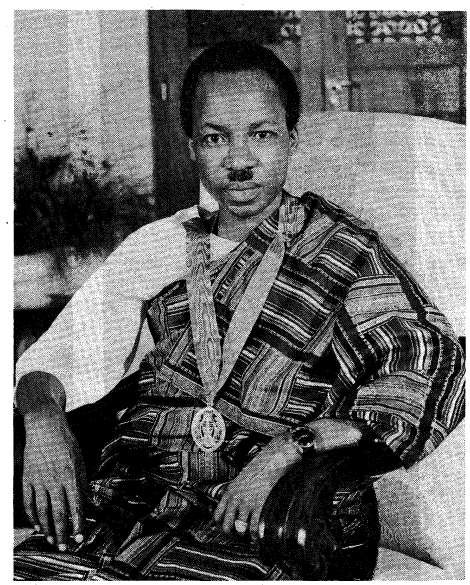
BIBLIOGRAPHY. C.M. TAYLOR and O.F. PYE, *Foundations of Nutrition*, 6th ed. (1966), an excellent introductory textbook; S. DAVIDSON and R. PASSMORE, *Human Nutrition and Dietetics*, 4th ed. (1969), a popular intermediate textbook; M.G. WOHL and R.S. GOODHART (eds.), *Modern Nutrition in Health and Disease*, 4th ed. (1968), the best advanced textbook with special reference to clinical aspects; D.B. JELLIFFE, *Child Nutrition in Developing Countries* (1968), written for field workers in developing countries; J. MAYER, *Overweight* (1968), a popular paperback describing nutritional problems in a prosperous community (written for the general reader, but technically reliable); R.A. MCCANCE and E.M. WIDDOWSON, *The Composition of Foods*, 3rd rev. ed. (1960), a good explanatory text, with tables giving analyses of 663 foods; U.S. NATIONAL RESEARCH COUNCIL, FOOD AND NUTRITION BOARD, *Recommended Dietary Allowances*, 7th ed. (1968); and DEPARTMENT OF HEALTH AND SOCIAL SECURITY, *Recommended Intakes of Nutrients for the United Kingdon* (HMSO, 1969), two reports giving a table of recommendations for nutrients with good explanatory text and many references.

(R.Pa.)

Nyerere, Julius

Julius Kambarage Nyerere became the first prime minister of an independent Tanganyika in 1961, and, when that nation merged with Zanzibar to form the new state of Tanzania in 1964, Nyerere became its first president. A strong believer in Pan-Africanism and a major force behind the Organization of African Unity, he has consistently advocated reason and moderation to achieve radical ends. From 1954 he led his country through all stages of independence without the loss of a life.

Nyerere was born in March 1922 at Butiama, Musoma District, Tanganyika, the son of a chief of the small



Nyerere.
Pictorial Parade

Zanaki tribe. He was educated at Tabora Secondary School and Makerere College, Uganda. A convert to Catholicism, he taught in several Catholic schools before going to Edinburgh University, where he was graduated in liberal arts in 1952. In 1953 he married Maria Magige, by whom he had five sons and two daughters.

By the time Nyerere entered politics, the old League of Nations mandate that Britain had exercised in Tanganyika had been converted into a United Nations trusteeship, with independence the ultimate goal. Seeking to hasten the process of emancipation, Nyerere joined the Tanganyika African Association, becoming its president. In 1954 he converted the organization into the politically oriented Tanganyika African National Union (TANU). In 1955 and 1956 he journeyed to the United Nations in New York as a petitioner to the Trusteeship Council and the Fourth Committee on trusts and non-self-governing territories. After a debate that ended in his being granted a hearing, he asked for a target date for independence. The British administration rejected the demand, but a dialogue was begun that established Nyerere as the pre-eminent nationalist spokesman for his country.

Formation of TANU

The British administration nominated him a member of the Legislative Council, but he resigned in 1957 in protest against the slowness of progress toward independence. In 1958 he accepted a plan for tripartite electoral representation of the three communities, African, Asian, and European. In subsequent elections, his organization (TANU) put up candidates from each of the three communities and won an overall majority. Progress toward independence owed much to the understanding and mutual trust that developed during the course of negotiations between Nyerere and the British governor, Sir Richard Turnbull. Once, when Nyerere had determined on a course of passive resistance involving nonpayment of taxes, he agreed to call off the campaign, though it meant jeopardizing his political leadership, in favour of constitutional steps proposed by the Governor. When Tanganyika gained responsible self-government in September 1960, Nyerere became chief minister and in December 1961 became the first prime minister of independent Tanganyika. When Tanganyika became a republic a year later he was elected president and in 1964 became president of the United Republic of Tanzania (Tanganyika and Zanzibar).

Nyerere's views and inspiration often have had effects beyond the borders of East Africa. When the white government of Northern Rhodesia threatened repressive measures against the African organization headed by Nyerere's friend, the Zambian leader Kenneth Kaunda, Nyerere took a leading part in mobilizing 5,000 unarmed volunteers to cross the border if necessary. Kaunda did not feel it necessary to call on the volunteer force and

went on to become the first prime minister of independent Zambia. Nyerere was a strong advocate of economic and political measures in dealing with the apartheid policies of South Africa. He was prominent in formulating the Lusaka Manifesto, calling for black and white cooperation in the peaceful development of all Africa.

Soft-spoken, unpretentious, small of stature, and quick to laugh, Julius Nyerere is widely credited with unusual powers of political perception. A key to his mind and understanding may be found in his two books, *Freedom and Socialism* (1968; *Uhuru na Ujamaa*) and *Freedom and Unity* (1967; *Uhuru na Umoja*), and in his choice of two plays of Shakespeare, *The Merchant of Venice* and *Julius Caesar*, for translation into Swahili. He has defended the one-party state, entrenched in his own and several other African nations, on the grounds that a young country, creating an independent political economy, cannot afford to threaten its foundations with elections. The defect of the system seemed to be illustrated by the trial in 1971 for treason of a number of Nyerere's closest colleagues and especially by the secret conclusion of the trial and imprisonment of the defendants for life. Nyerere has continued to insist, however, on an idealistic basis for his government. "We must try to find a method which will enable us in Africa to avoid the weakness of the 'national' state, and make it an instrument for the unification of Africa," he has said. "African nationalism is meaningless, dangerous, anachronistic, if it is not, at the same time, pan-Africanism." Again, addressing himself to the deep conflict of the 20th century, he has said, "The choice before the free states of the world . . . is not between peaceful change and no change. The choice is between peaceful change and conflict."

His defense of the one-party state

BIBLIOGRAPHY. For the history of Tanganyika and Tanzania and the part played in it by Julius Nyerere, see the selections from his writings and speeches collected in *Freedom and Unity* (1967) and *Freedom and Socialism* (1968). Other relevant books are ALEXANDER MACDONALD, *Tanzania: Young Nation in a Hurry* (1966); B.T.G. CHIDZERO, *Tanganyika and International Trusteeship* (1961); ROLAND OLIVER, *The Missionary Factor in East Africa*, 2nd ed. (1965); MARGARET L. BATES, "Tanganyika," in GWENDOLEN M. CARTER (ed.), *African One-Party States* (1962); MICHAEL F. LOFCHIE, *Zanzibar: Background to Revolution* (1965); and WILLIAM HAROLD INGRAMS, *Arabia and the Isles*, 3rd ed. enl. (1966). An excellent, three-part profile by W.E. SMITH may be found in the *New Yorker* (October 16, 23, and 30, 1971).

(G.M.S.)

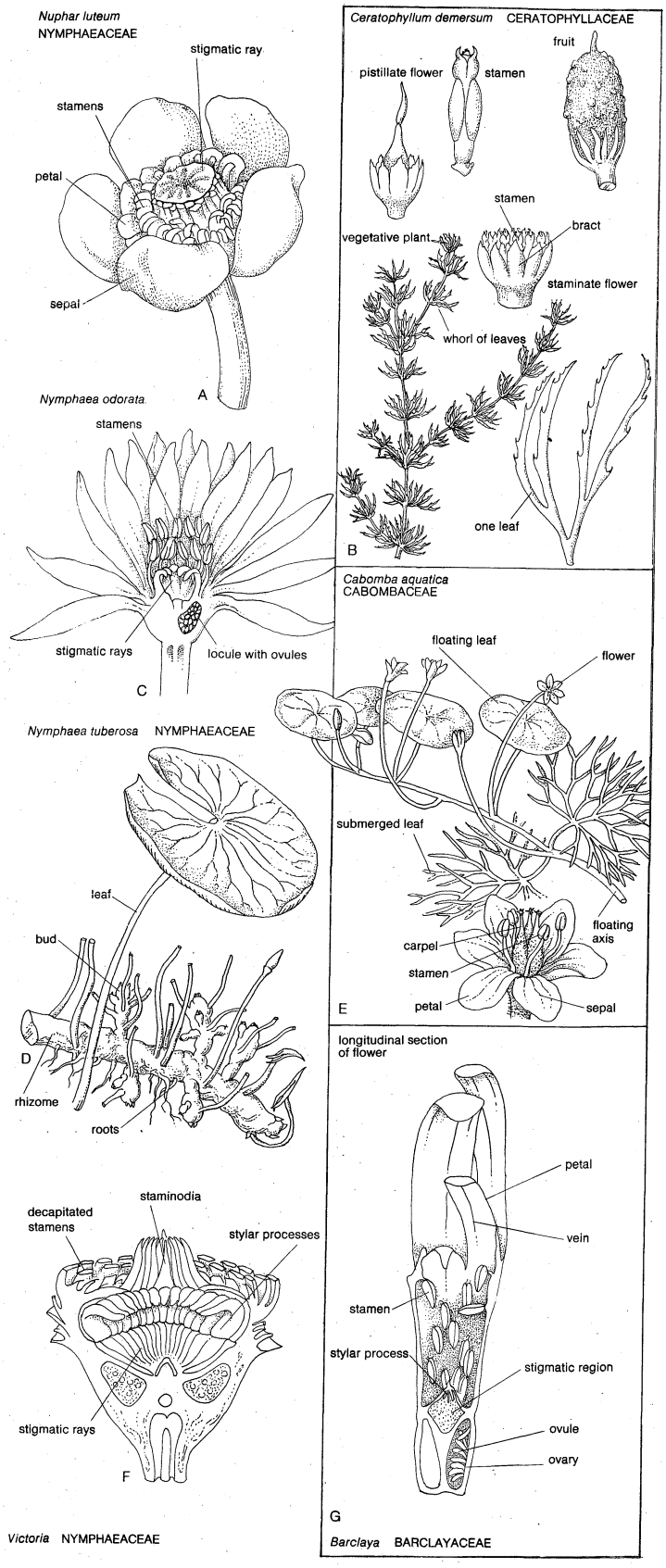
Nymphaeales

Nymphaeales is a plant order belonging to the dicotyledonous group of flowering plants. (One of two such groups, the dicots are roughly equivalent to flowering plants having broad, net-veined leaves—most trees, roses, daisies, etc., being familiar examples.) Its members are found in aquatic habitats, and its best known representatives are the water lilies, from which the order derives its common name, the water-lily order. The aesthetic appeal of the water lilies, especially those of the genera *Nymphaea* and *Victoria*, enhanced by their aquatic habitats, has led to their worldwide cultivation and familiarity. There are 4 families, 8 genera, and about 70 species.

GENERAL FEATURES

Plants of the Nymphaeales order are aquatic. In the genus *Ceratophyllum*, the plants are either rooted or free floating but submerged; in the other genera, they are attached to pond or stream bottoms, but their leaves and flowers may be either submerged, floating, or standing above the water. They range in size from the small, fragile fanwort (*Cabomba*) that has floating leaves less than an inch in diameter, to the hornwort (*Ceratophyllum*) with small leaves and richly branched stems reaching 12 to 15 feet (3.5 to 4.5 metres), to *Victoria* with its floating, circular, shieldlike leaves 6½ feet (2 metres) across and flowers 15 inches (38 centimetres) in diameter.

Representatives of the order occur, often abundantly and frequently, in quiet freshwaters over most of the earth. The genus *Cabomba*, for example, is found in the New World tropics and subtropics; *Brasenia*, the water



Representative plants from the four families of the water lily order.

Drawing by M. Pahl based on (*Nuphar luteum*, *Cabomba aquatica*, flower) E. LeMaout and J. Decaisne, *Traite general de botanique descriptive et analytique*, (*Nymphaea odorata*, *Ceratophyllum demersum*, flowers and fruit) G.M.H. Lawrence, *Taxonomy of Vascular Plants* (Copyright 1951); The Macmillan Company, (*Ceratophyllum demersum*, leaves) N.C. Fassett, *A Manual of Aquatic Plants* (1957); University of Wisconsin Press, (*Nymphaea tuberosa*) reprinted from Walter Conrad Muenscher, *Aquatic Plants of the United States*. Copyright 1944 by Comstock Publishing Company, Inc. Used by permission of Cornell University Press, (*Cabomba aquatica*, branch) J. Hutchinson, *The Families of Flowering Plants;* The Clarendon Press, Oxford

shield, is native to all continents except Europe; *Victoria* is found in tropical South America; *Euryale* and *Barclaya* live in the tropics and subtropics of Asia and Indonesia; the genus *Nuphar*, consisting mostly of the large yellow-flowered pond lilies, ranges over most of the Northern Hemisphere; *Nymphaea*, consisting mostly of smaller, white-flowered water lilies, is cosmopolitan, except in the Pacific islands and New Zealand; and the hornwort, *Ceratophyllum*, is also cosmopolitan.

The leaves; the starchy, horizontal, creeping stems; and the protein-rich seeds of the larger species have been used as food by man throughout his history and are used today, commonly by primitive peoples, and even more widely during famines. The seeds are eaten by fish, and all parts are eaten by grazing or browsing hoofed mammals. Several genera supply cover for fish and attachment and protection for their spawn. Fruits and seeds provide food, and the plant parts, extending above the water, provide food and cover for aquatic birds.

Cabomba, *Brasenia*, *Ceratophyllum*, and *Nuphar*, and other genera less commonly, may grow so rankly as to clog lakes and irrigation ditches. The Figure illustrates some of the structures characteristic of the order.

NATURAL HISTORY

Life cycle. All genera are perennial (*i.e.*, they live through more than one growing season) in their natural habitats, although plants of *Victoria* persist for only a few years. Those genera with creeping, horizontal stems may multiply vegetatively, as older branches die. Some species of *Nymphaea* produce detachable plantlets on their leaves.

Seed dispersal

The seeds of all genera are distributed passively. The fruits of *Nymphaea* species are thrust underwater by the arching growth of the floral stalk and are ruptured by the swelling of mucilage contained in the spaces around the seeds. The seeds of *Nuphar* species are released to float, in packets of soft tissue enclosing air, after the decay and opening of the fruit; the seeds eventually sink when released from the packets. Pulpy, bladderlike tissues produced by the seed stalks around the seeds of *Nymphaea*, *Victoria*, and *Euryale* carry the seeds until ruptured. In the genus *Ceratophyllum*, the spiny fruits are transported by submerged water currents or by fur animals, and the spines eventually anchor them in the substratum. Fish and seed-eating or fish-eating birds also distribute the seeds of various species.

Leaves, flowers, and young fruits of attached genera remain afloat or extend above water level, whichever is normal, even with rising water by means of continuing growth at the bases of leaf or floral stalks, respectively. Cutin and other waxes on floating leaves and flower parts prevent wetting of tissues and their sinking. A coating of mucilage covers the leaves and stems of species of the Cabombaceae family, probably as a protective mechanism. Sharp, strong prickles protect all exposed organs from animals in the genera *Victoria* and *Euryale*.

Ecology. The genus *Ceratophyllum* grows in stagnant or slow-moving water and may be either attached, by burial of the older end of the stem and its branches, or free-floating. All attached species characteristically grow in deep, silted, saturated soils, with little or no free oxygen, below quiet bodies of water. The creeping stems (rhizomes), submerged leaves, and roots are tolerant of oxygen-poor conditions. Air passages extending in the ground tissues from floating leaves, or leaves above the water, into the roots provide for some gaseous exchange between atmosphere and tissues. *Nymphaea* and *Nuphar* are able to withstand fairly severe dehydration without injury and can flourish in semiterrestrial, moderately aerated soils.

Insect associations

Many plant-eating insects, all of them flying types in the adult form, visit Nymphaeales species for food and to deposit eggs. No insect that lives on submerged plants or parts depends on a single species of plant; but there are specific relationships between emergent or floating plant species and flying insects. Some insects, dependent on floating plant parts, are not associated with submerged plant species since they cannot deposit eggs on them. Injury is due to sucking of food materials by bugs (Heteroptera), insertion of eggs and subsequent decay by dragonflies and damselflies (Odonata), removal of epidermis and leaf mining by beetles (Coleoptera), and gross destruction by the caterpillars of moths and butterflies (Lepidoptera). The transfer of pollen by flying insects is the usual pollinating mechanism, except in the genus *Ceratophyllum*. Pollinating visitors, also often plant-eating, are primarily beetles and flies (Diptera), but bees (Hymenoptera) and bugs have also been recorded as pollinators.

Pollination mechanisms

Natural mechanisms insuring cross-pollination and hybridization have been described for *Nymphaea*, *Nuphar*, *Victoria*, and *Ceratophyllum*. The stigmas (female pollen-receiving structures) of a flower become receptive before the anthers (male pollen-producing structures) of the same flower shed pollen. In the flowers of *Nymphaea*, *Nuphar*, and *Victoria*, the inner stamens (the male structure of the flower consisting of the fertile anther and its supporting stalk) or staminodia (sterile, petallike stamens) form a closed cone over the stigmas apparently preventing insect invasion after pollination and often trapping those insects that delivered pollen, causing them to drown in liquids present in the stigmatic cup. During the period when the female stigmas are receptive to pollen, nectar, sweet liquid in the stigmatic bowl, scent, and colour attract insects. Later, when the male anthers are ripe, pollen attracts the insects. In the submerged species of *Ceratophyllum*, male flowers are situated above female flowers. Pollen is released and sinks slowly through the water. Some of the pollen eventually slides down a groove contained in the pistil of the female flower and is thus deposited in a pit where pollen germination occurs, leading to eventual fertilization of the female ovules.

FORM AND FUNCTION

Vegetative features. Some genera have extensive, buried or surface creeping rhizomes (rootlike horizontal stem structures), especially the genera *Nuphar*, *Nymphaea*, and *Brasenia*; others have short, erect, buried stems with (in *Brasenia*, *Cabomba*) or without (in *Victoria*, *Euryale*) associated horizontal floating branches. The genus *Ceratophyllum* has long, largely floating stems with many lateral branches, but a plant may be anchored at its basal end. Branching is not understood in all genera. In the Cabombaceae family, branching is sympodial; that is, a vegetative apex is transformed into a flower, and a lateral branch continues vegetative growth. In others, branching is monopodial; that is, the main axis is continued and the flowers are laterals. In *Nuphar*, flowers arise in leaf sites, replacing leaves, but they are not directly associated with and located above leaves (the axillary position); vegetative buds arise at prospective floral sites and are axillary. In the genus *Nymphaea*, flowers and buds arise similarly, but neither is axillary. In *Victoria* and *Euryale*, the flowers arise in a successive, helical arrangement, alternating with a leaf helix. Roots, which are absent only in the genus *Ceratophyllum*, arise in clusters from the stem at leaf bases.

Vascular systems

The vascular systems in the stems of the Nymphaeaceae and Barclayaceae families are extremely complex and have not yet been satisfactorily described. In the Cabombaceae and Ceratophyllaceae families, relatively simple vascular systems are present. Very little xylem (water-conducting tissue made of special hollow cells known as tracheids and vessels) is present in the vascular bundles. Tracheids (nonspecialized supporting and water-conducting cells) are present, although scarce in the Cabombaceae family, in the organs of all genera. Vessels (specialized water-conducting cells) are usually reported as being absent, but recent studies have indicated that they may be present in the roots of some species of the Nymphaeaceae family.

Flower characteristics. Flowers are solitary in all genera of the order, except that in *Ceratophyllum* two flowers may occur in the axil of one leaf. Floral parts are arranged radially in each flower (*i.e.*, they are "regular" or actinomorphic). Anatomical studies indicate that probably floral parts are attached in complex, multiple

helical arrangements, although the helices may be nearly suppressed, and some organs appear externally to be arranged in whorls (cyclic). The flowers of the Nymphaeaceae, Barclayaceae, and Ceratophyllaceae families pose special problems in the study of their structures, which have been partly resolved by anatomical investigations. In the Nymphaeaceae and Barclayaceae families, the carpels (structures making up the female part of the flower, the ovary or pistil) are partially fused in some species of *Nymphaea*, or completely fused in a subwhorl around a central core except in the stigmatic (pollen receiving) regions. The ovary is multilocular (has several chambers) with few to many ovules attached widely over the side walls. The stigmatic region occurs as an upwardly widening cone, with a basal conical projection in some genera, and it has prominent to reduced extensions of the carpels termed stylar processes. The pollen-receptive stigmatic tissues occur on the inside surface of the stigmatic cone as laterally discontinuous, centrally grooved, radiating zones (rays), each above a locule (chamber of the ovary). Stigmatic-like tissue is continuous from the rays to the bottom of the inner lining of the locules, providing the pathway for pollen tubes. In *Nymphaea* and *Nuphar*, the perianth (flower petals and green leaflike sepals considered together) and stamens are hypogynous (attached below the carpels) and are free from one another, although some stamens are fused to the ovary in *Nymphaea*. In *Victoria, Euryale*, and *Barclaya*, the perianth and staminal organs are, reportedly, fused laterally and basally into a common tubular structure that is fused to, and indistinguishable from, the fused ovary wall. In *Victoria* and *Euryale*, therefore, the free lobes of the perianth and stamens are epigynous; that is, they arise from the top of the fused carpels. In *Barclaya*, the stamen-perianth tube extends above the fused carpels and the free staminal parts are attached to the inner wall of the tube where they hang downward. The inner stamens of *Victoria*, which assist in the cross-pollination mechanism, are, however, sterile; their basal, fused parts arch over the recurved stigmatic rays and are fused with them, and their free upper parts stand erect, after pollination.

The young female structure in *Ceratophyllum*, generally reported as being unicarpellary, is composed of two separate and unequal carpels, which fuse during development. The product is unilocular (single-chambered).

Biochemistry. Latex cells have been reported for most genera, except *Ceratophyllum*. All genera have a reddish anthocyanin pigment that turns purple instantly on exposure to air or to acid. Recent investigations have revealed an insulin-like compound in *Nymphaea nouchalii*, minute quantities of a spasmolytic alkaloid in the rhizomes of *Nymphaea alba*, and ellagitannins (dark-coloured carbohydrates) in *Nymphaea* and *Nuphar*, characteristic of other primitive dicotyledons, but absent in monocotyledons (dicots and monocots are the two great groups of flowering plants, the latter apparently derived from the former through some unknown primitive ancestor).

EVOLUTION AND PALEONTOLOGY

Origin of the order

On the basis of widespread fossil deposits of rhizomes, leaves and fruits of the genus *Brasenia* from Cretaceous deposits (about 100,000,000 years old) and of *Euryale, Nymphaea*, and *Nuphar* from Tertiary deposits (about 50,000,000 years old), apparently the Nymphaeaceae family has enjoyed a wide distribution over the earth since the early Tertiary Period, and possibly since the early Cretaceous Period. Some of the oldest angiospermous (flowering plant) pollens reported (reports not unchallenged, however) by paleobotanists are those of *Nuphar* and *Nymphaea* from the Jurassic Period (about 150,000,000 years ago).

Because of the relative lack of environmental stresses in their aquatic habitat, few striking structural changes from the ancient form have occurred in the Nymphaeales order during its long existence other than the evolution of fused carpels and enclosed ovaries (in some) to protect the seeds against plant-eating insects, and the adoption of aquatic features, including reduction of the xylem, development of air canals, and the rhizomatous habit.

CLASSIFICATION

The following characters of the Cabombaceae, Nymphaeaceae, and Barclayaceae families contrast with those of the Ceratophyllaceae family: adventitious roots (roots arising in unusual places such as along stems) are present; flowers are bisexual (male and female parts present in the same flower) and generally with differentiated petals and sepals; embryos are small; endosperm (the starch and oil containing tissue of seeds) is mealy but scanty; perisperm (tissue surrounding the endosperm) is abundant.

Annotated classification.

ORDER NYMPHAEALES

Aquatic herbs, usually perennial, from stout rhizomes or lacking roots and free-floating, juice sometimes milky, leaves simple, peltate (shieldlike, the stalk attaching near the centre of the disk), or in whorls, floating or emergent or submersed and then with dissected margins. Flowers bisexual, or unisexual, solitary, parts trimerous or numerous and arranged spirally; sepals, petals and stamens in some intergrading in structure. Sepals 3, 4, 5, or numerous; petals 3 to many, or none. Stamens 3 to many in cyclic or spiral arrangement, anthers splitting lengthwise. Carpels 3 to many or 1, usually free but sometimes coalescent; ovules 1 to 5 or many; ovary superior or inferior. Fruit a follicle, nutlet, or leathery berry.

Family Cabombaceae

Two genera, *Brasenia* and *Cabomba* containing only about 8 species. Leaves petiolate (*i.e.*, with a stalk); floating leaves of both genera are peltate. Submerged leaves of *Cabomba* are finely and dichotomously divided. Flowers small, barely floating, and hypogynous. Sepals and petals are apparently cyclic, bi- or trimerous. Stamens numerous, separate, apparently cyclic, and each is erect. Anthers split longitudinally. Carpels 3 to 6, or many, separate, unilocular, with 1 to 3 pendulous, anatropous, bi-integumental ovules attached laminally or rarely dorsally. Fruit a follicle.

Family Nymphaeaceae

Four genera (*Nymphaea, Nuphar, Victoria*, and *Euryale*), with about 53 species in all. Leaves are long-petioled, submerged, floating, or slightly emergent, peltate in *Euryale* and *Victoria*, but lanceolate (lance shaped) to orbiculate (round) with cordate (heart shaped) bases in the other genera. Flowers are relatively large, showy, generally floating or emergent, but commonly submerged and cleistogamous (not opening at maturity) in *Euryale*. The perianth is apparently cyclic. Sepals free and trimerous to 7 (*Nymphaea*), or to numerous (*Nuphar*), basally fused and epigynous (*Victoria, Euryale*). Petals showy, numerous, and separate (*Nymphaea*), basally fused and epigynous (*Victoria, Euryale*), or small, bracteate (with bracts), with abaxial nectaries (*Nuphar*). Appendages, which are transitional between petals and stamens, exist in *Nymphaea, Nuphar japonicum*, and *Victoria*. Stamens numerous, reaching 700 in some *Nymphaea*, separate, and generally hypogynous (*Nymphaea, Nuphar*), or basally fused and epigynous (*Victoria, Euryale*). Anthers split longitudinally and introrsely (directed toward the flower axis). Carpels partially fused (some *Nymphaea*) to syncarpous with 5 to 35 carpels (7 to 12 in *Euryale*), with multilocular ovaries, and with conical stigmatic regions. Ovules anatropous (*i.e.*, the ovule is bent back along the funiculus so that the micropyle opening is against it), except in *Euryale* with 2 to 3, suspended, and attached laminally. Fruit a tardily opening, leathery berry enclosing bi-integumental seeds with arils.

Family Barclayaceae

One genus, about 3 species, *Barclaya*. This family conforms to the generalized characters of the Nymphaeaceae family, except in the particular ones which follow. Leaves moderately long-petioled, floating, linear to orbicular, with cordate bases. Flower with 4 or 5 whorled, separate appendages, considered by some authorities to be sepals, at the base. Stamens basifixed (attached by their bases), pendulous, and attached to the inner wall of the extended epigynous perianth tube. Anther dehiscence longitudinal and introrse. Seeds without arils.

Family Ceratophyllaceae

One genus (*Ceratophyllum*) with about 6 species. Leaves sessile, 2 to several times coarsely and dichotomously divided, the margins serrulate. Roots absent. Flowers small, submerged to floating, unisexual (plants are monoecious), and lacking perianth. Each flower has an involucre (a whorl of bracts at the base of the flower) of 10 to 15 basally fused bracts (small leaflike appendages). Stamens 5 to 27, erect, with nearly sessile anthers. Anther dehiscence longitudinal and extrorse. Carpels 1 with 1 essentially orthotropous ovule which is

pendulous from near the top of the single locule. Fruit a spiny nutlet, tipped by a long style, containing 1 seed with 1 integument and no aril.

Critical appraisal. The delimitation of Nymphaeales, as presented here with the exclusion of the genus *Nelumbo*, frequently placed with this order, seems to be sound. The order is considered to consist of related groups, derived from a ranalian (*i.e.*, a group of dicotyledonous flowering plants showing certain primitive characteristics) ancestor, but semiherbaceous, rather than woody and treelike. The two genera *Victoria* and *Euryale* of the Nymphaeaceae family, with their massive connation (fusion) of floral parts, epigyny (*i.e.*, ovary enclosure by the fused bases of the sepals, petals, and stamens), and petal-like stamens, seem rather segregated from *Nuphar* and *Nymphaea* of the same family; on recent biochemical evidence this segregation is more justified than that of the genus *Barclaya* into its own family separated from *Victoria* and *Euryale*. Although *Ceratophyllum* seems closely related to the Cabombaceae family, the pollen types do not support such a relationship. Recent considerations based on biochemical, serological, and anatomical researches do not support the derivation of monocotyledons from the Nymphaeales order.

BIBLIOGRAPHY. V.I. CHEADLE, "The Occurrence and Types of Vessels in the Various Organs of the Plant in the Monocotyledoneae," *Am. J. Bot.*, 29:441–450 (1942), the first comprehensive study on this aspect of the subject; A.P. DE CANDOLLE, "Nymphaeaceae," in *Prodromus Systematis Naturalis Regni Vegetabilis*, 1:113–116 (1824), a systematic treatment of angiospermous families known at the time; A. ENGLER, *Syllabus der Pflanzenfamilien*, 12th ed. rev. by H. MELCHIOR, 2 vol. (1964), the most recent systematic survey of plant families; J. HUTCHINSON, *The Families of Flowering Plants*, 2nd ed., 2 vol. (1959), a modern systematic treatment of angiosperm families; H. KOSAKAI *et al.*, "Morphological Studies of the Nymphaeaceae: V. Does *Nelumbo* have Vessels?," *Am. J. Bot.*, 57:487–494 (1970), a recent research report on primitive vessels in the genus *Nelumbo*; G.H.M. LAWRENCE, *Taxonomy of Vascular Plants* (1951), a modern plant taxonomy textbook, with a survey of angiospermous families arranged largely according to Engler's concept of their evolution; H. LI, "Classification and Phylogeny of Nymphaeaceae and Allied Families," *Am. Midl. Nat.*, 54:33–41 (1955), a taxonomic study of the Nymphaeales order, including the most extensive available subdivision of the groups within the order; A.D.J. MEEUSE, "The Descent of the Flowering Plants in the Light of New Evidence from Phytochemistry and from Other Sources," *Acta Bot. Neerl.*, 19:61–72, 133–140 (1970), a review of recent phytochemical research on major groups of the flowering plants, and reports of different phyletic interpretations of the results; F.C. RICHARDSON, "Morphological Studies of the Nymphaeaceae: IV. Structure and Development of the Flower of *Brasenia schreberi* Gmel.," *Univ. Calif. Publs. Bot.*, 47:1–101 (1969), a recent research paper on the vegetative, floral, and developmental anatomy of the genus *Brasenia*; J.P. SIMON, "Comparative Serology of the Order Nymphaeales. II. Relationships of Nymphaeaceae and Nelumbonaceae," *Aliso*, 7:325–350 (1971), a serological study of relationships among all genera of the order Nymphaeales.

(M.F.Mo.)

Ob River

One of the greatest rivers of Asia, the Ob flows across Western Siberia in a twisting diagonal from its southeastern sources in the Altai Mountains to its northwestern outlet through the Gulf of Ob (Obskaya Guba) into the Kara Sea of the Arctic Ocean. It is a major communications artery, crossing territory at the heart of the Soviet Union that is extraordinarily varied in terms of physical environment and the character of its peoples: even allowing for the barrenness of much of the region surrounding the ice-threatened lower course of the river and the inhospitable waters into which it discharges, it drains a region of great economic potential, much of which is being realized under long-term Soviet development plans.

The Ob proper is formed by the junction of the Biya and Katun rivers, in the foothills of the Soviet sector of the Altai, from which it has a course of 2,287 miles (3,680 kilometres); but, if the Irtysh River is regarded as part of the main course rather than as the Ob's major tributary, then the maximum length, from the source of the Black (Cherny) Irtysh in China's sector of the Altai, is 3,362 miles (5,410 kilometres) or, if the Gulf of Ob be included, 3,959 miles (6,370 kilometres). The catchment area is approximately 1,150,000 square miles (2,975,000 square kilometres) down to the delta's end or 1,343,000 square miles (3,479,000 square kilometres) if the gulf be included—the aggregate 172,000 square miles (445,000 square kilometres) of undrained land in the south of the basin being included in both totals. The Ob's catchment area constitutes about half of the whole Kara Sea; it is the largest Soviet catchment area and the sixth-largest in the world.

The Basin. The West Siberian Plain covers about 85 percent of the Ob Basin, the rest of which is occupied in the south by the terraced plains of Turgay (Kazakhstan) and the small hills of northernmost Kazakhstan, and in the southeast by the Kuznetsky Alatau, by the Salair Range (Salairskoye Kryazh), by the Mountains of Shoria, and, behind them, by the Altai Mountains.

There are more than 1,900 rivers within the basin, their aggregate length being about 112,000 miles. The Irtysh, a left-bank tributary 2,639 miles long, itself drains 634,362 square miles (a somewhat larger area than that drained by the Upper and Middle Ob before the Irtysh confluence); and some 70 percent of the whole basin is drained by left-bank tributaries.

The huge basin of the Ob could be taken, geographically, as a representative cross section of Soviet territory: in the far south, around Lake (Ozero) Zaysan (recipient of the Black Irtysh and source of the Irtysh proper), there is arid semidesert; in the central regions of the West Siberian Plain—that is to say, over more than half of the basin—there is the swampy coniferous forest known as taiga, with very large expanses of marshland; and in the north there are vast stretches of the icy, treeless plains known as tundra.

The mainstream and its tributaries. The Upper Ob runs from the Biya–Katun junction to the confluence of the Tom, the Middle Ob from the Tom confluence to the Irtysh confluence, the Lower Ob from the Irtysh confluence to the Gulf of Ob.

The Biya and the Katun both rise in the Altai Mountains: the former in Lake (Ozero) Teletskoye, the latter to the south, among the glaciers of Mount (Gora) Belukha. From their junction the Upper Ob at first flows westward, receiving the Peschanaya, the Anuy, and the Charysh tributaries from the left; for this reach, the river has low banks of alluvium, a bed studded with islands and shoals, and an average gradient of one foot per mile. From the Charysh confluence the Upper Ob flows northward on its way to Barnaul, receiving another left-bank tributary, the Aley, and widening its floodplain as the valley widens. Turning westward again at Barnaul, the river receives a right-bank tributary, the Chumysh, from the Salair Range: the valley hereabouts is three to six miles wide, with steeper ground on the left than on the right; the floodplain is extensive and characterized by diversionary branches of the river and by lakes; the bed is still full of shoals; and the gradient is reduced, but the depth increases markedly. At Kamen-na-Obi, however, where the river begins to loop northeastward, the width of the valley shrinks to two to three miles and that of the floodplain to less than one mile, and reefs of rock emerge in places from the bed. Just above Novosibirsk another right-bank tributary, the Inya, joins the Upper Ob; and a dam at Novosibirsk forms the great reservoir known as the Ob Sea. Below Novosibirsk, where the river leaves the region of forest steppe to enter a zone of aspen and birch forest, both valley and floodplain broaden notably, till at the Tom confluence they are respectively 12 and three or more miles wide. The depth of the Upper Ob (at lower water) varies between six and a half and 20 feet.

The Upper Ob

The Middle Ob begins where the Tom flows into the mainstream, from the right. Taking at first a northwesterly course, the river henceforth becomes much deeper and wider, especially after receiving its mightiest right-bank tributary, the Chulym, shortly below the confluence of the Shegarka from the left. Successive tributaries of the

The Middle Ob

northwesterly course, after the Chulym, include the Chaya and the Parabel (both left), the Ket (right), the Vasyugan (left), the Tym (right), and the Vakh (right). Down to the Vasyugan confluence the river passes through the southern belt of the taiga (marshy forest country); thereafter it enters the middle belt. Below the Vakh confluence the Middle Ob changes its course from northwesterly to westerly and receives more tributaries: the Tromyegan (right), the Great (Bolshoy) Yugan (left), the Lyamin (right), the Great (Bolshoy) Salym (left), the Nazym (right), and finally, at Khanty-Mansiysk, the Irtysh (left, as has been said). In its course through the taiga the Middle Ob has a minimal gradient, a broadening valley (18–30 miles wide), and a correspondingly broadening floodplain (12–18 miles), through which it flows in a complex network of channels, with the main bed widening from less than one mile on the higher reaches to nearly two miles at the end and becoming progressively free of shoals. Low-water depths vary between 13 and 26 feet; and at high water there are great floods every year, sometimes spreading from 15 or even 50 miles across the valley and lasting from two to three months.

The Lower Ob

From its start at the Irtysh confluence the Lower Ob flows northwestward as far as Peregrebnoye and thereafter northward, crossing the northern belt of the taiga till it enters the zone of forest tundra in the vicinity of its delta. The valley is wide, with slopes steeper on the right than on the left, and the vast floodplain is much intersected by channels of the river and dotted with lakes. Below Peregrebnoye the river divides itself into two main branches: the Great (Bolshaya) Ob, which receives the Kazym tributary and the Kunovat from the right; and the Little (Malaya) Ob, which receives the Northern (Severnaya) Sosva, the Vogulka, and the Synya from the left. These main branches are reunited below Shuryshkary into a single stream nearly 12 miles wide; but after the confluence of the Poluy (from the right) the river divides itself again to form a delta, the two principal arms of which are the Khamanelskaya Ob, which receives the Shchuchya from the left, and the Nadymskaya Ob, which is the more considerable of the pair. At the base of the delta lies the Gulf of Ob, which represents a 500-mile extension of the river's valley invaded by the sea, with a width reaching 50 miles at certain points and its own catchment area (forest tundra and tundra proper) of more than 40,000 square miles.

Climate and hydrology. The Ob Basin has short, warm summers and long, cold winters. The average temperature for the year ranges from 14° F (−10° C) on the shores of the Kara Sea to 34° F (1° C) and 36° F (2° C) in the forest zone and to 38.8° F (3.8° C) on Lake (Ozero) Zaysan. The absolute maximum temperature, in the arid south, is 104° F (40° C); the minimum, in the Altai Mountains, is −76° F (−60° C). Rainfall, which occurs mainly in the summer, varies between averages of 12 inches a year in the north to 20 inches in the taiga zone and 100 inches on the steppes, but the western slopes of the Altai may receive as much as 62 inches a year. Snow cover, which lasts for 240–270 days in the north and for 160–170 in the south, is deepest in the forest zone (24–36 inches) and in the mountains (80 inches), much shallower on the tundra (12–20 inches), and very thin on the steppe (8–16 inches).

Annual flooding

On the Upper Ob the spring floods begin very early in April, when the snow on the plains is melting; and they have a second phase, ensuing from the melting of snow on the Altai Mountains. The Middle Ob, scarcely affected by the Upper Ob's phases, has one continuous spring–summer period of high water, which begins in mid-April. For the Lower Ob, high water begins later in April or early in May. Levels in fact begin to rise when the watercourse is still obstructed by ice; and maximum levels, which will have been attained by May on the Upper Ob, may not be attained till June, July, or even August on the lower reaches. For the Upper Ob, the spring floods are over in July, but autumnal rains bring high water again in September–October; for the Middle and for the Lower Ob, the spring–summer floodwaters gradually recede until freezing sets in (on the lower reaches, flooding may last four or five months). The rising of levels on the Ob proper and on the Irtysh obstructs the drainage of the minor tributaries' individual catchment areas.

The autumnal ice drift on the Ob lasts from about October 31 to November 10; then the lower reaches begin to freeze solid; by November 25 the whole river is frozen; and the upper reaches remain so for some 150 days, the lower for 220. The thawing of the ice, which takes longer than the freezing, lasts from the end of April (upstream) to the end of May; and the spring drift (about five days in duration) produces considerable ice jams.

The difference of level between high water and low is approximately ten feet on the Upper Ob; 36 feet on the Middle Ob down to Aleksandrovskoye, but only about 28 feet between Surgut and the Irtysh confluence; and at most 39 to 40 feet on stretches of the Lower Ob, but less than 20 feet at the mouth of the river.

The water is warmest in July, reaching a maximum of 82° F (28° C) in the vicinity of Barnaul.

In terms of drainage, the Ob is the third-greatest river of the U.S.S.R.—after the Yenisey and the Lena. Every year it pours about 515,000,000,000 cubic yards (394 cubic kilometres) of water into the Arctic Ocean—about 12 percent of that ocean's total intake from drainage.

The volume of flow at Salekhard, just above the delta, is nearly 56,000 cubic yards (42,800 cubic metres) per second at its maximum, 2,600 (2,000) at its minimum, while for Barnaul, on the Upper Ob, the corresponding figures are 12,673 cubic yards (9,690 cubic metres) per second and 211 (161). Most of the water comes from the melting of seasonal snow and from rainfall; much less of it comes from ground drainage, from mountain snow, and from glaciers.

The waters of the Ob are only slightly mineralized: dissolved substances account for an annual outpouring of 30,200,000 tons of ions into the Kara Sea. The total amount of solid matter brought down by the Ob every year amounts only to 50,000,000 tons.

Vegetation and animal life. Rich meadows extend in bands one to two miles wide for great distances along the banks of the Ob and cover many of the numerous islands. Pine, cedar, silver fir, aspen, and birch also grow on the banks and occasionally constitute isolated forests on the higher ground of the floodplain; and large areas near the river are covered with willow, snowball trees (*Viburnum*), bird cherry (*Prunus padus*), buckthorn (*Hippophaë*), currant bushes, and wild roses. Arable crops and vegetables are cultivated on some stretches of the banks.

Of some 50 species of fish to be found in the river or in the gulf, the most valuable economically are sturgeon, sterlet, and such "whitefish" as nelma (*Stenodus leucichthys nelma*), musksun (*Coregonus muksun*), chir (*C. nasus*), and pelyad (*C. pelea*); pike, burbot, Siberian dace, carp, and perch are also caught. For lack of oxygen in the water, however, many fish die every winter in the reaches between the Tym confluence and the delta.

Fur-bearing mammals of the Ob Valley include European and Siberian mole, Siberian and American mink, ermine, fox, wolf (in the taiga), elk, white hare, water rat, muskrat, otter, and beaver. Among more than 170 species of birds breeding in the floodplain are grouse, partridge, goose, and duck.

The human imprint. Politically, most of the Ob Basin belongs to the Russian S.F.S.R., but the south of it forms the northernmost part of Kazakhstan. Russians, Ukrainians, and Belorussians constitute the majority of the population, but there are, of course, numerous non-Slavic peoples also. These include the Kazakhs in the south, the Altay and Shorian peoples of the mountains, the Tatars of the Irtysh Basin, the Khanty and the Mansi, whose *natsionalny okrug* (national area) occupies part of the taiga; and the Nenets, Nganasan, Enets, and Selkup peoples of the north. The valleys of the river are more densely populated than other parts of the basin.

The earliest known sailing directions for the Lower Ob are those appended to Pyotr Ivanovich Godunov's *Chertyozh* ("Chart") of Siberia (1667); further details were

Exploration and development

provided by Semyon Ulyanovich Remezov in his *Chertyozhnaya Kniga Sibiri* ("Chart-Book of Siberia"; completed 1701); and the Russian scientists of the Great Northern Expedition (1733–43) investigated the Lower Ob as well as other Siberian rivers. For the next 150 years the river system was investigated chiefly for purposes of communication within the basin. Hydrological studies, however, were inaugurated by the end of the 19th century and were pursued intensively in the 20th; and during the Soviet period the hydroelectric potential of the rivers was not only studied but also developed.

The Ob's total hydroelectric potential is estimated at 250,000,000,000 kilowatt-hours. There were three stations operative in the early 1970s: one on the Ob proper, at Novosibirsk (400,000 kilowatts, its reservoir having an area of 410 square miles and a capacity of 11,500,000,000 cubic yards [8.8 cubic kilometres]); the other two on the mountainous reaches of the Irtysh, at Bukhtarminsk and at Ust-Kamenogorsk.

The Ob, one of Western Siberia's principal means of communication, is navigable for 190 days of the year on its upper reaches, for 150 on its lower. It serves the basin both for importing and for exporting. The Trans-Siberian Railway crosses the Irtysh at Omsk and the Upper Ob at Novosibirsk. Railways going into Kazakhstan from Novosibirsk and from the foothills of the mountains cross the Upper Ob at Barnaul.

Several benefits would accrue if more dams were built on the Ob: (1) its hydroelectric productivity would be augmented, (2) its floods would be less extensive, (3) its navigation facilities would improve, and (4) some of its water might be diverted to irrigate the arid parts of Kazakhstan and Central Asia.

(L.K.M.)

Ocean Basins

The ocean basins and the continents constitute the largest relief features on Earth, with the ocean basins, covering three-fifths of the Earth's surface, dominant. Water, including that of the shallow seas, actually covers 71 percent of the Earth, but only 60 percent of this overlies the deep ocean basins, which occur below the 2,000-metre (6,500-foot) contour line. The average depth of the sea is 3,800 metres (12,500 feet), whereas the average height of the continents is 840 metres (2,760 feet). Thus, the total relief contrast is 4,640 metres (15,220 feet).

The distribution of ocean basins and continents is asymmetrical. Continents are generally antipodal (diametrically opposed) to ocean basins. The antipodal position of Australia, for example, is within the North Atlantic Basin, and Antarctica opposes the Arctic Basin. The ocean basins lie principally in the Southern Hemisphere, and the Antarctic Basin encircles the Earth. In a sense, the three major oceanic basins (Pacific, Atlantic, and Indian) may be regarded as huge gulfs that extend off the Antarctic Basin, with the Arctic Basin being a secondary northward extension of the Atlantic. Thus, unlike the separated and isolated continents, the ocean basins are all interconnected, so that there is really only one worldwide ocean basin. This ocean basin is subdivided into a number of individual ocean basins largely as a matter of convenience.

A classical view of natural philosophers held the oceans to be as deep as the mountains are high, which is roughly correct if only the greatest mountains are considered. The first deep-sea sounding was made in the central South Atlantic in 1840; a heavy plummet was lowered 2,425 fathoms (4,435 metres [14,500 feet]) on the end of a long line. The first generalized map of the ocean basins was fashioned in 1895, using the 7,000 oceanic soundings deeper than 2,000 metres (6,500 feet) that were then available. The advent of the echo sounder in 1920, with which precisely timed echoes are bounced off the bottom, revolutionized deep-sea soundings. A modern survey ship can obtain a sounding every second along its track. The problem of obtaining accurate position control, or fixes, under all conditions of weather and in any part of the world was solved recently through satellite navigation, determination of fixes continuously from low-orbited satellites. The bathymetry (depth measurement) of the entire ocean floor is rapidly being revealed and charted in the 1970s with literally millions of soundings.

Before about 1930 the ocean floor was regarded as flat, monotonous, and generally featureless. Charts since then show a topography remarkably varied in both shape and relief. Some seamounts (isolated conical submarine peaks) and escarpments (steep slopes) are higher and more rugged than any on land, whereas the abyssal (deep-sea) plains are the levelest surfaces on the face of the globe. It has also been learned that the ocean floor, like the surface of the Moon, is a distinct geomorphic domain (*i.e.*, occupied by distinctive landforms); constructional and depositional physiography are quite unlike that on land. Erosional landforms on the continents are sculptured by wind, ice, and running water, but only sluggishly moving water modifies the deep ocean floor. Terrestrial stream action is imitated beneath the sea by turbidity currents, mud-laden tongues of water that periodically pour down the continental shelves and slopes to the ocean floor. Weathering (rock disintegration by chemical and mechanical processes), as well as erosion, proceeds slowly beneath the sea, so that the sea-floor morphology (fault scarps, volcanic knolls, and other features) tends to retain a pristine appearance.

Undersea constructional topography commonly differs from that on land. There is, for example, no undersea equivalent of folded sedimentary mountains. The major features of the ocean floor are deep-sea trenches, rifts, and fracture zones that are created by the interaction along their boundaries of shifting rigid crustal plates. The giant volcanic cones that create seamounts are especially spectacular. The smaller features represent a variety of volcanic topographic forms, related mainly to fissure eruptions (along linear cracks rather than through a central vent) and rifting of the ocean floor.

Mineral resources

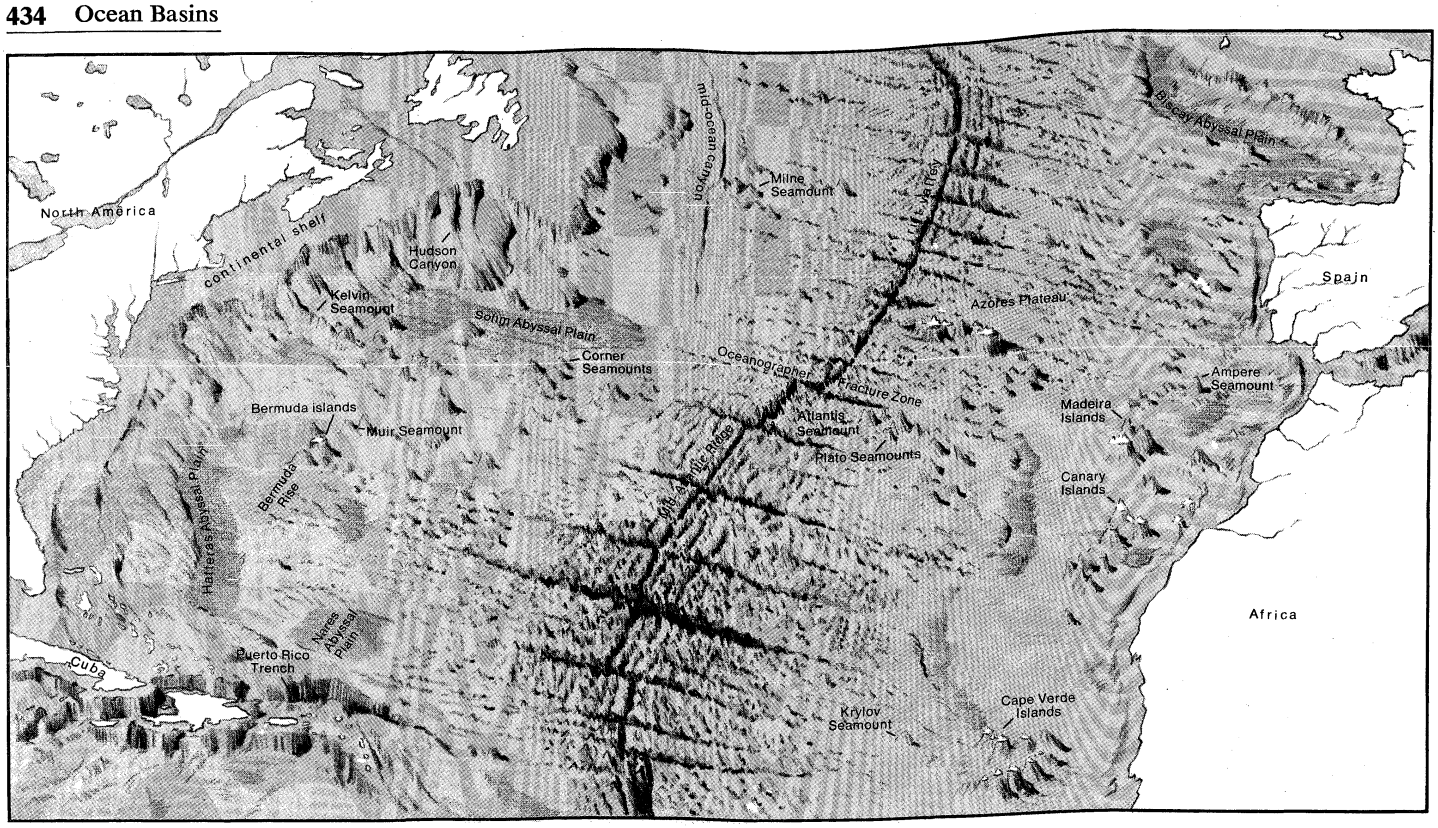
The mineral-resource potential of the deep ocean floor commonly has been considerably overstated. Basalt, which is the common rock of the oceanic crust and of seamounts, offers uninteresting mineral prospects. Iceland and Hawaii provide subaerial examples of oceanic rocks, and these, like all oceanic islands, have little to offer in the way of useful minerals. On the other hand, much interest has been accorded to marine phosphorites and manganese nodules, both of which are formed on the sea floor. The manganese nodules may someday be commercially recovered because of their high content of copper, nickel, and cobalt. Metalliferous sediments beneath pockets of hot brine in the Red Sea are another possibility. Also a highly mineralized layer commonly lies between the sediments on the sea floor and the underlying oceanic crust.

With the exception of precious coral, there is no mining of the deep-sea floor today. Precious coral occurs primarily around some of the coral atolls and seamounts of the central western Pacific, where it is gathered from deep crevices by the Japanese.

This article treats the geologic and geographic features of the ocean basins and includes a section on their origin. For further information on the areas marginal to ocean basins, see CONTINENTAL SHELF AND SLOPE, and for treatment of sediment transport from these areas to the ocean basins see CANYONS, SUBMARINE; DENSITY CURRENTS. See also OCEANS AND SEAS; OCEANIC RIDGES; and SEA-FLOOR SPREADING for additional detail on oceanic processes of relevance and EARTH, PHYSIOGRAPHY OF for an overview of the dimensions and interrelations of land and sea areas.

COMPONENTS OF OCEAN BASINS

The oceanic crust. The oceanic crust contrasts sharply with the granitic crust that makes up the continental blocks. The ocean floor is separated from the Earth's upper mantle (zone beneath the crust) by three distinct layers, which have been revealed by seismic refraction studies. These are: Layer 1 (zero–two kilometres [zero–one mile] thick), consisting of unconsolidated sediments that have been derived from the continents by submarine density currents or surface transport; Layer 2,



Midoceanic ridge, fracture zones, and sea-floor topography of the North Atlantic.
Drawing by E. Derdeyn

comprising Layer 2a (0.5 kilometre [0.3 mile] thick), consisting of pillow lava, a submarine form produced by the rapid quenching of lava as it poured from fissures onto the ocean floor, and Layer 2b (two kilometres [one mile] thick), a series of feeder dikes (dikes are solidified sheets of lava that intrude pre-existing rocks) that originally provided passages for the lavas onto the ocean floor; and Layer 3 (five kilometres [three miles] thick), consisting principally of gabbro, a coarse, crystalline, basic intrusic rock, a form of olivine basalt. Unlike the sediments that are added by deposition from above, these igneous rocks are all injected from below. They are derived by the partial melting of the primitive iron- and magnesium-rich rock of the upper mantle.

Relation of rocks and features to plate tectonics

The nature of the igneous crustal layers found below the sediments is in accord with the concept of sea-floor spreading (*q.v.*). As 100-kilometre- (60-mile-) thick lithospheric plates move apart at midocean rifts, mantle rock moves up from below to fill the void. Much of this injection takes place by viscous-solid flow, but there is also some molten rock present that is squeezed to the top. Outpouring of lava onto the sea floor is quickly chilled into bolster-shaped pods with glassy surfaces that are called pillow lavas (Layer 2). Beneath these pillow lavas are feeder dikes, along which the lava rose to the sea floor. With sea-floor spreading these dikes are pulled apart and injected with new dikes, forming a complex system of dikes (Layer 2b). These dikes are derived from an underlying layer of magma that cools slowly and crystallizes to form gabbro (the intrusive equivalent of olivine basalt) and associated suites of rocks.

This interpretation of the oceanic crust is not fully agreed upon by all authorities. According to this interpretation, however, all of the oceanic crustal rocks have the general composition of olivine basalt.

Major features. The major features within the ocean basins are the midocean ridges (rises), trenches, and fracture zones. These grand features are related, because they are all associated with the boundaries of the Earth's major crustal plates. These boundaries almost invariably lie within the ocean basins rather than on the continents. The trenches mark zones of underthrusting associated with the descent of the crust into the mantle; the midocean ridge forms along the pull-apart zones; and the fracture zones reflect lines of crustal disruption associated with great shearing action where one crustal plate has slid past its neighbour. Thus, the concept of plate tectonics (that the Earth's surface consists of a small number of crustal plates the boundaries of which are sites of major deformation) accounts for these three major types of features that dominate the ocean basins.

Midocean ridges. The midocean-ridge system actually forms a continuous swell throughout the world oceans with an overall length of 60,000 kilometres (37,000 miles), or more than the circumference of the Earth. It is a broad arch one to three kilometres high which may be either exceedingly rough or quite smooth. It is by far the largest and most extensive mountain range on the face of the Earth. Nearly everywhere it is deeply submerged, with only an occasional small island, such as Saint Helena or Ascension, marking its presence. Iceland alone is an exposed broad expanse of the spine of a midocean ridge. The expansion of rocks when hot apparently accounts in large part for the relief of the ridge above the normal level of the ocean floor. As the Earth's mantle loses this excess heat away from the ridge crest, the sea floor subsides to its normal depth, but this process of cooling takes several tens of millions of years.

This broad swell in the ocean floor forms a median ridge through the Atlantic and Indian oceans. This median position is almost precisely true for the Atlantic, where it faithfully follows the contours of the opposing continental slopes. The feature was first described and studied in the Atlantic and was appropriately termed the Mid-Atlantic Ridge. It is high, rugged, and marked by a prominent dorsal cleft, or rift valley. The ridge has a similar aspect in the Indian Ocean, although it is considerably more complex. Its overall form is like that of an inverted Y, but nevertheless the ridge maintains a roughly mid-ocean position relative to the surrounding continents.

The East Pacific Rise

The midocean-ridge system of the Atlantic and Indian oceans south of Australia joins with the East Pacific Rise (Albatross Cordillera). In the Pacific it is common practice to use the term rise rather than ridge, because, although this rise is simply an extension of the midocean-

ridge system of the other oceans, the Pacific swell has a smooth, low silhouette and generally lacks an axial rift valley. The East Pacific Rise also does not occupy a mid-ocean position but runs north–south, paralleling South America and intersecting North America at the mouth of the Gulf of California. This rise remains deeply submerged, and its presence is suggested by only a single volcanic island rising from its crest, Easter Island.

The difference in topographic expression between a ridge and a rise has been explained in terms of plate tectonics. Fast spreading, defined as the opening of the sea floor at a rate faster than six centimetres (two inches) per year (or three centimetres [one inch] per year on each spreading limb), as is characteristic of the Pacific, produces a rise. Slow spreading, on the other hand, results in the formation of a ridge. Sea-floor spreading is a symmetrical process that accretes new ocean floor equally to both flanks of a rift; when a former landmass splits apart, the ridge maintains a median position as the newly created ocean basin increases in size. This phenomenon occurred in the Atlantic and Indian oceans, but, in contrast, the rise in the Pacific did not rift a landmass when it was formed, and consequently there is no reason for it to be median.

Trenches. Oceanic trenches are long, narrow, arcuate depressions in the ocean floor. They occur principally around the periphery of the Pacific Basin, but examples are also found in both the Atlantic and Indian oceans. Individual trenches have lengths of thousands of kilometres, widths of roughly 100 kilometres (60 miles), and depths of two to four kilometres (one to two miles) below the adjacent ocean floor. Nearly all of the hadal regions, which are those deeper than 6,000 metres (20,000 feet), lie within trenches. Their continuity is remarkable—9,000-metre- (30,000-foot-) deep Tonga Trench is about five kilometres (three miles) wide, but it is continuous for 700 kilometres (400 miles). Typically, trenches have an asymmetrical V shape with a steeper slope toward land and a gentle slope toward the ocean basin; a low arch sometimes intervenes before the normal deep ocean floor is attained. Trenches and their associated island arcs, surmounted by explosive volcanoes, are the most active geological features on the face of the Earth. The great earthquakes and tsunamis (great sea waves produced by submarine earth movement or volcanic eruption) generated from them are invariably associated with trenches.

Greatest ocean depths

The greatest depths are found in trenches and so are near continental margins or island arcs rather than in the middle of the ocean basins. The deepest depression is the 10,850-metre (35,600-foot) Challenger Deep, discovered by HMS "Challenger II" in 1948, in the Mariana Trench not far from Guam. Some greater depths for this deep have been claimed, but they remain unsubstantiated. The oceanographers Jacques Piccard and Don Walsh descended to the bottom of the Challenger Deep in 1960 aboard the bathyscaphe "Trieste," a feat comparable to the ascent of Mt. Everest. Some other great deeps of the Pacific are Tonga Trench, 10,800 metres (35,400 feet); Kermadec Trench, 10,800 metres (35,400 feet); and Philippine Trench, 10,030 metres (32,900 feet). The greatest depth in the Atlantic Ocean is 9,200 metres (30,200 feet), found in the Puerto Rico Trench, just north of that island.

It is now commonly agreed that trenches are subduction zones—that is, zones where the outer, 100-kilometre- (60-mile-) thick outer shell of the Earth plunges into the mantle at angles from 30° to nearly vertical. The bending of this rigid crustal plate in adapting to the spherical geometry of the Earth is the cause of the arcuate shape of the trench. The descending lithospheric plate contains numerous rock types that are unstable in the regime of high pressures and temperatures that exists in the mantle. Hence, they cannot be consumed and are melted and returned to the surface as magmas and lavas that build up the arc behind the trenches. The Aleutians–Alaska Arc, which lies behind the Aleutian Islands, is a prime example; the Kuril and Mariana arcs are others (see further ISLAND ARCS).

Fracture zones. Since the early 1950s, bathymetric surveys have revealed a large number of horizontal lineations of high and rugged topography called fracture zones. These are long, narrow ridges and depressions that usually separate oceanic ridges of different depth. The fracture zones may be as much as 100 kilometres (60 miles) wide and 2,000 kilometres (1,000 miles) long. The first to be described was the Mendocino Fracture Zone, extending westward for 3,300 kilometres (2,100 miles) from Cape Mendocino, California. Subsequently, three other almost parallel extensive fracture zones have been surveyed off western North America—namely, the Murray, Molokai, and Clarion fracture zones. These appear to be scarps associated with offsets of a former extension of the East Pacific Rise, which was overridden by the westward drift of North America. In the Atlantic Basin there are also numerous fracture zones that offset the axial rift of the Mid-Atlantic Ridge. These fracture zones also extend far beyond the limits of the offsets and, in some cases, can be traced nearly to Africa or North America before their trace is lost beneath the thick-lying blanket of sediments along the continental margins.

Earthquake activity indicates that these fracture zones are actively shearing (moving) today only where they connect segments of actively spreading ridges or where they connect a rift to a trench. The extensions of the fracture zones onto the adjacent segments of the ocean floor are dead (immobile), although the rugged relief along their trends remains as evidence of earlier faulting and crustal slippage.

Mountains beneath the sea

Seamounts and guyots. A seamount is a mountain beneath the sea, generally in the form of an isolated, conical elevation of the sea floor at least one kilometre high. Seamounts are the most prominent and striking features on the ocean floor. More than 2,000 seamounts have been reported, and many more await future discovery. There remains no doubt that seamounts are nearly all volcanoes (mostly extinct), because when dredged the bedrock is always basalt, and their shapes and slopes are like that of a volcano on land. They are composed of alkaline basalts derived from depths of 150 kilometres (90 miles) or more within the deep portion of the upper mantle.

The Northeast Pacific Basin is especially rich in seamounts that commonly trend northwest to southeast in long festoons. Many of these chains are entirely submerged, such as the Magellan Seamount Group in the far western Pacific. Others, such as the Hawaiian chain, are mixed groups of islands, banks, and seamounts. Forming an extension of this chain is the giant, deeply submerged Emperor Seamount Chain off Japan. Each of these seamounts is named after a semi-mythical Japanese emperor.

Oceanic islands, those rising from the deep ocean bed beyond the continental shelves, may be classified as either high islands or low islands. High islands are simply the tops of giant seamounts that are both tall enough to pierce the surface and young enough not to have been eroded away by wave action. These are nearly all active, dormant, or recently extinct volcanoes, because erosion reduces an island to a shallow bank after a few million years. Strings of islands, such as the Hawaiian chain, may be formed when the Earth's outer crust drifts over a deep, stationary lava pipe. Thus, the volcanically active island always lies at one end of the chain. Low oceanic islands, those lying essentially at sea level, generally are coral atolls in tropical latitudes. As Darwin surmised about the mid-19th century, these atolls have formed by the deposition of limestone upon a subsiding, extinct volcano. The upward growth by corals and lime-secreting algae offsets subsidence and maintains the atoll precisely at sea level, so that these limestone edifices are "the gravestones of departed volcanic islands."

Some large, deeply submerged seamounts, especially in the central North Pacific, are flat-topped and are termed guyots, after Arnold Henry Guyot, a 19th-century Swiss–U.S. geologist. An especially large cluster of guyots is that of the Mid-Pacific Mountains, which stretch from west of Hawaii to Wake Island.

Because truncation of a seamount can occur only by wave action at sea level, guyots are thought to be

drowned ancient islands that have subsided one or more kilometres beneath the sea surface. They are found in regions unfavourable to coral growth, so that no atoll was built up to offset their subsidence. This sinking usually is largely a result of regional subsidence of the ocean floor and of the horizontal drift of the seamount as the ocean crust moves down the flank of a rise. Local subsidence, or foundering, caused by the load emplaced on the crust by a volcanic seamount, may also play a role, but sinking caused by a relative rise of sea level (addition of new water to the oceans) apparently is not important. Although the oceans probably are growing deeper with time as new water is squeezed out of the Earth's mantle, this deepening is exceedingly slow, probably not more than a few centimetres per million years.

Abyssal hills and plains. Hills and knolls on the ocean floor are termed abyssal; they are protuberances smaller than seamounts, rising to heights from a few tens to several hundred metres above the ocean floor in regions largely devoid of sediment. Extensive regions of chaotic roughness especially characterize the Pacific floor along the flanks of the midocean ridges. Where careful surveys have been made, these hills commonly display an elongate form. They are caused by faulting of the oceanic crust, volcanic extrusions, and other kinds of deformation.

Featureless plains on the sea floor

Extensive regions of the ocean floor are abyssal plains, flat, featureless, sedimentary plains with slopes of less than one part in 1,000. Over broad reaches these plains will not vary in depth by as much as one metre, and they are the levelest regions on the face of the Earth. This nearly perfect flatness is derived by the long-continued deposition of sediments by muddy bottom flows, which pond in the deepest hollows, burying any existing irregularities. These plains are found in all ocean basins but are best developed near continental margins and in the Atlantic Ocean, where deposition rates are high. Fine examples off the eastern United States are the Hatteras and Nares plains, lying at 5.5 kilometres (3.4 miles) depth, which have been developed by sediments shed from North America. The world's greatest abyssal plain is probably that underlying the Bay of Bengal, which has been built up by the muddy Ganges and Irrawaddy rivers.

Sediments of the ocean floor. The ocean floor is blanketed in most places with a sedimentary cover. Two basic types are recognized: namely, terrigenous sediments (sands, silts, and clays) that are shed from the continents or from islands and pelagic, or open-sea sediments—the finely suspended clays and remains of pelagic (floating-form) plants and animals that "rained" gently on the bottom.

The terrigenous sediments generally are deposited near the base of the continental slope as sedimentary fans or aprons. They are mostly turbidites laid down by turbidity currents (a type of density current in which the density contrast arises because of the high sediment concentration of the flow), and they attain great thickness. Although covering a larger area, the pelagic oozes form a thinner blanket (zero–two kilometres thick). The most common is globigerina ooze, composed of minute calcium carbonate shells of protozoans, mostly of the genus *Globigerina*. This ooze covers vast expanses of the Atlantic and Indian oceans. Calcium carbonate dissolves in the deeper portions of the oceans so that in much of the central Pacific Basin calcareous oozes are replaced by red clay. Diatom ooze, composed of opaline siliceous shells of marine algae, are found mainly beneath colder waters. A belt of diatom ooze cordons the world in the Antarctic region, and another zone lies across the far north Pacific. Another siliceous ooze, radiolarian ooze, is typical of tropical regions. A belt of this ooze extends across the Pacific beneath equatorial waters. It is composed of the minute remains of radiolarians, a pelagic protozoan (see further MARINE SEDIMENTS).

Basin boundaries. *Continental slopes.* The outer edge of the continental slope is marked by an abrupt brink where the sea floor plunges three to five kilometres (two to three miles) to join the abyssal floor. The continental slopes are the longest, highest, and straightest boundary walls in the world. Only the lofty Himalayan rampart facing India attains the scale of a continental slope. If the ocean waters were removed, the continents would stand as pedestals everywhere. The continental slopes are the margins of the continents and, hence, are the boundaries of the ocean basins. Their declivity is typically 3° to 5°.

Some continental slopes can be classified as accretionary, because they were created by oceanic crust underthrusting the continental margin. This origin applies to much of the Pacific margin. Other slopes are modified scarps or faults related to continental breakup and continental drift. This origin applies to much of the Atlantic and Indian ocean margins. Such slopes have commonly been extensively modified by sedimentation.

The continental rise. In many parts of the world the continental slope is separated from the abyssal ocean floor by a broad apron. This continental rise is the top of a prism of sediments shed from the continents and laid down mostly by turbidity currents on the deep ocean floor. Such rises are particularly characteristic of the Atlantic and Indian oceans. A particularly fine example is developed off the eastern United States, extending as a smoothly sloping apron for about 250 kilometres (150 miles) from the 1,000-fathom (2,000-metre [6,000-foot]) level to the deep abyssal plain. Geophysical investigations indicate that this continental-rise prism may attain a thickness of more than six kilometres. This huge sedimentary prism is regarded as a potential major oil province of the future. After initial deposition by turbidity currents, the sediments of the continental rise may be extensively reworked and are deposited by deep bottom currents. This condition is particularly true along the western sides of an ocean basin. The Blake Nose, a huge lens of sediment projecting outward from the continental rise off the south central United States, is an example of this bottom-current effect. Because the deep currents tend to move parallel to the bottom slope, they are called contour currents and their deposits contourites.

ORIGIN OF OCEAN BASINS

The new concepts of plate tectonics and sea-floor spreading have worked a revolution in the Earth sciences. Among other things, the concepts provide an adequate explanation for the origin of ocean basins. The theory holds that the Earth's outer shell is broken into about eight large, rigid, spherical caps, plus many small subplates. Except for the Pacific plate, which includes much of the Pacific Ocean, each major plate contains a separate continent embedded within it. An ideal plate may be envisioned as being rectilinear. Along one edge, where the plate is heavy, it dives into the Earth's mantle along a trench called a subduction zone. Opposing this zone of lithospheric descent, a rift is formed. As the rift grows larger, the void left behind is constantly healed by the upwelling of mantle rock that emplaces new ocean crust, a process known as sea-floor spreading. New ocean crust is presently being generated at about 1.5 square kilometres (0.6 square mile) per year, a rate sufficient to repave the entire ocean basin in 200,000,000 years. To accommodate this crustal movement, the rift and the trench are connected by large zones of shear or slippage called transform faults. Thus, crust is consumed at the trenches, created at the rifts, and conserved along the transform faults. The crustal plates and the continents embedded within them undergo drift, and this condition provides the mechanism for continental drift (*q.v.*). Typical drift rates range from one to several centimetres a year, a remarkably rapid geologic process.

Formation of ocean basins between the continents

To understand why the Earth has ocean basins, the origin of continents must be considered, because the ocean basins are simply those depressed crustal regions that lie between the isolated continents. The continental plateaus are slabs of granitic rock or sial (siliceous or acid igneous rock), and they literally float in the Earth's denser mantle like blocks of wood floating in water. Following the principle of Archimedes, the continents adjust themselves to a level at which the weight of the mantle rock displaced is equivalent to their own weight, which is called isostatic (equal-weight) equilibrium. The 35-kilometre (22-mile)

thickness of the continents and their density contrast with sima (mantle rock) is such that a five-kilometre (three-mile) relief results between oceanic and continental levels.

The generation of the granitoid rocks that form the continents requires at least two stages of melting and gravitational differentiation (separation according to density differences) of the low-density rock from the heavier rock. First, as the midocean rifts pull apart, new oceanic basalt fills in the void, repaving the ocean floor. This basalt is a partial melt or differentiate of the primitive, heavy, ultrabasic material of the Earth's mantle. Because of crustal drift this basalt is eventually consumed within trenches, subduction zones that carry the crust downward into the hot regime of the upper mantle. There the oceanic basalt is remelted, giving off sialic (lighter) rock that rises up to the surface as lava or granitic intrusions. This new rock generally forms an island arc that ultimately becomes accreted to the margin of a continent. In this manner the continents grow and the regions between these sialic plateaus become the ocean basins.

BIBLIOGRAPHY. The literature on the ocean basins is extensive and is growing at a rapid rate. Only a brief sampling of English-language publications can be included here. Some general nontechnical discussions about the ocean floor are: B.C. HEEZEN and C.D. HOLLISTER, *The Face of the Deep* (1971), a magnificent monograph of deep sea photos with explanatory text; F.P. SHEPARD, *The Earth Beneath the Sea* (1959), which emphasizes the geomorphology of the ocean floor; D.H. and M.P. TARLING, *Continental Drift* (1971), summarizes the now accepted concept of drifting continents in terms of plate tectonics; H.W. MENARD, *Anatomy of an Expedition* (1969), tells how a modern scientific expedition is organized to study a portion of the Pacific Basin. Some popular books that deal with the oceans generally are: J. DUGAN, *Man Under the Sea*, rev. ed. (1965), which concerns man's entry into the subsurface world; and J. PICCARD and R.S. DIETZ, *Seven Miles Down* (1961), a documentary account of the bathyscaphe "Trieste" and its ultimate dive to the bottom of the Challenger Deep. Some important textbooks and basic reference works include the following: The classic text on oceans generally is by H.U. SVERDRUP, M.W. JOHNSON, and R. FLEMING, *The Oceans* (1942). An excellent treatment of the Pacific is presented by H.W. MENARD in *Marine Geology of the Pacific* (1964). F.P. SHEPARD, *Submarine Geology*, 2nd ed. (1963); and P.H. KUENEN, *Marine Geology* (1950), discuss the ocean floor generally. Erosional forms are covered in F.P. SHEPARD and R.F. DILL, *Submarine Canyons and Other Sea Valleys* (1966). The mineral potential of the ocean floor is described in J.L. MERO, *Mineral Resources of the Sea* (1965). Some more general reference works are: R.W. FAIRBRIDGE (ed.), *Encyclopedia of Oceanography* (1966); and M.N. HILL and A.E. MAXWELL (eds.), *The Sea*, 4 vol. (1963–70). Charts of the coastal regions of the United States may be obtained from National Oceanic and Atmospheric Administration (NOAA), National Ocean Survey, Washington, D.C.; chart and bathymetric maps for other countries are issued by the Naval Oceanographic Office, Washington, D.C. Physiographic maps of the Atlantic, Pacific, Arctic, and Indian oceans may be obtained from the National Geographic Society; and original oceanographic data for selected regions is available from NOAA, Environmental Data Service, National Oceanographic Data Center, Rockville, Maryland.

(R.S.D.)

Ocean Currents

The great water masses that cover nearly 71 percent of the earth's surface are interconnected by a rather orderly system of ocean currents. The driving forces for this water motion are wind friction at the sea surface and horizontal and vertical differences in the density of seawater. Differential heating and cooling, precipitation, and evaporation produce differences of water density at the sea surface. Overturning of water masses, vertical convection, and mixing provide the mechanisms for distributing density differences from the sea surface into deeper layers. This creates horizontal density differences and, consequently, horizontal pressure differences and currents at all depths in the oceans. As far as the driving forces are concerned, a distinction is made between these two components of the general oceanic circulation (wind friction and differential density); they are not independent of each other, however.

Because the oceanic circulation is so closely linked to the atmosphere and its behaviour, the great oceanic current systems cannot be as stable or as steady as might be expected. The usual charts depicting ocean currents show only the prevailing current directions and speeds, not their variability. Such charts represent the average trend of water displacements and are comparable to climatic wind charts.

This article treats the nature, distribution, and causes of ocean currents and oceanic circulation, and oceanic–atmospheric interactions, including climatic effects. Related aspects of these topics are discussed in the articles WATER WAVES; TIDES; OCEANS AND SEAS; ATMOSPHERE; WINDS AND STORMS; and CLIMATE.

DISTRIBUTION OF OCEAN CURRENTS

Charts that show the horizontal distribution of ocean currents are based on a large number of direct and indirect current observations. Most of the data are obtained from ship drifts, however. If careful record of the course and speed of the ship is made then a "dead reckoning position" is established. This would be the true position if no drift by currents has affected the course and speed of the ship. The difference between dead reckoning and the actual position as determined by astronomical or electronic means indicates the average current. Whenever possible, these differences are determined at least 24 hours apart.

The general pattern of currents throughout the world's oceans is shown in Figure 1 for the Northern Hemisphere winter. More detailed maps are available for individual oceans or for parts of the oceans, and the adjacent seas. Figure 1 shows that the distribution of currents in the upper strata of the oceans tends to coincide with the prevailing winds of the world. It should be recognized, however, that this smooth, average representation of the oceanic circulation only outlines the general trend of horizontal water movements.

Vertical components of currents

Because ocean currents are three-dimensional water displacements, the vertical components of currents are also of interest. In general, vertical speeds are much smaller than horizontal speeds. Nevertheless, vertical motions are important for an accurate description and explanation of the oceanic circulation system. Significant vertical branches in the current structure occur mainly where horizontal currents either diverge or converge. The phenomenon of upwelling water in the surface strata of the oceans is caused by diverging surface currents in the open ocean, or near the coasts of continents where a supply of water from deeper layers is required to compensate for the seaward motion of surface currents. The formation and spreading of water masses in the deep sea also depends upon vertical branches in the current system for explanation.

THE CAUSES OF OCEAN CURRENTS

The equations of motion in hydrodynamics make use of one of Newton's fundamental laws of mechanics when applied to a continuous volume of water. They state that the product of mass and current acceleration equals the vector sum of all forces that act upon the mass.

Besides the external force of gravity, the most important forces that cause and affect ocean currents are horizontal pressure gradient forces, the Coriolis forces, and frictional forces.

Pressure gradients. The hydrostatic sea pressure, p, at any depth below the sea surface is given by the product $\rho = g\,\rho\,z$, where g is the acceleration of gravity, ρ is the density of seawater, and z the depth below the sea surface. Because differences in density at any fixed depth are due to differences in temperature and salinity, the sea pressure, p, will vary correspondingly from region to region. This part of the total internal pressure field in the oceans is called the relative field of pressure.

In a homogeneous ocean of constant potential density, horizontal pressure differences are only possible if the sea surface is tilted against level surfaces. Level surfaces are everywhere perpendicular to the force of gravity (the plumb line). In this case, surfaces of equal pressure, called

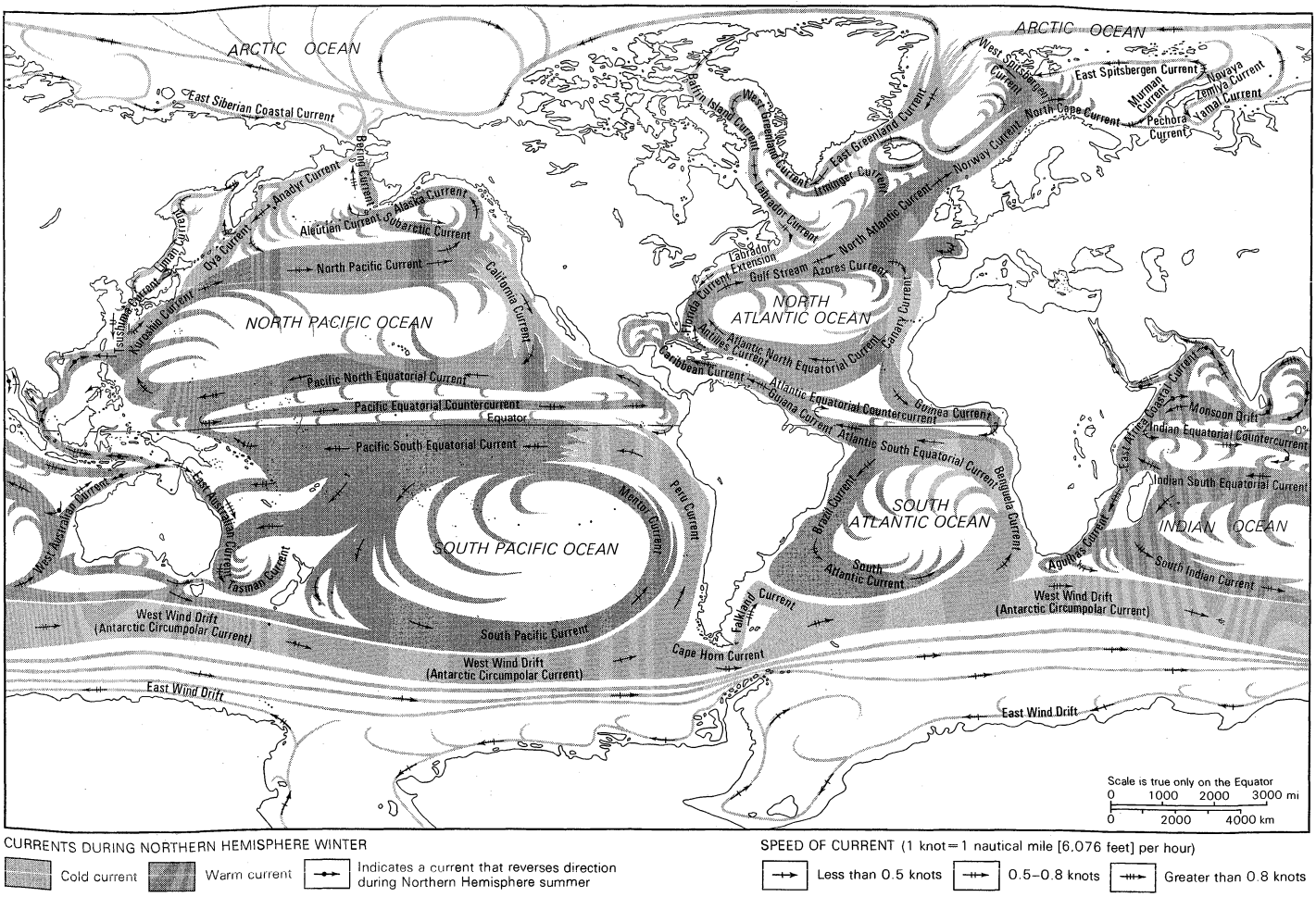


Figure 1: Major ocean current systems of the world.

isobaric surfaces, are also tilted against level surfaces in the deeper layers. This part of the internal pressure field is called the slope field. Thus, the total oceanic pressure field is composed of two parts: the relative field and the slope field. The slope field of pressure can be the result of the wind, for example, when wind-driven surface currents pile up water against a coastline or between opposing winds in the open sea.

The rate of change of pressure, Δp, along a horizontal distance, Δn, is called the horizontal pressure gradient, $\Delta p/\Delta n$. Although much smaller than the vertical pressure gradient, $\Delta p/\Delta z$, the horizontal pressure gradient nevertheless is most important with respect to ocean currents.

Coriolis effect. To an observer in space a moving body would continue to move in a straight line unless the motion were acted upon by some other force. To an earthbound observer, however, this motion cannot be along a straight line because the reference system of coordinates is fixed to the rotating earth. This is similar to the effect that would be experienced by an observer standing on a large turntable if an object moved over the turntable in a straight line. An apparent deflection of the path of the moving object would be seen. If the turntable rotated clockwise, this apparent deflection would be to the left, relative to an observer looking in the direction of the motion.

Because the earth rotates from west to east about its axis, an observer in the Northern Hemisphere would notice a deflection of any moving body to the right. In the Southern Hemisphere this deflection would be to the left. At the Equator there would be no apparent horizontal deflection but only a vertical deflection for zonal currents. This most remarkable effect of the earth's rotation, the Coriolis effect, is named after Gaspard Coriolis, a 19th-century French mathematician.

It can be shown that the Coriolis force always acts perpendicular to the motion. Its horizontal component, CF, is proportional to the sine of the geographical latitude and the speed, c, of the moving body. It is given by $CF = 2\omega c \sin \varphi c$, where ω (7.29×10^{-5} radian per second) is the angular velocity of the earth's rotation. Coriolis forces affect all motions on or near the earth's surface, including missiles in flight, a Foucault pendulum, air currents, and ocean currents.

Frictional forces. A faster moving fluid layer tends to drag along a slower moving layer. A slower moving layer will tend to reduce the speed of a faster moving layer. This is the result of momentum transfer between layers of different velocity. Also, the wind blowing over the sea surface transfers momentum to the water and causes a drift of water. This frictional drag at the sea surface (the wind stress) is transmitted to deeper layers by friction within the water. Together with the Coriolis force this friction causes pure wind-driven currents.

Geostrophic currents

A simple equation of motion can be derived for the case where friction is insignificant, the currents are nonaccelerated, and the driving forces for ocean currents are due to horizontal pressure gradients only.

On a nonrotating earth, water would be accelerated by the pressure gradient and would flow, as in a river, from higher to lower level, or from high pressure to low pressure. On a rotating earth, however, the Coriolis force deflects the motion, and the acceleration ceases only when the speed, c, of the current in a given latitude is just fast enough to produce a Coriolis force that can exactly balance the pressure gradient force. Therefore: $\Delta p/\Delta n = 2\rho\omega c \sin \varphi$. From this balance, it follows that the current direction must be perpendicular to the pressure gradient, because the Coriolis force always acts perpendicular to the motion. In the Northern Hemisphere this direction is such that the high pressure is to the right when looking in

current direction, in the Southern Hemisphere it is to the left. This type of current is called a geostrophic current. If the pressure gradients are the result of differences in density of ocean water, this simple equation provides the basis for an indirect method of computation of ocean currents.

If $\Delta p/\Delta n$ is known, the velocity, c, of the resultant current can be computed. Because $\Delta p/\Delta n = g\rho(\Delta z/\Delta n)$, where $\Delta z/\Delta n$ is the slope of an isobaric surface, the speed, c, of the current is obtained from $c = g(\Delta z/\Delta n)/2\omega \sin \varphi$.

Across the Gulf Stream, the slope of the sea surface over a distance of 100 kilometres is about one metre, thus $\Delta z/\Delta n = 10^{-5}$. With $g = 9.8$ metres per second squared and $2\omega \sin \varphi = 10^{-4}$ in middle latitudes, it follows that $c = 0.98$ metre per second, or about one metre per second. This is a reasonable average surface speed of the Gulf Stream. Because the Gulf Stream often is less wide and somewhat "streaky," however, its speed in some parts of its course can be higher.

Wind-driven currents

The frictional drag of the wind at the water surface, and between layers of water, sets ocean water in motion. If a steady wind blows long enough over a large ocean, pure wind-driven currents are the result. They show some remarkable properties. The currents do not follow the wind direction, but are deflected at the sea surface by 45° to the right in the Northern Hemisphere, and to the left in the Southern Hemisphere. With increasing depth this deflection increases while the current speed decreases. At a certain depth, roughly around 75 or 100 metres in middle latitudes, the current even flows against the surface current; however, its speed has reduced to about $\frac{1}{23}$ of its surface speed.

In 1902 the Swedish oceanographer V. Walfrid Ekman deduced these results in a theoretical model. He considered a homogeneous ocean where frictional forces are balanced by Coriolis forces, and where resulting currents are nonaccelerated. His results show that the end points of the current vectors from surface to bottom, when vertically projected on a horizontal plane, produce a logarithmic spiral, now termed the Ekman spiral.

Pure wind-driven currents of this type do not penetrate deeply into the ocean. Essentially, they occupy a surface layer called the Ekman layer, about 100 metres deep. The average total water transport in the Ekman layer is 90° to the right of the wind direction in the Northern Hemisphere, to the left in the Southern Hemisphere, and proportional to the wind stress. In the ocean these drift currents must either converge or diverge because they will meet coasts or, in the open ocean, winds of opposing direction and speed. In regions of surface water convergence, water is not only piled up but also is pushed down. In regions of divergence the sea surface is lowered and deeper water rises to the surface. This produces slopes of the sea surface and slope currents. Superposition of pure wind-driven currents and slope currents can result in a complicated current system. If density differences contribute to a relative pressure field in addition to a slope field, ocean currents are further modified.

Since 1946 considerable progress has been made in the study of the wind-driven ocean circulation and the thermohaline circulation produced by density differences. One of the most modern theoretical models that tries to take into account all of the driving forces, wind-driven and thermohaline, is the numerical model by Kirk Bryan and Doak Carey Cox (1967). It has succeeded in explaining many of the major features of the general oceanic circulation, for the wind-driven surface strata and for the deep sea. Laboratory models such as those conducted by William Sterling von Arx that employ rotating water basins have helped to substantiate theoretical results on the dynamics of ocean currents.

GENERAL SURFACE CIRCULATION

The system of surface currents. The circulation of the upper strata of the oceans divides into gyres that rotate either clockwise or anticlockwise (Figure 1). A clockwise rotation in the Northern Hemisphere with higher pressure in the centre of rotation is called anticyclonic. An anticlockwise rotation with lower pressure in its centre is called cyclonic. In the Southern Hemisphere the sense of rotation is opposite, because the effect of the Coriolis force has changed its sign of deflection.

This system of surface currents corresponds closely to the average climatological wind system over the oceans. It is, however, affected by the interference of coastlines with respect to the winds.

Westward intensification of currents

The tendency of ocean currents in big, elongated, anticyclonic gyres is to displace their centres of rotation to the west, and to form strong western boundary currents like the Gulf Stream and the Kuroshio. This westward intensification of ocean currents was explained by Henry Melson Stommel (1948) as a result of the fact that the horizontal Coriolis force increases with increasing latitude.

Walter Heinrich Munk (1950) and others explained theoretically many of the major features of the wind-driven ocean circulation by using the mean climatological wind stress distribution at the sea surface as a driving force.

Effects of bottom topography or of a restricted vertical extent of wind-driven currents by strong vertical density gradients in the oceans have not been considered in these models, however. These additional effects may help to explain why the westward intensification of ocean currents is not equally strong in the Southern Hemisphere.

Western and eastern boundary currents. Individual branches of the surface current pattern are indicated by their geographical names in Figure 1. The poleward flow on the western side of the great subtropical gyres in the North Atlantic and North Pacific, the western boundary currents, is fast, narrow, and deep; whereas the return flow toward the Equator, the eastern boundary currents, is slower, broader, and less deep. In the Southern Hemisphere, however, western boundary currents like the Brazil Current and the East Australian Current are not as strongly developed as their Northern Hemisphere counterparts.

Western and eastern boundary currents associated with subtropical gyres in the world's oceans are listed in the Table.

Western and Eastern Boundary Currents of the World's Oceans

boundary currents	Atlantic Ocean		Pacific Ocean		Indian Ocean
	north	south	north	south	
Western	Gulf Stream	Brazil Current	Kuroshio	East Australian Current	Somali Current
					Mozambique Current
					Agulhas Current
Eastern	Canary Current	Benguela Current	California Current	Peru Current	West Australian Current

Gulf Stream and Kuroshio currents

The best explored western boundary currents are the Gulf Stream and the Kuroshio. The Gulf Stream system is composed of the Florida Current, the Gulf Stream proper between Cape Hatteras and the Grand Banks, and the North Atlantic Current. The corresponding divisions of the Kuroshio system are the Kuroshio, south of 35° N, the Kuroshio Extension, and, farther east, the North Equatorial Current.

The Gulf Stream and Kuroshio are usually sharply defined by an oceanic front along their continental side that marks a transition between the warm and more saline water masses of these currents and the cooler, less saline water of coastal regions. This front, often called the cold wall, is distinguished by a sudden change in temperature and sea colour.

Recent observations have shown that both currents are much narrower and faster than previously supposed. Sometimes they break up into two or more streaks that are considerably variable in their position and speed, often forming meanders and eddies that may detach from the main current.

Eastern boundary currents are relatively shallow and often indicate the presence of numerous eddies and countercurrents. These currents are closely connected with the process of upwelling water in coastal regions. Outstanding areas for this phenomenon are the coasts of northwest Africa, southwest Africa, California, and northern Chile and Peru. The upwelling water is usually restricted to the upper layer of 200 metres or 300 metres depth.

The equatorial current system. The circulation of the upper ocean strata in tropical and subtropical regions is dominated by the North and South Equatorial currents, Equatorial countercurrents, Equatorial undercurrents, and in some parts by Monsoon currents. Pronounced seasonal variations of the equatorial current system are the result of strong seasonal changes of the wind system, particularly in the Indian Ocean. This wind system is governed essentially by the trade winds and in some areas by monsoon winds. Between the fairly steady trade winds of both hemispheres a more or less broad belt of light variable winds or calms develops, often called the doldrums.

From G. Neumann, "Oceanography of the Tropical Atlantic," *Anais da Academia Brasileira de Ciencias*, vol. 37, Supplement, pp. 63–82 (1965)

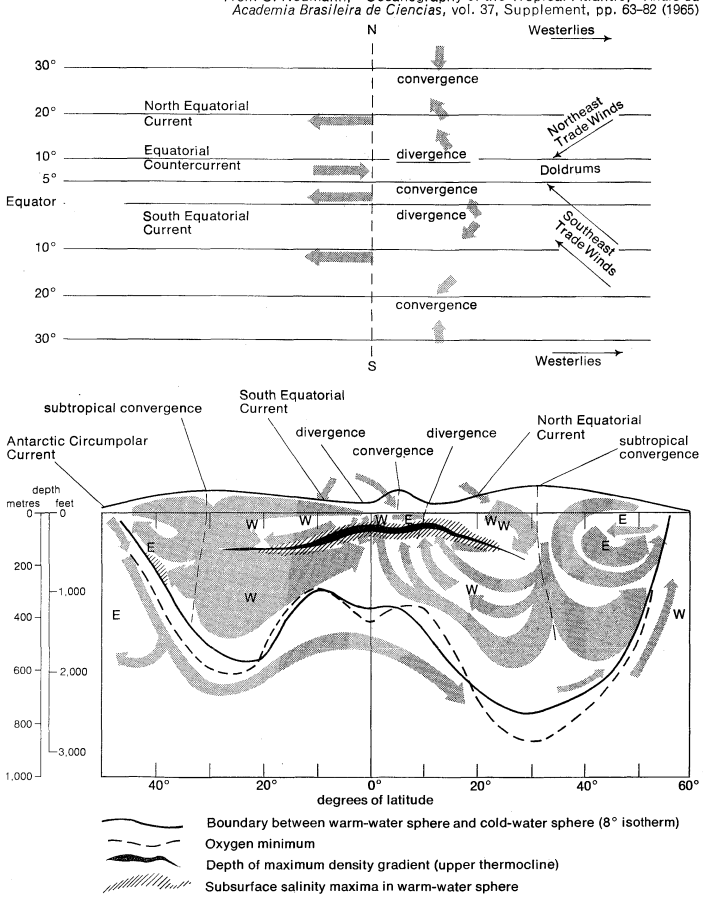


Figure 2: (Top) Explanation of equatorial and subtropical current system. (Bottom) Cross-sectional view of warm-water sphere (equatorial and subtropical regions) showing meridional circulation by arrows and zonal currents by W and E, respectively. Main features of stratification as related to currents are also shown.

Equatorial countercurrents

Among the outstanding branches of the equatorial current system are the Equatorial countercurrents. The most impressive is the North Pacific Equatorial Countercurrent. The Equatorial Countercurrent of the Atlantic is fully developed only during the Northern Hemisphere summer when it extends from about 50° W longitude into the inner Gulf of Guinea. During winter (Figure 1), it occupies only the eastern part of the tropical Atlantic.

In the Indian Ocean, during the time of the northeast monsoon (Northern Hemisphere winter), a strong countercurrent is found between about 2° S and 8° S, embedded in the westward flowing North and South Equatorial currents. During this season, the Somali Current, near the African coast, flows southward, where at about 7° S it feeds part of its water into the countercurrent (see Figure 1). During the Indian southwest monsoon (Northern Hemisphere summer), however, currents have completely reversed their direction in the tropical western and northern part of the Indian Ocean. The Equatorial Countercurrent has disappeared and the Somali Current has reversed its direction and flows northeastward, sometimes even faster than the Gulf Stream. This remarkable current now feeds its water masses into the eastward flowing Southwest Monsoon Current in the northern Indian Ocean.

The doldrums

The locations of the Equatorial countercurrents are closely related to the locations of the doldrums. A qualitative explanation of the tropical current system for the case where doldrums occupy latitudes between 5° N and 10° N is as follows. Wind-driven water transport in the Ekman layer leads to convergences and divergences of surface water between the major winds and to a piling up or depression of the sea surface, respectively. The resulting slopes of the sea surface in a meridional section from about 25° N to 10° N are downward. Between 10° N and 5° N it must be upward because of the convergence at 5° N. Near the Equator is a region of divergence and, therefore, the sea surface must rise from the Equator not only toward 25° S but also toward 5° N.

These slopes of the sea surface produce horizontal pressure gradients and slope currents that superpose the pure wind-driven currents. Because the resultant currents must always flow in such a direction that the higher pressure is to the right in the Northern Hemisphere and to the left in the Southern Hemisphere, westward flowing currents must occur between 25° N and 10° N, and between 5° N and 25° S. These currents represent the North and South Equatorial currents. The surface slope in the doldrums belt between 10° N and 5° N, however, requires a narrow, eastward flowing current, the Equatorial Countercurrent.

Equatorial Undercurrent and the Cromwell Current

The Equatorial Undercurrent is another extremely narrow, eastward setting current branch in the equatorial current system. It is centred on the Equator with a width of about 200 to 300 kilometres and a vertical extent of 200 to 300 metres. Most often, it does not reach the sea surface and is, therefore, not represented in surface current charts. Maximum eastward speeds of 100–150 centimetres per second may occur at depths between about 50 and 150 metres, while at the sea surface westward setting currents are observed. This current was first extensively studied in the Pacific Ocean where it is named the Cromwell Current. During recent years, oceanographic work in equatorial regions has established the worldwide existence of this remarkable current. In 1960 G. Neumann pointed out evidence for an Equatorial Undercurrent in the Atlantic Ocean.

The Antarctic Circumpolar Current. Currents around Antarctica flow mainly from west to east. Only in a narrow zone close to the continent are westward flowing currents observed (Figure 1). Both currents are largely a response to the prevailing winds.

The great eastward flow is called the Antarctic Circumpolar Current. In contrast to the wind-driven currents in tropical and subtropical regions, this current is deep and in some parts of the ocean it reaches the bottom at 3,000–5,000 metres depth. Therefore, its course is strongly affected by bottom topographic features, like submarine ridges. The surface speeds are rather small (15–20 centimetres per second), but as a result of the great depth, the water volume transported exceeds that of any other oceanic current system. V.G. Kort (1962) estimated the Antarctic Current transport to be 150,000,000 and 190,000,000 cubic metres per second. This is about twice the rate of transport of the Gulf Stream. The Antarctic Circumpolar Current constitutes an important link between the Atlantic, Indian, and Pacific Oceans, with respect to the oceanic deep-sea circulation and stratification.

DEEP-SEA CIRCULATION

Knowledge of the deep-sea circulation is essentially based on indirect methods. The German oceanographers Albert J.M. Defant and George A.O. Wüst used such methods extensively to analyze the deep-sea currents in the Atlantic

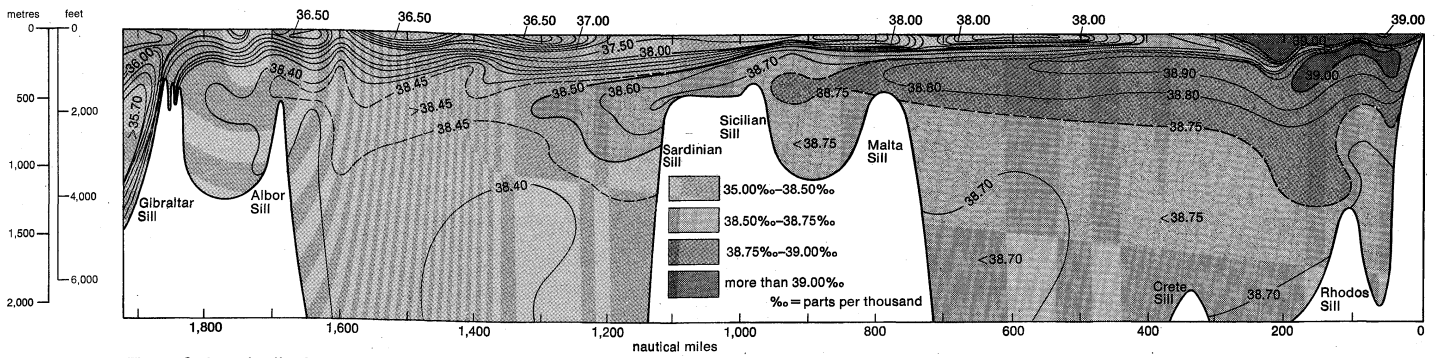


Figure 3: Longitudinal sections of summer salinity in the Mediterranean Sea along the axis of the Levantine Intermediate Water, showing the origin of the water that forms the so-called Upper Deep Water in the Atlantic Ocean.
From Wust (1961) in G. Neumann and W.J. Pierson, Jr., *Principles of Physical Oceanography* (1966); Prentice-Hall, Inc.

Ocean. The distribution of temperature, salinity, oxygen, and other chemical factors have provided much information on currents and water mass transport throughout the oceans. Some direct current measurements from the deep sea also have been used and it has been concluded that the deep-sea circulation is essentially a thermohaline circulation that depends on differences of temperature and salinity. It is not independent of the wind-driven circulation, however.

Influence of highly saline waters

Compared to the Pacific and Indian oceans, the vertical structure of temperature, salinity, and oxygen in the Atlantic is more differentiated. This is because the exchange of water with the Arctic Ocean is many times greater than that of the Pacific. Moreover, the Atlantic is bordered by a number of adjacent seas, among which the Mediterranean Sea is the most important. Small, but continuous, intrusions of Mediterranean water of high salinity and relatively high temperature through the Strait of Gibraltar create a special "climate" in the Atlantic deep sea.

The Mediterranean Sea, in general, is located in an arid climate zone. Evaporation exceeds precipitation, and this causes an increase of surface salinity, especially in its eastern parts. This high salinity water near Asia Minor sinks into subsurface layers because of its higher density. From there, it moves with subsurface currents at about 500 metres depth westward and, finally, flows through the Strait of Gibraltar into the Atlantic Ocean. This Mediterranean water sinks quickly to about 1,500 to 2,500 metres depth and spreads out into the Atlantic. It can be traced across the Equator as far south as the Antarctic Ocean, where it is entrained in the Antarctic Circumpolar Current. Along its path this water mass, termed Upper Deep Water, gradually decreases in salinity while mixing with less saline water above and below its level of spreading. In the Indian Ocean a similar water mass originates in the Red Sea and the Persian Gulf, whereas it is missing in the Pacific Ocean.

A schematic block diagram (Figure 4) for the western part of the Atlantic Ocean summarizes knowledge of its surface currents and deep-sea circulation. The spreading of the Mediterranean water from about 30° N and 1,500 metres depth is indicated.

Middle and Lower Deep Water masses

Other important deep-sea movements from north to south are associated with the spreading of Middle and Lower Deep Water masses. These water masses originate near Greenland. They are cooler (3°–4° C) than the Upper Deep Water and can be traced by their higher oxygen content while spreading southward between about 2,500 and 4,000 metres depth. Wüst estimated current speeds of 10–15 centimetres per second at about 3,000 metres depth in western equatorial regions. There seems to be no source of cold deep water in the North Pacific.

Deepwater movements from north to south in the Atlantic are sandwiched between water masses that spread northward with relatively low salinities (less than 34.85‰ [parts per thousand]; see Figure 4). The upper water mass is called the Antarctic Intermediate Water. It is formed at the sea surface near the Polar Front, or Antarctic Convergence, around 50° S. From there it spreads northward at about 700 or 1,000 metres depth, mainly in the western part of the Atlantic, where it is characterized by low salinities. It can be traced as far north as 25°–30° N.

Antarctic Bottom Water

The deepest water mass of all, Antarctic Bottom Water, originates on the shelf of the Antarctic continent. Because of its high density, it slides down to the ocean bottom and then spreads northward and eastward. Its movements are channelled by submarine ridges and troughs, and it spreads mainly along the western half of the Atlantic (West Atlantic Trough), crossing the Equator while passing some of its water through a narrow pass (the Romanche Trench) in the Mid-Atlantic Ridge. Average velocities of 10–15 centimetres per second near South America have been obtained from geostrophic current computations. Bottom-water current speeds can be more than twice as great on occasion, however. Relatively strong bottom-current speeds, up to 60 centimetres per second, have been inferred by marine geologists from ripple marks, scour marks, and rock outcrops beneath the Antarctic Bottom Water, in the Drake Passage south of Cape Horn, and in other parts of the oceans, including the Pacific Ocean.

OBSERVATION OF OCEAN CURRENTS

Meaningful current measurements require knowledge of a fixed point in the open ocean or accurate determination of the geographical position from which the measurements are made. The fixed point can be an anchored ship, a moored buoy, or a platform construction. Submerged deep-sea diving vessels like the bathyscaphe "Trieste II" also can be used to make current observations near the sea floor while the ship has settled on the bottom.

Methods of measuring ocean currents can be divided into two basic groups. The Eulerian method measures the speed and direction of currents at fixed points by means of current meters. If observations of this kind are available for many points, current charts can be constructed that depict the instantaneous movement of the water in a given area. The current can be shown by a current vector at each observation point that indicates direction and speed. The combination of many of such current vectors in a chart gives a streamline representation of currents, where at each point the current vector is tangent to the streamline.

Drift measurements

The second group of observational methods is termed Lagrangian; basically, the path of a drifting water parcel is followed over a longer time interval. The traced path of a water particle or of a freely floating object is called a trajectory, and the combination of many of such trajectories permits construction of a trajectory chart, in distinction to the streamline chart obtained by the Eulerian method. Only when currents are stationary, where they do not change with time at a given place, will trajectory charts agree with streamline charts.

Trajectories of ocean currents usually represent only horizontal trajectories, and the vertical motion of water is neglected. This is because the observations are taken by means of drifting objects, called tracers, and these are bound to a certain water layer and, therefore, do not follow exactly the three-dimensional path of drifting water.

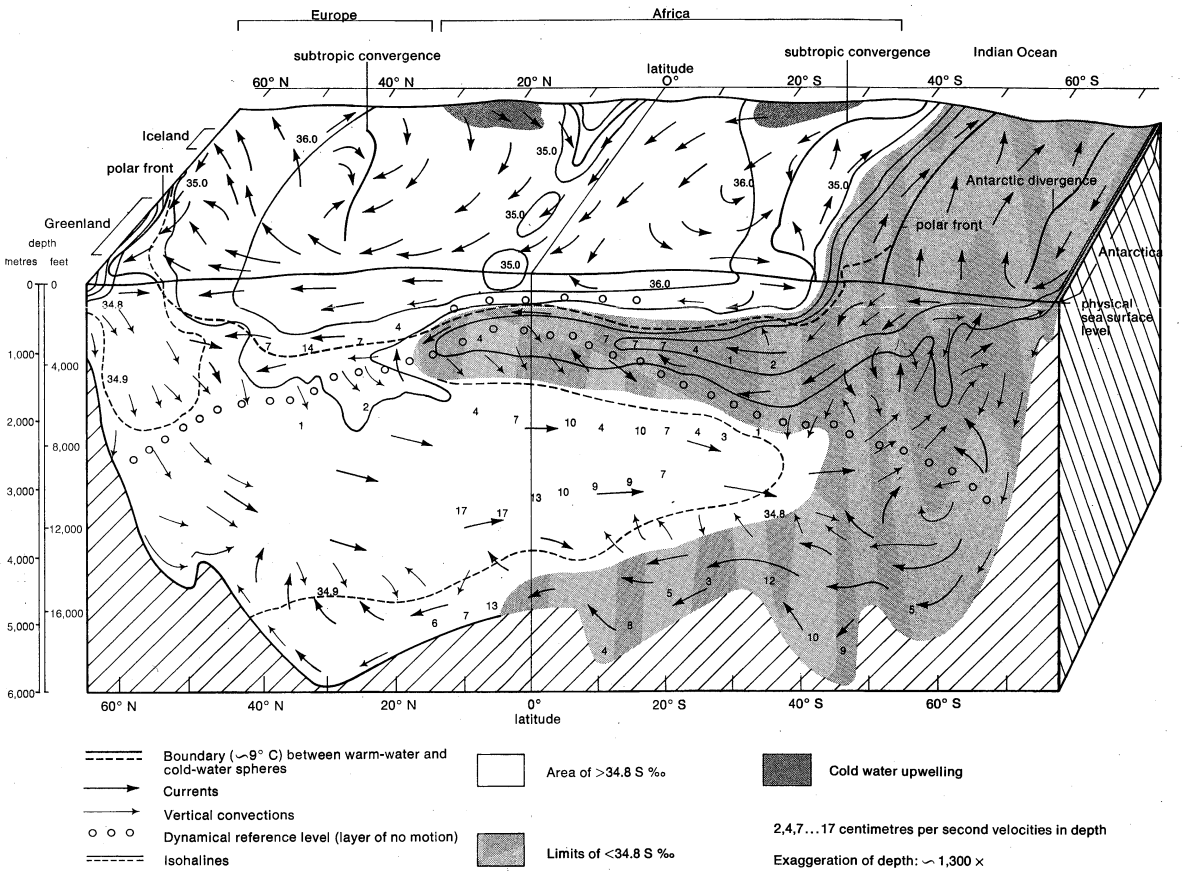


Figure 4: Essential aspects of deep-sea circulation and spreading of water masses in the Atlantic Ocean.
From Wust (1961) in G. Neumann and W.J. Pierson, Jr., *Principles of Physical Oceanography* (1966); Prentice-Hall, Inc.

The oldest method of drift measurements involves the use of drift bottles or drift cards. Together with ship drift observations, where the ship itself is used as a tracer, these simple methods have helped to describe many of the main features of the surface circulation in the oceans. Because drift measurements require the continuous tracing of freely floating objects along their path, the obvious disadvantage of the method is that in many cases only the end point of drift can be related to the starting point. Modern radio buoys can be traced continuously by radio direction-finding receivers from shore stations, ships, or aircraft.

Deep drogues to measure the path of moving water can be either canvas, wood, or metal crosses or sheets that offer the largest possible drag at the level of current measurement. Parachutes that are designed to open at a prescribed depth of measurement have been widely used. These deep drogues are attached to a small surface float that can be followed by an attending ship.

The first successful subsurface current drift tracer without surface connection was developed and used in 1955 by J.C. Swallow. It is a neutral buoyancy float, properly balanced to float freely at a predetermined level, where the weight of the displaced water equals the weight of the float. The subsurface float acts as an acoustical source, and sends signals to the surface ship. These signals or "pings" can be followed by a surface ship equipped with hydrophones suspended into the sea.

Electromagnetic and temperature–salinity methods

If seawater, which is an electric conductor, moves in the earth's magnetic field, an electric field is induced in the water. The intensity of this electric field can be measured by a pair of electrodes in tandem, about 100 metres apart, towed behind a ship. The electric potential difference between the two electrodes, e, is recorded on board the ship by means of a potentiometer. The velocity, v, of the moving water is proportional to e, and $v = ke$, where k is a factor that depends on the distance between the electrodes, the specific electrical resistance of seawater (and of the bottom in shallow water), and on the vertical intensity of the geomagnetic field. Manoeuvring the ship by 90° changes in course provides the necessary data for determining current speed and direction. The advantage of this method is that it does not require a fixed point for making current measurements.

The velocity of ocean currents also can be obtained from measurements of the temperature and salinity of seawater from surface to bottom. Knowledge of the distribution of temperature and salinity permits computation of the density of seawater. The relative pressure field that provides the driving force for ocean currents can be derived from the density field. If the pressure field in the oceans is known, certain hydrodynamic relationships can be used to compute the speed and direction of currents.

OCEAN CURRENTS, CLIMATE, AND WEATHER

The mutual influence of atmosphere and ocean manifests itself in climate and weather phenomena, as well as in the circulation and stratification of the oceans. Ocean currents are largely responsible for the temperature distribution at the sea surface, which, in turn, affects the overlying atmosphere. The difference between the temperature of the air and water basically determines evaporation rates, atmospheric humidity, precipitation, and the vertical stability of air masses moving over the ocean surface. As air cools when flowing over a cold ocean current, fog is likely to form. When cool air flows over a warm current, the air is heated from below and may rise. This leads to condensation of water vapour in the rising air, cloud formation, and often precipitation. Such atmospheric convection processes can sometimes be violent, forming towering cumulonimbus clouds and thunderstorms.

Interaction of atmosphere and ocean

Meteorological consequences of the interaction of atmosphere and ocean are obvious in regions of upwelling water near the west coasts of the continents. The upwelling water is accompanied by relatively cool sea surface temperatures. Because the air is cooled and stabilized over these regions, the climate is cool, with haze or fog, but

little or no precipitation. From the biological point of view, these coastal regions offer a striking contrast between land and sea. Off the coast of northern Chile and southern Peru, for example, the land is barren, dry, and almost wholly infertile (Atacama desert); the sea, however, teems with sea life of every type. Its cool waters provide some of the world's best fishing grounds, because water from deeper layers with a high nutrient content is brought up to the sea surface by the upwelling process.

The oceanic circulation also supports the atmospheric circulation in equalizing temperature differences between the tropics and the poles. The oceans can store great amounts of heat because of the great specific heat of water. The heat stored in tropical regions is transported over great distances, especially by western boundary currents. It is released to the atmosphere in higher latitudes chiefly as latent heat of evaporation. The fact that northwest Europe has such an agreeable climate is ultimately the result of the Gulf Stream system, a warm-water heating system that is supported by, and mutually related to, a warm-air heating system that is provided by the prevailing winds. In the Pacific Ocean, the Kuroshio Current system plays a similar role with respect to the climate of the northwestern part of North America.

Disturbances of the oceanic current system and its associated heat transport can produce far-reaching effects that are reflected in climatological anomalies; droughts or floods, abnormally cold and severe winters, or extremely mild ones are examples. Attempts at long range weather forecasting (*q.v.*) depend largely on the general oceanic circulation, its variation through time, and its effect on physical conditions at the air–sea interface.

BIBLIOGRAPHY. General textbooks, with extensive information and bibliographies on ocean currents, include: H.U. SVERDRUP, M.W. JOHNSON, and H.R. FLEMING, *The Oceans* (1942); A. DEFANT, *Physical Oceanography*, 2 vol. (1961); W.S. VON ARX, *Introduction to Physical Oceanography* (1962); G. DIETRICH and K. KALLE, *General Oceanography* (1963); H.J. MCLELLAN, *Elements of Physical Oceanography* (1965); and G. NEUMANN and W.J. PIERSON, JR., *Principles of Physical Oceanography* (1966).

Specialized textbooks on currents are H. STOMMEL, *The Gulf Stream: A Physical and Dynamical Description*, 2nd ed. (1965); L.M. FOMIN, *The Dynamic Method in Oceanography* (1964); and G. NEUMANN, *Ocean Currents* (1968), with an extensive bibliography.

Information on this subject may also be found in M.N. HILL (ed.), *The Sea*, vol. 1 and 2 (1962–63), containing technical articles on the dynamics of ocean currents, current measurements, and special current systems; R.W. FAIRBRIDGE (ed.), *Encyclopedia of Oceanography* (1966); and in R.W. STEWART, "The Atmosphere and the Ocean," *Scient. Am.*, 221:76–86 (1969).

(G.N.)

Oceania, History of

Oceania, in its broadest sense, comprises most of the insular territories of the Pacific Ocean, but conventionally comprises four divisions: (1) the island continent of Australia; (2) Melanesia, from New Guinea eastward to Fiji; (3) Micronesia, principally the Mariana, Caroline, Marshall, and Gilbert archipelagoes; and (4) Polynesia, a triangle with many islands, such as Hawaii, Tahiti, Samoa, Easter Island, and New Zealand. In this article, the history of Oceania is limited to Melanesia, Micronesia, and Polynesia. (For Australia and New Zealand see AUSTRALIA, HISTORY OF; NEW ZEALAND, HISTORY OF.) These names were coined in the second quarter of the 19th century by the French explorer Jules Dumont D'Urville from the Greek noun *nēsos* ("island") and the adjectives *melas* ("black"), *mikros* ("small"), and *polys* ("many"). But the history of Oceania began many thousands of years before this European classification, before even the first European sightings of the islands in the 16th century. Prehistory, the period before written records survive, began in Melanesia at least 20,000 years ago, and the migrations and voyages that led to the populating of every habitable Pacific island extended from that time until the first quarter of the 2nd millennium AD. After European discovery, which extended from the early 16th century until the end of the 18th, the prehistoric world of Oceania was transformed, at first by casual visitors, then by more permanent ones, such as beachcombers and castaways, and then, at the end of the 18th century, by missionaries. From the 19th century the European impact was accelerated by missionaries, traders, settlers, and by the penetration of European governments into the area. By the 20th century, Oceania had been completely involved in the European world, even as European governments began to withdraw from the region.

Historiographical problems. Both the prehistory and the history of Oceania present problems. Prehistory is primarily dependent for its reconstruction upon archaeology. This has scarcely begun in Oceania, so that any conclusions must be tentative and subject to modification with each new site that is examined. Linguistics, upon which the prehistory also rests, is somewhat more advanced, but here, too, much remains to be done. A broad picture can be sketched, but the detail is still missing. Indeed, prehistory with a precise detailed chronology may never be possible. The period of Oceanic history for which documentary evidence is available also presents difficulties. The documents are chiefly of European origin and are, therefore, the products of people who may not accurately have recorded a culture different from their own—a culture they perceived and understood only imperfectly. This distortion can be corrected to some extent by using the findings of social anthropology and the oral traditions of the Oceanic people, but these sources are difficult because, by definition, they are contemporary. They may not describe the past accurately because they serve contemporary purposes; they do not record the past for its own sake. But the main historiographical problem of Oceania is its diversity. Some 10,000 islands scattered across 750,000 square miles of ocean, a great variety of cultures, hundreds of mutually unintelligible languages, and very diverse historical experiences make it very hard to generalize.

Geography and prehistory. The physical diversity means that contact between islands has never been easy because of the distances involved; it also means that the physical environment is not only isolated but varied. The large continental islands of Melanesia—such as New Guinea, New Caledonia, and Viti Levu—not only have a different physical basis of life and a wide variation in climate and fertility but their rugged terrain has made for social isolation. The smaller volcanic high islands, such as Samoa and Tahiti, have greater homogeneity and, although lacking the mineral resources of the continental islands, are very fertile, supporting a life well above the subsistence level. The coral atolls, the low islands of Oceania, support a much narrower range of vegetation; they are more exposed to bad weather and support an existence much closer to bare subsistence, except for the rich marine resources of their lagoons.

Physical environment does not determine the kind of society that exists, but it does set limits to it. The large islands of Melanesia produce marked differences between people of the coast and those of the interior. Their long coastlines have acted as a filter for many different arrivals in Oceania. The valleys have perpetuated differences. Thus Melanesia is characterized by many small groups of people, divided from each other by language and custom. Political and social organization has been small in scale. The margin above subsistence has not been great enough to allow for elaborate ceremonial. The high volcanic islands of Polynesia offered no such barriers to social and political unity. Their fertility allowed the development of elaborate social, religious, and political ceremonial. The low islands also allowed homogeneity and wide social groupings, but their land resources offered no great margin for ceremonial life except as they were supplemented by skill as fishermen. These contrasts within Oceania are obvious, as they were to early European visitors, but they conceal a similarity: whether small, with leadership a matter of acquiring influence rather than hereditary position, or larger, with chiefs surrounded with awe and reverence, Oceanic societies all rested on the principle of reciprocity. Every gift or service had to be returned.

Ethnography

The diversity of Oceania not only conceals a basic similarity but it has also misled historical enquiry. The useful division of Oceania into Melanesia, Micronesia, and Polynesia has prompted a search for different origins of the people of those areas. They have been assumed to be different races—Negrito, Mongoloid, Caucasoid. They have been traced as different waves of migration out of Southeast Asia; and the Polynesians have been given, by at least one authority, a South American origin that is denied by the linguistic and botanical evidence. These earlier theories of waves of immigrants from Malaysia and elsewhere are very doubtful in the light of the archaeology. What recent archaeology suggests is that the threefold division happened in Oceania itself: it is no evidence of different origins in some area outside Oceania. New Guinea was certainly settled at least 20,000 years ago, presumably by people from Southeast Asia. But the earliest dates for the settlement of the rest of Oceania are as yet no older than the 2nd millennium BC, and this settlement extended into the Christian Era. Linguistic evidence supports this. People on the islands of Oceania speak languages of the Austronesian stock, which includes Indonesia, Malaysia, the Philippines, and the aboriginals of Taiwan and Madagascar. Languages of non-Austronesian type are found only in small numbers in the interior of the islands of New Guinea and its close neighbours.

The broad picture that the archaeological work suggests is, therefore, that there was an early human settlement in New Guinea, perhaps 25,000 years ago. Then, at some later date, there was a new complex of settlement, which was not that of hunters and gatherers but of horticulturalists who had domesticated animals. This complex was taken, fully formed, with the first settlers of other Pacific islands, together with a ground-stone technology. The pottery called Lapita ware (from the site at which it was first systematically examined in New Caledonia) indicates the track of these settlers. It has been found in New Britain, near Rabaul, and in the New Hebrides, in Fiji, and in Tonga. Its makers had settled in Fiji by 1000 BC and in Tonga by 500 BC. They were the formative influence in what has been classified as Polynesian culture. In the Marianas, in Micronesia, a related type of pottery goes back to the 2nd millennium BC. It seems, therefore, that what is distinguished as Polynesian was filtered through Melanesia to western Micronesia, and to western Polynesia, where it developed distinctive characteristics that were then taken by the original settlers to eastern Polynesia, where the Marquesas were settled about AD 500. On this evidence, the three divisions of Oceania that so struck the first Europeans originated within the islands themselves. And, in the case of some of the eastern Polynesian islands, settlement antedated that of Europeans by only a few hundred years.

Neolithic technology

Oceanic societies, at the time of European discovery, were Neolithic. They had developed a technology based on stone, bone, and shell; they cultivated tubers and tree fruits, all of which were of Southeast Asian origin, with the exception of the sweet potato, which derives from South America, and which culture had spread through most of Polynesia in pre-European times, but only marginally into Melanesia. This Neolithic cultivation was associated with three domesticated animals: the pig, dog, and chicken, also of Asian ancestry. The coastal people had developed techniques of fishing and considerable skills as sailors; although the fabulous voyages once ascribed to the Polynesians are doubtful, navigation between the closer islands was well-developed. At the same time, some skills were lost: pottery, for example, disappeared in the Samoas and the Marquesas shortly after initial settlement. And with the control of the environment that Oceanic technology offered, there is evidence of overpopulation, settlement spreading into less-favoured areas of the islands and being fortified, as in the Marquesas. (For a discussion of the cultures of the indigenous peoples, see OCEANIAN PEOPLES AND CULTURES; for further description of the geographic setting, see PACIFIC ISLANDS.)

European exploration. The Oceanic world was not a static unchanging one, but changes were slow compared with those that attended European discovery. Vasco Núñez de Balboa discovered the Pacific in 1513; seven years later Ferdinand Magellan rounded South America and sailed across the ocean, missing the main island groups but discovering one of the Tuamotu Islands and Guam; after his death in the Philippines, his expedition discovered some of the Carolines. These northern islands were further explored by the Spaniards as they established a galley trade from Manila to Acapulco (see PHILIPPINES, HISTORY OF THE). The next major Spanish discoveries were made by Álvaro de Mendaña de Neira and Pedro Fernández de Quirós. In 1567 the former set out from Peru to discover the great southern continent, which was believed to exist in the South Pacific. He found the Solomons but failed to find them again on his second journey, during which he died. In 1606 his chief pilot, Quirós, after discovering some of the Tuamotu Archipelago, found the northern Cooks, Tikopia, and the New Hebrides. One of his companions, Váez de Torres, found southeastern New Guinea and then the strait (later named for him) between that island and Australia, although the discovery was unknown to later sailors. These Spanish expeditions were motivated by the search for riches; by zeal to extend religion; and, in the case of Quirós, discovery for its own sake. But with his voyage, the Spanish effort was ended. Thereafter, the Dutch, who were already established in Indonesia, entered the Pacific. They too looked for a southern continent. In 1615–16 the Dutch navigator Jacques Le Maire came from the east through the Tuamotus to discover Tonga, the Bismarck Archipelago, New Ireland, and New Hanover. In 1642 Abel Janszoon Tasman, sailing from Batavia (now Djakarta, Indonesia), the Dutch headquarters in the East Indies, discovered New Zealand, Tonga, some of the Fijis, and New Britain. The Dutch were primarily interested in commerce; they found none. Tasman thought that New Zealand was part of the great southern continent. The effect of these visitors on Oceania was transitory. They stayed for periods of at most a few months. Their contacts with the islanders were those of simple barter, but the demands they made upon food supplies often caused hostilities in which some European and many islanders' lives were lost, as on Guadalcanal and in the Marquesas during Mendaña's visits.

During the early 18th century the extent of Oceania was further revealed. The English pirate William Dampier visited New Hanover, New Britain, and New Ireland in command of a Royal Navy ship. Dampier was the forerunner of scientific exploration, and he proved that those islands were separated from each other and from Australia. In 1722 the Dutch admiral Jacob Roggeveen crossed the Pacific from east to west on a voyage of exploration that had also commercial objects. He found Easter Island, more of the Tuamotu Archipelago, the northern islands of the Society group, and some of the Samoan islands. These voyages were not essentially different from earlier ones, but they foreshadowed the scientific interest of the later 18th century. The execution of that interest was delayed by European wars. But in 1765 the English admiral John Byron (grandfather of the poet), who was sent by the British Admiralty in search of the supposed southern continent, found more of the Tuamotus and the southern Gilberts. In 1767 Samuel Wallis and Philip Carteret followed, but their ships were separated as they entered the Pacific. Wallis discovered Tahiti, more of the Tuamotus, and the Society Islands, while Carteret found Pitcairn and rediscovered the Solomons of Mendaña, although he did not so identify them. This was left to the French following Louis-Antoine de Bougainville's visit in 1768, during which he also discovered some of the New Hebrides and Rossell Island in the Louisiade Archipelago.

These explorers were more important for the knowledge they spread in Europe than for most of their discoveries. Dampier's *New Voyage round the world* and Bougainville's description of the noble savage in Tahiti were particularly influential. The English captains had not carried out their precise instructions, but the interest their jour-

nies created was in part responsible for the instructions given to the greatest of all 18th-century explorers of Oceania, James Cook. In three voyages he left others little to do but fill in occasional details of Oceania. Cook was sent in 1769 to observe the transit of Venus at Tahiti and then to search for the great southern continent. Cook found some of the Society Islands, but he also circumnavigated New Zealand, and he defined the limits of eastern Australia. During his second voyage (1772–75), he proved that there was no southern continent, but he also made further discoveries in Oceania: in the Tuamotus, the Cooks, the Marquesas, Fiji, New Caledonia, the New Hebrides, and Norfolk Island. His third voyage (1776–79) was mainly concerned with the North Pacific, during which he found some of the Tongan group, Christmas Island, and Hawaii between. He had completed the main work of discovery with an exactitude hitherto unknown. Although his contacts with islanders were not in essence different from his predecessors, his relations with them were nevertheless more prolonged and more humane. And his exploration of eastern Australia, through the account of his naturalist, Joseph Banks, was of great importance in Oceania, for it led to European settlement close to the islands.

Importance of Cook's travels

Traders and vagabonds. With the establishment in 1788 of the Australian settlement, Oceania became a source of supply. In 1793 pigs from Tahiti were landed at Sydney, and until 1826 the trade, although subject to fluctuations due to the competition of other cargoes and Tahitian wars, was important. The competition among Europeans for sandalwood, pearl shell, and bêche-de-mer (sea cucumbers), valuable cargoes that attracted ships from the Australian colony, further involved Oceania with the European world. In 1804 sandalwood was found in Fiji and for the next ten years attracted European traders. The sealing industry drew men to New Zealand, just as the fur traders, in the 1790s, wintered in Hawaii. All of these contacts began to affect Oceanic societies because they were sustained and prolonged. Together with the castaways and beachcombers who had begun to live in the islands from the days of first European contact but who increased in numbers with commercial shipping, these European contacts began to transform Oceania.

The "Bounty" mutineers

Castaways, such as the HMS "Bounty" mutineers who went to Tahiti in 1789, began with their muskets to alter the political state of the islands they lived in by the support they gave to the chiefs who befriended them.

Missionary and trading societies. Such Europeans had an important effect, but they were dependent upon the people they lived with for their survival. Very different were the first Europeans to come to Oceania permanently and with the deliberate intention of changing Oceanic society: the missionaries. In 1797 the London Missionary Society (LMS), now the Congregational Council for World Mission, sent a party to Tahiti. After some vicissitudes it converted the chief Pomare I who controlled the area of Matavai Bay, where Europeans had called since Wallis' discovery. The LMS failed in its first attempts at Tonga and the Marquesas, although it was more successful in Huahine, the Tuamotus, the Cooks, and later in Samoa. Other missionary societies followed. In 1822 the Methodists began to work in Tonga; in 1835 they went to Fiji. In 1843 Roman Catholic missionaries began working in New Caledonia. The Church of England began to penetrate into Oceania from New Zealand in the 1840s. These missionaries encountered societies in Polynesia that already had a problem of law and order from the effects of European beachcombers and traders. Its solution was to create missionary kingdoms, in the case of British missionaries, or the establishment of direct political control, in the case of the French.

In Tahiti, Hawaii, and Tonga, native chiefs whose power was established by their access to European arms and support not only became kings but took missionary advisers and missionary-designed codes of law. In 1819 the native ruler Pomare II of Tahiti promulgated such a code. In Tonga, Taufa 'ahau took the name of George in 1833, and in 1845, when he took the Tongan title of Tu'i Kanokupolu, he became "king" of Tonga; in 1862, under the influence of the Rev. Shirley Baker, he adopted a constitution. By attempting to enforce a scriptural code of law, these missionary kingdoms were an answer to the problems of European lawlessness in the islands. If the missionaries could not prevent the sale of arms, they could at least make sure that these passed into the hands of friendly chiefs. But the authority of these "kings" was challenged from two sides. By becoming Christian they had cut themselves off from the mana (a Polynesian religious concept sometimes described as an all-pervasive energy) that came from the old gods, and this produced heathen reactions. In Tahiti there was a revolt against the new Christian order by supporters of the old ways in 1830; in Tonga there was a similar reaction in 1831. In Samoa, where the holder of the Malietoa title had embraced Christianity from Tahitian missionaries, there were heretical movements. If heathen beliefs thus resisted the chiefs and their missionary supporters, the European traders also resisted the political authority of the kings. Dissidents and heretics looked to these Europeans for leadership, and these Europeans looked to their own national governments for protection.

In Melanesia the story is somewhat different. In Fiji, the missionaries who landed in 1835, accompanied by an envoy from George of Tonga, made no headway with the rising chief Cakobau, who was not converted until 1854, when his fortunes were at a low ebb and he needed Tongan support. Elsewhere in Melanesia, the absence of chiefs meant that missionary work had to be conducted with small groups of people and repeated every few miles. There was no wholesale conversion of the kind that had happened in Polynesia. The attempt of the London Missionary Society in the New Hebrides in the 1840s came to nothing. The Anglican Melanesian mission in the Solomons made slow progress in the 1850s. In New Guinea, mission work, divided into four spheres of influence in Papua, did not begin systematically until the 1870s. Micronesia was a backwater. The Spaniards had established missionaries in the Marianas in 1668, but the missionaries in the Carolines were killed in 1733. The main effort came from the Hawaiian Evangelical Mission in the 1850s. The general effect of mission activity was, nevertheless, the same as it was in Polynesia. It dissolved the old ties of society by attacking the supernatural sanctions that supported leadership and social mores. It altered the political structure of Oceanic societies. It incidentally introduced both European goods and the desire for them. The missionaries themselves acted as intermediaries between Oceanic societies and other Europeans—as political advisers, as agents, and as interpreters.

Growth of trading communities

Beachcombers and castaways preceded missionaries in many of the islands, but the growth of trading communities was in part the result of the missionaries' work in restraining native violence. Those traders were initially pork traders in Tahiti, but European captains followed valuable cargoes from island to island. When the supply of sandalwood was depleted in Fiji by 1813, the traders then found it in Hawaii in the 1820s, in the New Hebrides in 1825, and in New Caledonia in 1840. Pearl shell attracted traders to the Tuamotus in 1807. The sandalwood trade declined as supplies were exhausted. Pearling declined as natives took reprisals for the atrocities that had accompanied both trades. The demands of the Oceanians also changed the character of trade. Once native polities were established, the demand for muskets fell off; under missionary influence, the demand for alcohol was limited. What the islanders now wanted was clothing and hardware. The exchange traders were not guilty of the cruelties as those looking for sandalwood or pearls were apt to perpetrate to obtain their cargoes, and exchange trading encouraged the growth of resident agents in the islands, a development that met the needs of the whalers who came ashore to refit their vessels. After 1840 it also met the needs of the staple trade of the islands—coconut oil. Copra trading, from which the oil came, became the mainstay of European trade because even islands that had no other resources had coconut palms. In the early 1840s new techniques enabled the oil to be used for soap and candles. Such commerce promoted the growth of the port

towns and of resident trading communities. Papeete in Tahiti, Apia in Samoa, and Levuka in Fiji became European centres, including not only respectable traders but also lawless men who might be escaped convicts from New South Wales or others escaping from the rules of settled societies. These were frontier towns and could be regulated only with difficulty by native kings or by visiting European captains. Law and order became a problem with which the Oceanic institutions were unable to cope and with which captains could deal only intermittently.

European settlement. The problem became more urgent with the advent of European settlers. In Fiji, for example, after Cakobau's first offer of cession of the islands to Great Britain in 1858, Europeans came to establish plantations, at first of coconuts, then, during the U.S. Civil War, for cotton and afterward sugar. The development in Samoa was similar. But settlers needed land on a much larger scale than traders, and they needed labour in much greater quantities to work the plantations. Both land sales and labour recruitment caused friction, for "ownership" was not an Oceanic concept; thus, land titles were disputed or resented, and the recruitment of labour often caused the breakup of Oceanic societies if too many males left their communities and the creation of foreign racial communities if they did not. The settlers were a more permanent element in Oceanic societies. By 1870 they numbered 2,000 in Fiji. Politically the settlers had an interest in stability, and economically they needed security of title to land and a supply of labour. Neither requirement was satisfied by the missionary kingdoms. Nor was it satisfied by native governments that were not missionary guided. In Tahiti, in Tonga, in Samoa, and in Fiji no native authority was able to keep order in the novel circumstances created by European enterprise; in any case, these native kings were themselves open to challenge within their own societies. Pomare encountered revolt in Tahiti, Samoan politics were always a matter of rivalry between chiefs, and Cakobau's government was threatened by the Tongan chief Ma'afu, who had established his own confederacy in the Lau Islands.

Establishment of plantation societies

Such internal conditions in Oceania began to draw in European governments, all of which acknowledged some responsibility for the protection of their nationals and their property. The French government was the first to intervene, after the expulsion of two Catholic missionaries from Tahiti in 1835. In the following year two more were deported from Hawaii. In 1839 the Archbishop of Chalcedon suggested regular association between the Catholic missions and the French navy, but the French government was also aware of the need for a good naval station for the fleet and for French commerce and for a place of penal settlement. Abel DuPetit-Thouars thus took possession of Tuahata and the southeast Marquesas in 1842 and in the same year persuaded the Tahitians to ask for a French protectorate, which was formally accepted in 1843. In 1853 the presence of French missionaries in New Caledonia led to French annexation, possibly for fear of British action, certainly to establish a penal colony (to which convicts were transported until 1897). Other European nations intervened for different reasons. In 1857 August Unselm, as agent for J.C. Godeffroy und Sohn, set up the company's depot at Apia, and Samoa became the greatest trading centre in the islands; and even when Godeffroys failed in 1879, the Deutsche Handel und Plantagengesellschaft took over, and Samoa remained the favourite colony of the colonial party in German politics. British nationals had trading and plantation interests in the islands; to give some protection to these interests, the British government had appointed consuls to those islands governed by recognizable rulers, but their powers to maintain order were limited and, except for the visits of warships, unenforceable. The United States also appointed consuls. The rivalry between these officers, between European entrepreneurs, and the involvement of both in the internal politics of Oceanic societies merely emphasized to metropolitan governments the disordered condition of the islands. In Tahiti the problem was resolved by French annexation. In Samoa, after a tripartite supervision set up by the Samoa Act of 1889 came to grief in European rivalries and Samoan factionalism over chieftainships, an agreement of 1899 divided the Samoa group between Germany and the United States; Britain received compensation elsewhere. Britain's main concern was in fact with the activity of its nationals: in Fiji, where it accepted the offer of cession of 1874, it did so primarily because native authority had broken down. But Britain also had been concerned with the labour trade by which the Queensland plantations took islanders, who were sometimes recruited under doubtful or brutal conditions. In the 1860s this trade flourished in the New Hebrides, and there was violence which led the missionaries to protest. Then the labour trade moved north to the Solomons, where again there was violence, including the murder of the Anglican bishop in the Santa Cruz group, from which five men had been taken by recruiters. The British solution was the Western Pacific Order in Council (1877), which empowered the governor of Fiji to exercise authority over British nationals and vessels in a wide area of the western Pacific. The problem still remained, however, of non-British nationals in islands that had neither native kings nor European governors, especially those of Melanesia.

Intervention of European governments

Recruitment of labour

European government, like both mission and commercial enterprise, had been slower to penetrate Melanesia. Missionary activity did not begin in New Guinea until 1873. There was not much labour recruiting. The first main activity was the gold rush of 1877, but German traders had come to the northern coast in 1873 followed by the firm of Hernsheim in 1875. Such foreign interest produced a demand in the Australian colonies for annexation for reasons quite unconnected with the internal situation in the islands. German interests were marked in Micronesia, French in the New Hebrides. A number of groups in Australia also looked on New Guinea as a rich possession. But the British government, notwithstanding Queensland's abortive attempt at annexation in 1883, would not annex unless the Australian colonies paid the cost of administration, the same argument it was applying to New Zealand's interest in the Cook Islands. When the Australian colonies agreed to pay, the British government acted. Southeast New Guinea was declared a protectorate in 1884 and annexed four years later; the Cooks became a protectorate in 1888 and were annexed in 1901. Germany annexed northeast New Guinea in 1884, including some of the Solomons, over the rest of which the British established a protectorate in 1893. In Micronesia, the Germans, after an attempt to annex the Spanish possession of the Carolines in 1885, finally bought them from Spain with the Palaus and the Marianas in 1899, having annexed the Marshalls in 1885 and, under a convention with Britain of 1886, the phosphate island of Nauru. By that convention Britain's interest in the Gilberts was recognized, although no protectorate was declared until 1892. The Ellices were added in the same year. The process of partition was completed by the New Hebrides. A convention of 1887 set up a mixed British and French naval commission, but its authority was limited, and in 1906 a Condominium was agreed upon by which such difficult legal questions as land title were settled and a joint administration set up.

Patterns of colonial government. With the exception of Tonga, which remained an independent kingdom under British protection (from 1900) with a consul who was not to interfere in internal affairs, the whole of Oceania passed under the control of European powers and the United States between 1842 and the end of the century. Having acquired colonies, these powers governed Oceania with metropolitan institutions modified to a degree by local circumstances. Thus Britain reproduced in the islands the pattern of crown colony government, which derived from its own political development: a governor who represented the king; an executive council of senior officials; and, where the European population justified it, a legislative council to advise the governor. Within this form of government, administration was adapted to local conditions. Thus, Sir Arthur Gordon, the first governor of Fiji, set up a system of native administration that incorporated

the chiefs: the island was divided into provinces and districts that, on the information available to him, represented the old divisions of Fiji, and over each he tried to select the chief to take administrative office. Even in Melanesia, where chieftainship was not highly developed, the British attempted to appoint chiefs who were men of influence. The first administrator of British New Guinea was a former officer in Gordon's government, William MacGregor, who first tried appointed chiefs and then settled for village constables. The Australians, who took over British New Guinea in 1906 and rechristened it Papua, followed the British pattern. The first Australian governor, Sir Hubert Murray, although he introduced measures of native development, still preserved the British pattern of colonial government, as did New Zealand in the Cook Islands. Beneath the governor, district administration tried to incorporate both native leadership and the technical and professional direction of specialized departments (such as agriculture and health), as well as those departments that dealt with the questions raised by European settlement (such as land and labour).

The Germans in Samoa

Other nations had different patterns. The Germans, if only because domestic German politics made colonies an incident of European policy, tried to administer their colonies through commercial companies. In New Guinea, the Neu Guinea Kompagnie was commissioned to administer the colony as a commercial enterprise. Only when it failed did the imperial government assume responsibility (1899). In the Marshalls, the German firms known as the Jaluitgesellschaft became a chartered company under a government commissioner in 1885. In Samoa, in the first decade of the 20th century, the governor Wilhelm Solf attempted to control the importation of Chinese labour for the plantations and tried to enlist Samoan interest for the government, but he was subject to the pressure commercial interests were able to exert in Germany itself. In the French territories, colonial rule meant assimilation to French institutions. The governor was analogous to the prefect of a French *département*, assisted by an administrative council and from time to time by a general council drawn from French citizens. Where such a council existed, its powers were limited to an optional section of the budget, the rest of which was obligatory. In effect, the governor ruled by administrative decree. In the United States territories, there was also a marked assimilation to United States metropolitan forms of government. When Hawaii was annexed in 1898, the president of the republic became a U.S. governor. When eastern Samoa was given to the United States under the convention of 1899, President William McKinley, in 1900, placed it under the authority of the Department of the Navy, the commanding officer of the station also becoming governor and administering the islands with the help of his technical officers and the advice of a Samoan fono, or legislature. These colonial governments were adapted to local circumstances. In the Polynesian islands and in Fiji, Britain and Germany attempted to incorporate the authority of the chiefs into their governments, both as advisers and as local officials in the districts, as did the United States in its part of Samoa. But in both Hawaii and Tahiti the old system of rank had broken down under the impact of missionaries, traders, and settlers, so that it could not be used for administrative purposes but had to be replaced by appointed local officials. In Melanesia, where there was in general no chiefly authority that could be used because influence was acquired by criteria that made it fleeting, the colonial powers had no choice but to use appointed local headmen. The Germans and the British did this in New Guinea and the Solomons, but the real instrument of government was the patrol by European officers with an escort of armed native police, who enforced peace.

The impact of World War I. This pattern of colonial rule in Oceania was altered by external circumstances. With the outbreak of World War I in 1914, an Australian force took German New Guinea, and a New Zealand force took German Samoa. Japan took the Carolines, the Marshalls, the Palaus, and the Marianas. At the end of the war these German territories were retained by the occupying powers as mandates under the League of Nations. The professed aim of colonial rule was to help the people of these territories to stand on their own feet under the strenuous conditions of the modern world.

This stress on Oceanic interests was not new, but it was now an international standard. Still, the first step was the establishment of government control and of law and order before any other measures could be taken. In New Guinea, in the Solomons, and in many parts of Melanesia, the interior was often unknown, let alone controlled. So the attention of colonial government was concentrated on this. The discovery in 1933 of the grass valleys of New Guinea presented the Australian administration with the problem of 750,000 new people. Thus in the Australian territories the resources for doing anything beyond this initial control were spread thinly. Health and education were left to the missions. By contrast, Fiji was more developed. With the importation of Indian labour to work the sugar plantations from 1879, and with the reforms in the indenture system (which came after a commission of 1909), the large Indian population of Fiji received the attention of the government in education and health matters, but the increase in this population raised the difficult question of the Fijians' future. They played a minor part in the economic life of the colony, and the trend of official policy was to preserve them within their villages under a separate system of administration, which was reorganized in 1944 as the Fijian Administration—to be virtually a state within a state. But the government's resources were not enough to introduce welfare measures and to promote development on any great scale. A good deal depended on the missions and other private organizations. In Samoa, where the old society had retained its organization, there was resistance to change. The New Zealand administration of Samoa had begun with the objective of promoting the welfare of the native race, which meant health, education, and better use of the land. By recognizing Samoan councils it tried to ensure Samoan support, but the policies broke down in execution. In American Samoa, the navy provided welfare services as part of its routine work but could do little more than that when its principal concern had to be the smooth running of the naval base. In French Polynesia native policy aimed at making the people French citizens. The welfare services were directed from Paris, but they were limited by the resources available.

Problems of health and welfare

Oceania since World War II. Such limited resources and the competition between different objectives of colonial policy plainly restricted what could be done. Mission resources were obviously limited. Moreover, they were affected by external events such as the Great Depression and the fluctuations in world markets for copra, sugar, and other products of Oceania. The principal achievements were to check population decline by control of the introduced European diseases that had ravaged the islands and by increasing control of endemic diseases (such as malaria in Melanesia) and to hold a rough balance between European and indigenous interests. But welfare policies and island administration were both interrupted by World War II. The Japanese had been established in the north of Oceania, where they had treated their mandates as part of Japan itself. In 1941 they advanced into the rest of Oceania, reaching and controlling most of New Guinea as well as much of the Solomons at the peak of their advance. New Caledonia, the New Hebrides, Fiji, and Polynesia were not occupied but the effects of the war made colonial government secondary to military operations. After the war, the Trusteeship Council of the United Nations replaced the mandates, but all of the colonial powers accepted that independence or self-government was the aim of their rule. The Oceanians themselves had been exposed to a more intensive European (and Japanese) impact; their horizons had widened. The colonial powers felt a greater urgency to promote development and to make available greater resources to achieve it.

Politically, colonial governments were reorganized to give indigenous people a part in government. In Western Samoa, the Legislative Council was given, in 1947, a

Beginning of self-government

Samoan majority and considerable powers. In American Samoa, naval rule was replaced in 1951 by civilian control, and a legislature of two houses was set up, which by 1960 became a lawmaking body of Samoans. In French Polynesia and New Caledonia, elected assemblies were given considerable local autonomy in 1956; both territories opted to stay within the French community in 1958. In Fiji the Legislative Council was reformed, in 1962 and again in 1966, to provide an unofficial majority; but elections, in view of the racial composition of Fiji, were still to be communal. In Papua and New Guinea, administered by Australia as a single territory after World War II, an elective representative assembly was introduced in 1964. The following year the Cook Islands were granted internal self-government by New Zealand. In Micronesia, the Trust Territory of the Pacific Islands, composed of the former Japanese mandates administered by the United States, the advisory congress of two houses was set up in 1964 when the islands ceased to be a strategic concern, and in 1967 a Future Political Status Commission began to discuss its relations with the United States. The Gilbert and Ellice islands received a constitution in 1967 that provided for the Governing Council and a House of Representatives, the latter having advisory powers.

The pace of political development in Oceania thus speeded up. In 1962 Western Samoa became independent; in 1968 the phosphate island of Nauru; in 1970 Fiji and Tonga. Hawaii received United States statehood in 1959. Economically, the change has been less dramatic. Since World War II the colonial powers have devoted much more money and attention to agricultural development. New crops, such as coffee and cocoa, have become important. Tourism has expanded. But copra in most islands and sugar in Fiji have remained the basis of agricultural production, even when their value has been overlaid by mineral wealth, such as copper in Bougainville. There still remains the problem of diversifying the economies of Oceania. There remains, too, the problem of education. In 1965 a university was established in Papua and New Guinea. Two years later the University of the South Pacific was set up in Fiji. Yet New Guinea still has a major problem of literacy and of elementary and secondary education; and, although a high level of literacy exists in Samoa and the French territories, education is largely still at the basic primary level. There is a shortage throughout Oceania of skilled artisans and an even greater shortage of professionally trained people.

Political development has been faster partly for external reasons. International pressure in the shape of a United Nations Visiting Mission, such as that of Sir Hugh Foot to New Guinea in 1962, forced the pace, although he was pressing an Australian government already moving in that direction. A growth of regional feeling, fostered by the international South Pacific Commission set up in 1947, increased political awareness. The main dynamic of the advance has been the governments' perception of circumstances in the islands themselves. Oceania has produced none of the mass national movements of Africa and Asia. The reaction to European rule has usually taken the form of nativistic movements, or cargo cults, in which, by ritual, Oceanians await the coming of cargo diverted from them by Europeans. There have been attempts to remedy the frustrations felt at European material superiority. But they have not been political movements, except in so far as they now underlie the electoral processes of Oceania. The Mau, or opinion, movement that declined cooperation with the government in Western Samoa in the 1920s and 1930s was more overtly political, but it was not in any usual sense a nationalist movement. The nationalism that has existed has been the work of political parties formed as a result of the creation of electoral processes and assemblies. In French Polynesia, the Rassemblement Démocratique des Populations Taïtiennes began as a political opposition but split when it became a government. In Samoa, politics have followed kinship groupings rather than nationalist principles. In Fiji, politics have been racial rather than nationalist. In Papua and New Guinea, political parties have been slow to form. The reasons for this absence of nationalism are only partly that the colonial governments have on the whole maintained native interests in the face of European pressures and thus have prevented landless, detribalized people from emerging. The reasons lie much more in the local nature of Oceanic societies, which still look on politics as a matter of kinship or conformity to correct behaviour as determined by their own standards. Locally there are exceptions, but in general Oceania has been shielded from the impact that might have destroyed its societies. All of them have changed by their involvement in the European world, but the changes have been adapted so that the new is still interpreted in terms of the old.

Nativistic movement

BIBLIOGRAPHY

Prehistory: The picture is changing so rapidly with every new site that there is no comprehensive account. The best work is J. GOLSON, *Prehistoric Research in Melanesia* (1967). A traditional view is presented by P.H. BUCK, in *Vikings of the Sunrise* (1938), now very doubtful.

Exploration: The standard work is J.C. BEAGLEHOLE, *The Exploration of the Pacific*, 3rd ed. (1966). A. SHARP, *The Discovery of the Pacific Islands* (1960), lists the first European sightings of every island. An older popular account is J.A. WILLIAMSON, *Cook and the Opening of the Pacific* (1946).

Missionaries: The only general book is A.A. KOSKINEN, *Missionary Influence As a Political Factor in the Pacific Islands* (1953); but W.P. MORRELL, *Britain in the Pacific Islands* (1960), includes accounts of missionary activity.

Traders: The trading history of Oceania has yet to be written, but studies of particular aspects of trade are well described in H.E. MAUDE, *Of Islands and Men* (1968). D.S. SHINEBERG, *They Came for Sandalwood* (1967), deals with one crucial trade.

Colonialism: Colonial rule in the Pacific has received much attention. W.P. MORRELL, *Britain in the Pacific Islands* (1960), is the standard account. F.J. WEST, *Political Advancement in the South Pacific* (1961), deals with the post-World War II political development in several territories. Individual islands so far studied are: J.D. LEGGE, *Britain in Fiji, 1858–1880* (1958); and R.P. GILSON, *Samoa 1830–1900: The Politics of a Multi Racial Community* (1970). J.W. DAVIDSON, *Samoa Mo Samoa: The Emergence of the Independent State of Western Samoa* (1967), is the view of a participant in the independence movement. R.G. CROCOMBE, *Land Tenure in the Cook Islands* (1964); and N. MELLER, *The Congress of Micronesia* (1969), discuss aspects of islands so far little studied. F.J. WEST, *Hubert Murray: The Australian Pro-Consul* (1968), gives an account of Papua under Australian rule; and S.W. REED, *The Making of Modern New Guinea* (1943), describes the mandated territory.

(F.J.W.)

Oceanian Peoples, Arts of

Art is an imprecise concept, especially if the economic connotations that accompany it in a given culture are not regarded. In the West, indeed, anything sold as art is considered such, especially whatever sells best. Well-known "aestheticians" have gone so far as to treat the bark paintings of Arnhem Land (Figure 1) as mere commercial objects and yet to regard as "art" certain pieces that, after a lengthy period in European hands, had lost their colour and acquired an "admirable" patina after being zealously waxed!

Some art historians have tried to understand Oceanic art —like other supposedly "primitive" arts that now look so modern—as rising from the same kind of inspiration that gave birth to Western and other cultures. They have spoken of magical art, then of religious art, using concepts and connotations—positive or negative, black or white —that properly belong to the religious and theological universe of the Western world. It is better to discard such hand-me-down ideas and to look at the way each piece of Oceanic art—including those objects that the West regards as beautiful and, also, those that the Oceanic peoples themselves appreciate as such—finds its place within the specific society and culture. Within the concept of Oceanic art is included not only the plastic works—sculpture, painted plaques, monumental structures—but also any object in daily use that has received a finished or intentionally fashioned form (Figure 2). It likewise in-

cludes the vocal arts, with all the forms and uses of a construed sentence, music, and many of the elaborate formal gestures of the dance.

This article is divided into the following sections:

I. General considerations

THE SOCIAL MILIEU

The influence of the social milieu on art is a constantly recurring problem in aesthetic studies, perhaps because the meaning of the term social milieu is so elusive. It is, indeed, better to question how aesthetic expression is introduced into and operates within the social group, from which it can only be detached for the sake of convenience and in formal terms; for there can be no aesthetic manifestation unless there is at least one person to serve as the public. There is, however, a permanent and balanced relationship between art and society in the sense that, while aesthetic expression profits from an organization that ensures its reproduction or transmission, the artist derives advantages that do not belong solely to the level of art: prestige or gain, if not both of them together.

Commercial values. A discovery of recent decades is that among the Oceanian peoples art has its commercial value and is bought and sold. The Melanesians, at any rate, have a system that, for want of a better term, might be called "copyright." This system is as complicated as that in the West, with both negative and positive aspects. It is forbidden under pain of supernatural punishment to take that which does not belong to you, and this extends to the arts: all stylistic elements, whether plastic forms, painted decoration, musical themes, poems, or sequences of dance steps are copyright. The usual owner is the local group or clan attached to a given area, and the group is represented by some chosen dignitary. Ownership may be by inheritance, the result of an expensive purchase, or both. Literature sells least well. This is because sung verse usually embodies the myth of a particular clan, and the recitation of a myth to which one has no right is considered theft—it is seen as an attempt to deprive the owners of their land. One does, however, come across song texts with the original music, either in the original language or in that of the borrowers, sometimes paired and sometimes not. (Chants can also be transmitted from one place to another without an accompanying translation and without the new singers necessarily knowing the meaning of the words.) Dance steps can also be borrowed and used with a different chant. The important thing is to pay for what is taken and (so that there can be no misunderstandings about this) to pay in public—ritually and with goods whose value is undisputed, such as pigs or traditional treasures.

The Oceanic conception of ownership

A work of plastic art may also be sold, but the sale need not be of the piece itself. It may be one or other of its elements that provides the attraction—an outstanding portion, an astute technique, an aspect of decoration, or a combination of all. Each legitimately bought element can be combined with others obtained elsewhere to produce a new artifact that will appeal to the customs and fashions of society.

Art and social institutions. Aesthetic expression is thus influenced by society in the wide sense and makes use of its institutions and forms of control, but there is hardly any evidence to suggest that a mode of government can have been the originating force of any particular aesthetic tradition. The combined nuances within an organized society can exert such a force, however, and a good example of how they do so is found at the level of the organization of trade. Craftsmen specialize by village in the Admiralty Islands, making a specific category of goods, and this ensures a stability of form, with variations occurring only in the fine detail. Individual gifts and abilities are not especially important. On the other hand, societies that offer a premium to initiative or to individual inspiration (as in Middle Sepik, a region of central New Guinea)

By courtesy of the Museum of Primitive Art, New York; photograph, Lisa Little

Figure 1: X-ray-style bark painting of kangaroos, earth pigments on eucalyptus bark. From Arnhem Land, Australia. In the Museum of Primitive Art, New York City. Height 1.03 m.

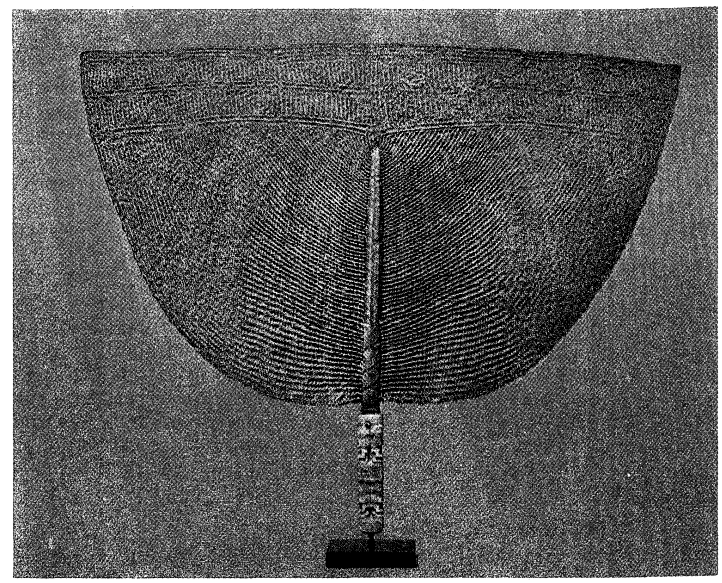


Figure 2: Fan, pandanus and whale ivory. From the Marquesas Islands, Polynesia. In the collection of Mr. and Mrs. R. Wielgus, Tucson, Arizona. Height 46.4 cm.
By courtesy of the Art Institute of Chicago

foster the means to maximum creative power, and the importance of individual talent in this situation is paramount.

But this factor alone does not seem to account for the comparative artistic richness of certain areas. Wide circulation of information and the buying or selling of artistic ideas and elements are equally important factors and favour those places where ideas and themes are most readily interchanged. Another essential factor in the situation is competition for prestige among individuals (Middle Sepik) or among groups (New Hebrides, New Ireland), which, it is generally agreed, results in some of the finest art forms of all. Where there is a frozen hierarchy of power and prestige, such as there is in Tonga, a sophisticated theological structure is evolved, but a diversity of art forms is not (though some does exist). The complex theogonies of the Maoris, for example, were not the inspiration behind sculptures and bas-reliefs of New Zealand (Figure 3); the fact that rival schools of sculpture flourished there offers a better explanation for their exuberant form.

By courtesy of the Hamburgisches Museum fur Volkerkunde, Hamburg; photograph, Ralph Kleinhempel, Hamburg

Figure 3: Maori (Polynesian) wall panel from a meeting house, wood and reeds. From New Zealand. In the Museum für Völkerkunde, Hamburg. Height 21.6 cm.

Sexual attitudes seem to have had little connection with artistic productiveness: the New Caledonians, who are sexually reserved, have created many interesting works of sculpture; the inhabitants of the Loyalty Islands, who are permissive, have a very developed art of chanting and dancing (but know almost nothing of sculpture), whereas the Trobriand Islanders, also permissive, have not developed these arts.

DEMOGRAPHIC INFLUENCES

Coastal populations seem to be more productive than those of the interior. This is especially true of New Guinea, where distance from the coast can mean virtual isolation. Being by necessity involved in coastal trade and therefore well placed for cultural interchanges, the seafarers all knew some kind of sculpture, though of widely differing forms. But this is not a hard and fast rule. The people living inland in New Guinea, known for convenience under the designation of Middle Sepik, achieved the highest artistic levels, for instance, and the inhabitants of the highlands often produced forms, motifs, and colours of great interest. Their work, however, was either deliberately destroyed after ritual use or else decayed very quickly, for they used mainly fragile vegetable materials (Figure 4).

Works that were made to be destroyed

Works gathered by European collectors were usually taken from the coastal regions, and even these were those pieces that were most easily transported. Very little is known, therefore, about the interior from an aesthetic standpoint, and problems of transport still present a formidable obstacle in the way of learning very much more.

Isolation. The state of isolation, when it truly exists, inevitably works against inventiveness and encourages stereotyped repetition. Actually, only cases of relative isolation are known: even the people of Tikopia, for instance, travelled far in their boats, or "pirogues" as their crafts are called, along routes established by tradition, and they reached some destination. The extent to which harsh climatic conditions inhibit artistic expression is not known, if indeed they do at all—the abstract art engraved on stone plaques by aborigines of the inhospitable central Australian desert is, for example, especially fine. Nor is there much evidence that the rough or elaborate aspect of wood sculpture (Figure 5) has any relation to technological backwardness or otherwise: the basic technique of cutting wood (which is prepared by slow, controlled firing) with a stone hatchet is everywhere the same, and only the finishing tools—splinters of jade or obsidian, shells, dog incisors, and the like—vary.

By courtesy of the Museum of Primitive Art, New York; photograph, Lisa Little

Figure 4: Vegetable panel, sago spathe, bamboo, earth pigments. From the Lower Sepik area, New Guinea, Melanesia. In the Museum of Primitive Art, New York. Height 1.02 m.

Intercultural borrowings. It is the element of complexity and number of intergroup relations that seems to provide conditions most clearly favourable to aesthetic production in Oceania. Two thousand years of Western history teaches the same thing. Objects are "borrowed" by one culture from another, but the object borrowed does not always retain its original significance. The society borrowing, stealing, or plundering the object attributes to it the connotations that suit it, those that are available at the moment. A mask, for example, acquires quite a different significance from one place to another (Figure 6): it may be an object used in theatrical games, with the laughing participation of women and children; another group may use it as a symbol for the initiation of young men, with the appropriate respect (even if the oldest women know which man is hidden under the dis-

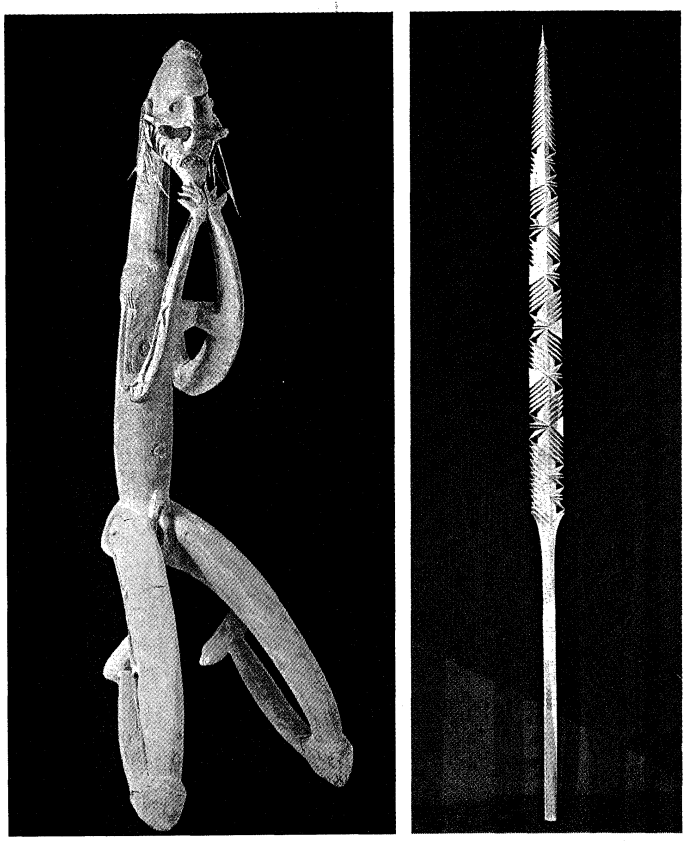


Figure 5: *Rough and elaborate styles of woodcarving.* (Left) Canoe prow from the Asmat area, New Guinea, Melanesia. In the Museum für Völkerkunde und Schweizerisches Museum für Volkskunde Basel, Switzerland. (Right) Spear from the Samoa Islands, Polynesia. In the Peabody Museum, Harvard University. Length 1.88 m.

By courtesy of (left) the Museum fur Volkerkunde und Schweizerisches Museum fur Volkskunde Basel, Switzerland, (right) the Peabody Museum, Harvard University, Cambridge, Massachusetts; photograph, (left) Hans Steiner

guise); somewhere else it represents a returning ancestor and is invested with emotion accordingly; and yet elsewhere it serves as the public symbol of an institution, guaranteeing a form of social control and serving to support the proclamation of decrees concerning public order. It never has the same connotations in one place as in another, and generalizing discourses on the mask, as a subject, have little basis in fact.

Western influence. The effect of the West on Oceanic culture after two centuries has been both overestimated and underestimated. During the 18th century, through the intelligence or determination of the parties concerned, the relationship between the two was much more equal than is generally believed, though at the cost of many human lives. Their artistic traditions were not stifled; indeed, when the European policy of conquest was abandoned, largely because of the territories' lack of useful natural resources, many societies underwent an aesthetic flowering (particularly in the visual arts) that has not always been appreciated. This took place because metal tools had been introduced, bringing about more efficient techniques in carving wood, and because new leisure time had been created, which had to be filled by a kind of ritual explosion and in which invention was given free rein (though the themes were, nevertheless, drawn from tradition).

Effects of Christianity on Oceanic art

In the 19th century, however, conversion to Christianity and forced colonialization brought about the progressive disappearance of all indigenous visual arts. The musical arts and the dance resisted better: despite the quickly adopted Western techniques, they displayed much originality in their reinterpretation. (Some of the most modern experiments in rhythm were foreshadowed in the orchestras of unknown villages in the mid-Pacific, where the instrument was no more than a string stretched across an empty gasoline can.)

HISTORICAL PERSPECTIVES

Oceania does not lack historical depth, though knowledge of it is not consistent, even among the people of the region: the great Tonga lineages stretch back for 40 generations, whereas some groups have difficulty in tracing themselves back more than two. Tonga is also exceptional in that its dynamism was turned outward, for it built up a maritime empire. But usually, within an insular society's stable population and fixed habitat, the complexities of family relationship were of major importance—particularly when determining an inheritance. Often the relationships worked out were not accurate, but this need not, in fact, have hindered the actual working of a system based on genealogical reference. In New Guinea, on the other hand, where a 1,000-year-old population movement had been in existence (partly from west to east, partly from the high, overpopulated mountain valleys to the less healthy coastal areas), the permanence of land tenure was never guaranteed, and the peoples there did not favour the father-to-son pattern of inheritance. This means that notions of patrimony and inheritance pertaining to other cultures must be very carefully reviewed and corrected before being attributed to Oceanic societies. Even the most specialized Western observers, in studying this phenomenon, have too often been coloured by value systems belonging to their own society.

Continuity of tradition. There has been a tendency to treat the Oceanian peoples a little as though they belong to fairy tales. The societies they constructed and made function were considered by ethnologists to be as fragile as they were precious. They were supposed to have operated rigorously and unchangingly over thousands of years until the days of European colonialism, which the specialist considered a noxious influence with an impact that inevitably destroyed the existing society. It is true that the West and Christianity made such a crushing advance that Oceanic societies appeared to be delicate structures, as easily broken as porcelain. No one imagined that they had adopted tactical, collective decisions in the face of the invader. These tactics were aimed at saving the essential structure, while letting go of the superfluous—the visual appearances that the observer so often regarded as basic. This fact helps to explain how, at different periods, mass conversions to Christianity (the missionaries' efforts being supported from afar and close at hand by irresistible naval forces) were followed after 50 years or so by an easy return to paganism (except where, disconcertingly to Western observers, the natives hurled themselves headlong into some kind of messianic movement). That societies, weak in the sense of lacking armed strength, might adopt tactical compromises when faced by a superior force ought not to be surprising, for it is a recurring feature of behaviour throughout the history of the world; Oceanian peoples, like all others, are part of the evolutionary pattern. Delicate and supple social structures can be destroyed, however, and it is little short of miraculous that Oceanic societies should still be alive and thriving, capable of independent life while preserving a cultural heritage sufficient to maintain their cohesion.

The Oceanic world view

Such a capacity for life implies that, before the arrival of the white man, Oceanic societies had a broad vision of the world that could accommodate any new situation they faced. It is not always easy to rediscover this vision through the distorting prism that now exists because of the mistaken conceptions of the first European observers and the considerable caution of their Oceanic informants. They were conscious that if they were to protect themselves, it was necessary for them to comprehend and assimilate the white man's concepts and way of thinking. The native's great gift for psychological penetration and his capacity for adaptation are thus largely responsible for the inadequate and untrustworthy evidence that stems from that time. This is especially true of the Polynesians, who quickly learned how to present themselves to romantically minded navigators as representatives of a classical culture like that of the Celts, a society with bards and hierarchies of kings and priests. As a result, they were often able—with the exception of Hawaii—to retain more of their land and to avoid collective conscription

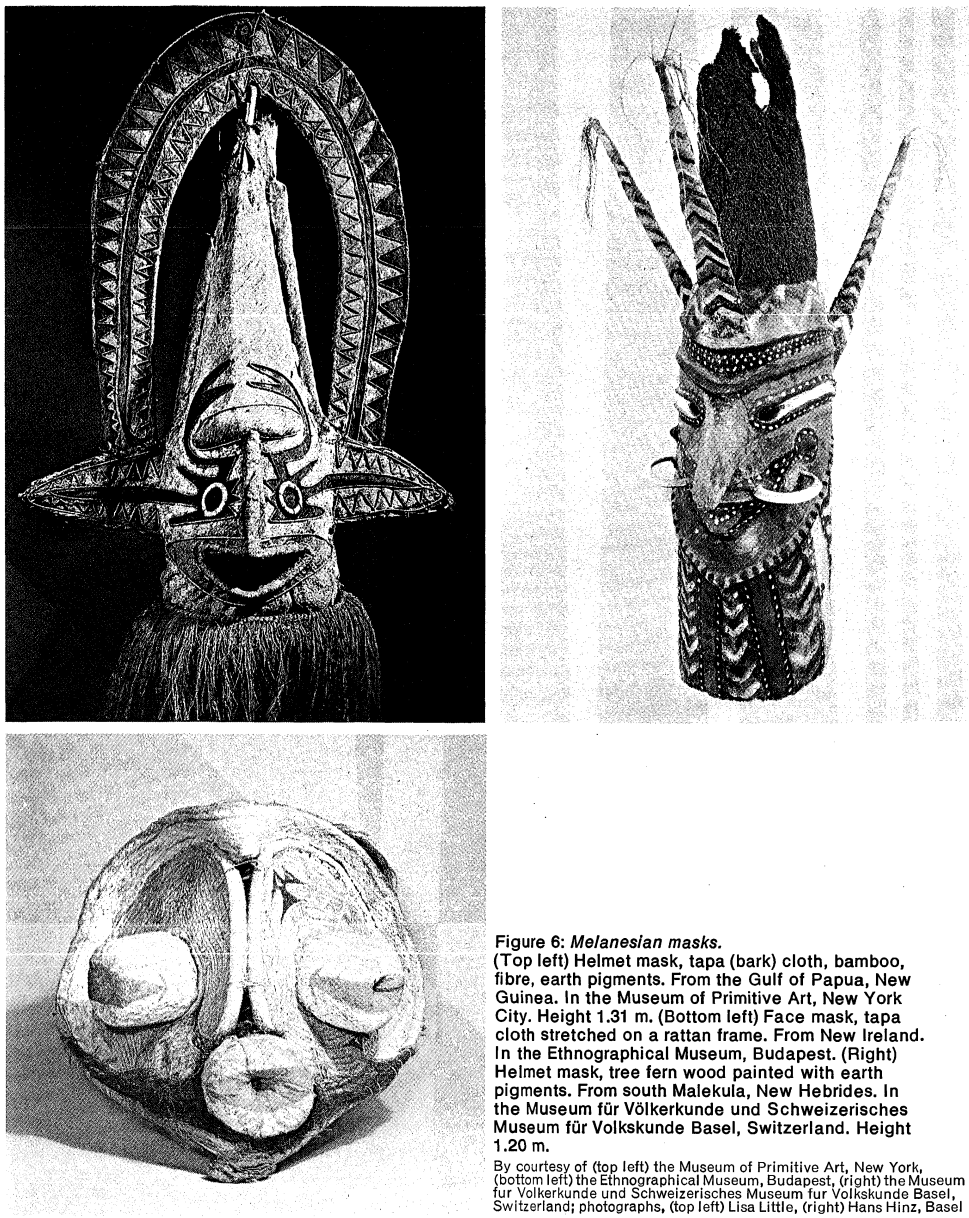


Figure 6: *Melanesian masks.*
(Top left) Helmet mask, tapa (bark) cloth, bamboo, fibre, earth pigments. From the Gulf of Papua, New Guinea. In the Museum of Primitive Art, New York City. Height 1.31 m. (Bottom left) Face mask, tapa cloth stretched on a rattan frame. From New Ireland. In the Ethnographical Museum, Budapest. (Right) Helmet mask, tree fern wood painted with earth pigments. From south Malekula, New Hebrides. In the Museum für Völkerkunde und Schweizerisches Museum für Volkskunde Basel, Switzerland. Height 1.20 m.

By courtesy of (top left) the Museum of Primitive Art, New York, (bottom left) the Ethnographical Museum, Budapest, (right) the Museum fur Volkerkunde und Schweizerisches Museum fur Volkskunde Basel, Switzerland; photographs, (top left) Lisa Little, (right) Hans Hinz, Basel

for forced labour. The explanation of this capacity for adaptation should be sought in the very heart of traditional structures, in their permanent receptivity to any external contribution, and in their tradition of welcoming the newcomer and providing him, in the form of land, spouse, and social status, with the means of becoming a part of society. The entry of European elements must also have been facilitated by a long-standing pre-eminence given to light skins—girls of high station were kept in the shade, for example. The Europeans were therefore looked on favourably as possible fathers. The symbolic contrast between black (accentuated in time of war by means of various vegetable pigments) and white (the colour of mourning, obtained by coating oneself with ashes or lime on the hair and upper part of the body) has played a constant role in the relations between Europeans and Melanesians. Esteemed like the dead, that is to say like gods, because of the colour of their skins, the whites benefitted from the prestige that this gave them until events proved that they were mortal and fallible, so that they came to be dreaded instead for their intentions and material strength. There was a later attempt to place the phenomenon of the white man's arrival within the context of destiny. Various explanations emerged, turning around a confusion of such ancient themes and biblical outlines as the departure of a Melanesian or Papuan Noah in a pirogue, or canoe, carrying with him the elements of power that would one day be restored to his people or the idea of the "cargo," material wealth that had been sent by the dead to their descendants, in chests whose labels were changed by the whites so that they could usurp the contents.

The Old World vision first gave way before the arrival and entry of outside elements, both from the sea in pirogues and from inland by migrations that drove the people of New Guinea from west to east and from the high, overpopulated mountain valleys toward the coast. Groups were now defined socially in terms of their physical—that is, topographical—position within the general system. The universe, broken down and divided into elements that could be listed, was apportioned among the various groups. Each group had symbolic control of its part of the known world—its animals, vegetable life, atmospheric elements, mythological figures—and located on its own territory fixed spots from which this control was asserted. Thus, the soil itself was of prime importance, and it is very important to understand that in societies without writing this should be so. For the soil offers subsistence and is literally the basis of social and religious rites—it supports communication routes and provides the sacred spots (ritual or mythological or both) among the areas assigned to lie fallow or be cultivated, to

be given over to hunting or the gathering of fruit and vegetables. This fact of the soil's importance cannot be neglected in favour of an arrangement of concepts more in line with the intellectual habits of other cultures. In order to begin understanding Oceanic culture, it is essential to realize that its peoples symbolically express their very society in what may seem to others no more than a dry list of topographical detail. The "list" is an embodiment of their basic myth, and the native inherits, as it were, a "key" to its interpretation.

Culture and tradition. There is an often bewildering multiplication of languages spoken in Oceania, with several hundred separate tongues for New Guinea alone and scores for each of the Melanesian archipelagoes. In contrast, the Polynesian languages are more obviously similar, a fact that may be connected with man's comparatively late arrival there. The languages of Oceania are divided into two large groupings. One is called, at present, Austronesian, which is considered to be relatively recent, and which includes Indonesian, certain Madagascan, Melanesian, Polynesian, and Micronesian languages. Formerly this grouping was termed Malayo-Polynesian. The other group, called Non-Austronesian, is considered to be older than Austronesian, and includes many of the languages of New Guinea, of New Britain, and a few isolated languages in New Ireland and the Solomon Islands. In older terminology these Non-Austronesian languages were subsumed under the term Papuan. While the Austronesian languages seem to be historically related, the Non-Austronesian group serves as a convenient catch-all grouping, the constituent languages of which are not necessarily related to each other. The presence of so many of the older Non-Austronesian languages in Melanesia makes this region an important one for research into historical linguistic relationships between older and newer language manifestations prevalent in Oceania.

Language groupings

Symbolism, imagery, and the aesthetic tradition. It is difficult to pinpoint any common general elements within the great variety of symbolic systems that the Oceanian peoples have built up. The bird symbol appears to be the one most frequently used, as much in speech, myth, dance, and song as in visual art. It seems to represent the political group, power, high social status, and also virility. Often it is a small sea eagle, common in coastal regions, or some other sea bird, such as the frigate bird. But the symbolic meaning is attached to the species and not the genus. A predatory bird of the fields, the sultana hen supplies a female connotation—it is the symbol of illicit love. Therefore one should not invoke the whole order of flying creatures to explain the above series of symbols, inasmuch as the large fruit-eating bat that feeds on the same fruits as the sultana hen is both a male symbol and a representation of life; the lizard, the water snake, and the shark (Figure 7) have the same connotations, and those of the shark can vary to the point of overlapping those of the sea eagle. The explanation of this apparent discrepancy lies in the division between, and the appropriation by, the different groups constituting a society of the whole of the known elements of the universe. The specific significance of the recourse to such a symbol will only be understood in terms of the whole symbolic lexicon and, simultaneously, of the overall aspects and details of its adaptation.

It is, likewise, difficult to extract from the overall culture anything that might be called a common aesthetic tradition. Landscape architecture might constitute one, for it is everywhere methodically planned and executed, though with quite different aims from one archipelago to another. Here, one will find a compact, close-cropped lawn around the houses; there, everything that encroaches on the black, lustreless, volcanic soil has been carefully uprooted. Here, the hierarchy of constructions will be organized vertically as a function of their attributes; there, grassy open spaces will be used to distinguish large categories of huts. The most elaborately worked of these glades, in terms of the techniques of arboriculture used to provide shade, will be the one that contains no huts, being used as a common dancing ground by several groups.

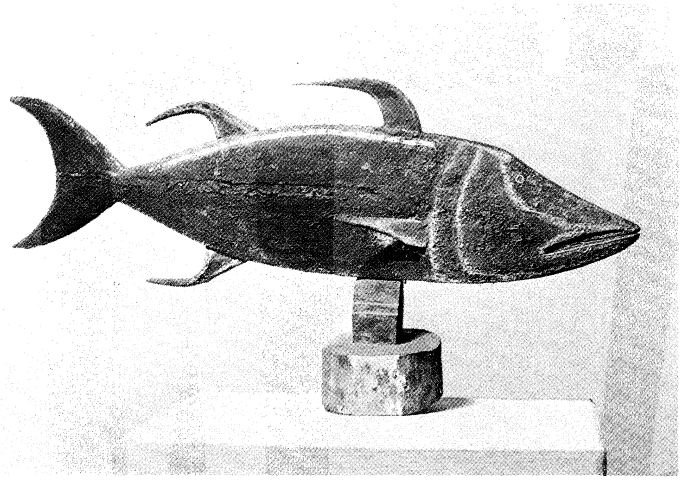


Figure 7: Carving of a shark, blackened wood with inlaid decoration of mother-of-pearl. From San Cristobal Island, Solomon Islands, Melanesia. In the Museum für Völkerkunde und Schweizerisches Museum für Volkskunde Basel, Switzerland. Length 57.8 cm.
Holle Bildarchiv, Baden-Baden, West Germany

CONTEMPORARY FEATURES

It has been suggested above that most Oceanic cultures, despite their confrontation by Western colonialists and missionaries, retained much more life and cohesion than observers generally supposed. It is, however, a fact that the expression of their culture was seen as a manifestation of godlessness by the white overlord and could not be maintained. Much opprobrium has since fallen on those missionaries who compelled the suppression of chants and dances and still more on those who burned the idols (Figure 8). But a great many Polynesian sculptures were

Figure 8: Wooden image of the war god Kukailimoku, one of the Polynesian works of art that escaped destruction by missionaries. From the Hawaiian Islands. In the British Museum. Height 76.8 cm.
By courtesy of the trustees of the British Museum

also destroyed by isolated Polynesian evangelists and their converts, who were eager to demonstrate their changed lives, and often were motivated further by the hope of attaining the evident wealth and power of the white man. Traditional dances and chants were generally condemned by Catholic and Protestant missionaries alike, but the extent to which they survived depended upon the

proximity of the missionary station and the zeal of its local representative. Many songs and dances were, in fact, preserved, often because European dignitaries expressed an interest: official ceremonies that they attended would incorporate music and dance. Tourists were also interested in folklore, and their patronage helped to keep alive some parts of it as well as to encourage pale copies and complete invention. But a tradition can only be maintained if the principals involved really desire it. Once the ritual itself has been abandoned, everything related to it has the greatest difficulty in retaining anything but a residual or revised function.

It is, however, astonishing to discover how much was not destroyed but lay dormant, eventually to emerge from an oblivion that might have been thought final as Christianity began to suffer a relative loss of prestige. Complex texts and ceremonies have been built up in a way that reveals how retentive is the collective memory.

Commercialization

The plastic arts, on the other hand, have largely succumbed to a process of commercialization whereby an ever greater number of objects—mostly of poor quality—are needed to supply the needs of a greatly expanded market (much of it touristic) that is controlled by a class of, usually, European middlemen of extraordinary rapacity.

II. Literature

GENERAL CHARACTERISTICS

Oceanian literature is oral. Its existence was discovered about two centuries ago in Polynesia, when European observers realized that complex cosmogonies and theological formulations, embodied in astonishingly long and complicated texts, were committed to memory by what they thought of as bards and priests. For a long time, however, no one felt any call to take this discovery further. The prevailing strain of European pre-Romanticism—and, later, Romanticism itself—which then tinged the outlook of most observers, unfortunately resulted in intellectual scruples being abandoned when they came face to face with attractively exotic Oceanian situations; they produced a flood of literary reconstruction, giving current European ideas free rein, so that the genuine Polynesian literary tradition was quickly swamped. To construct a "Polynesian" text it was enough to sprinkle it with proper names and a few other words or phrases borrowed from the language. Those who did not invent did translations, and the translated texts were even more dangerous because of innumerable misinterpretations and small mistakes. Few authors supplied the Polynesian text alongside their translation, a course that would be expected if the work were one of the genuine scholarship (although an early effort by serious scholars did succeed in organizing the body of New Zealand's vernacular, unexpurgated texts, presenting them in a line-by-line translation, with an arsenal of notes and commentaries, in such a way that the text could be both understood and critically appreciated). A properly serious and scientific approach to textual criticism was, on the whole, reserved to the classical and sacred texts of other cultures.

The rise of anthropological studies toward the end of the 19th century ought to have brought with it a new interest in Oceania's oral literature. But however much conditions varied from country to country, anthropology was never practically influenced by the methods of classical philology, the rigorous discipline of which it admits with the greatest reluctance even today. Vernacular vocabularies did become an object of sustained attention for certain scholars, but they were basically seeking to describe normative and conceptual systems; as a result, they preserved very little apart from those phrases that aptly served to demonstrate the truth of their own arguments and theories. The methodical collecting of documents was scarcely begun, and the task has now become one of extreme urgency before texts are completely forgotten. It is, moreover, the sort of work in which linguists and anthropologists ought to collaborate.

It is not easy for scholars belonging to other cultures to describe the content of Oceanic literature because their understanding of it is so limited. There are, however, one or two pointers. First of all, it has to be said that the purpose of literature is to communicate: it demands an audience. In the case of an oral literature, communication depends first on memory, and this usually means that such memory aids as rhythm and stock formulas and phrases are an important element of all texts. The majority of Oceanic texts keep closely to traditional forms and appear to be committed to memory and then communicated in a strictly unvarying manner. This is, however, only approximately the case because the various techniques of formalization can allow a quite fluid text. The tradition can be made evident at the lexical level, with the possibility of a great freedom of syntax. It only prevents prosodic elements from taking on primary importance. But it is these that have been studied last by modern linguistics.

Traditional forms of oral literature

Types. The literary occasions of the Oceanian peoples are, as in other cultures, reflected in sacred literature, political literature, and frivolous—even erotic—texts. This division, however, should not be taken to represent an attempt at classification; any such pigeonholing would be inconvenient, indeed, because too many texts defy neatness and straddle two categories. But there are certainly two poles between which the various forms of literary expression can be placed. On the one hand, there is a body of works that appeals to Western readers and is made accessible to them by its use of the poetic image. On the other hand, there are many texts, often brief, in which each word is frequently a complete image. This kind of text is part and parcel of the culture that has produced it and requires a veritable arsenal of commentaries of others to interpret the key words and unravel its significance. Texts are, basically, of two kinds: (1) recitatives, whose form is rigid; they can be expanded but not transformed; in this category belong all the songs or chants that accompany dances—whether the performers be standing or seated—funeral chants, songs that accompany children's games, and those with an erotic significance; and (2) public orations, in which the elements are formally but roughly organized, giving the speaker the right to vary the presentation within certain limits established for this literary genre. Such discourses, which can aptly be delivered as a high-level political oration and as a funeral eulogy or remembrance speech, can also, in a more simple form, commemorate such events as a birth or a marriage.

Themes. The themes are those that appear in other literatures of the world: love and death, defiance and hatred, nostalgia for the past, and the pleasure of the moment. Nature provides the necessary images. There is a barrier between Oceanic literature and that of other cultures because, although it presents a familiar mental universe, it does so in an often allusive manner that demands a knowledge of local place-names, local political geography, and land division before its meaning can be understood: the owl, for example, is symbolic of a given place, the lizard of some other, and the sea eagle associated with a third.

Although there is no specific body of Oceanic literature, there are fragments of different kinds that, once gathered, suggest a cultural coherence: definite conclusions cannot be drawn, but lines for further research are indicated. Even small texts, which have often been overlooked because they did not seem sufficiently elevated—such as those that were accompanied by a flute or a stringed instrument as part of individual rites for the purpose of securing a woman's favours, contribute to the total picture.

Importance of small texts

MELANESIAN LITERATURES

Melanesian literatures are better known today than Polynesian because they remained untampered with over a longer period by Western anthropologists and scholars and were, therefore, not distorted by them. Thus, they are of great interest, although they are only just being recovered. Great pioneering work was done by Maurice Leenhardt into the literature of New Caledonia, but it has scarcely been followed up. A prime difficulty is that of method. Leenhardt trained a few Melanesians, teaching

them to read and write their own language after inventing for them a system of transcribing their verbal utterances. Thus, he provided for himself scribes—and even authors—in the local language that he had chosen for study. The body of literature he assembled in this way, since published only in part, is the most complete available to date. Leenhardt's method (which had in fact been previously employed half a century earlier by British clergyman Robert Henry Codrington [1830–1922], who did not, however, leave to posterity the literature of the Mota language that had been given him by his pupils) makes it possible to obtain texts of very high quality over a long period. Written out by the trained student himself, who has had the opportunity to re-read and check for accuracy in the quiet of his hut, the method avoids the psychological tension inevitable in a dictation session. The results make for more confident study afterward, and the method is still the only sound one. Texts taken down directly on a tape recorder can too easily be garbled or abridged unless the informants are especially trained to use the machine by a linguist, so as to avoid a psychological block when faced by it.

Two generations of researchers have succeeded Maurice Leenhardt, working in the same region as he did. They have amassed a complementary body of literature astonishingly rich and have been able to establish just what the function of each text was in the life of the society they studied. The results of their work have obliged scholars to re-examine their conceptions of the oral literary tradition. It now emerges that not even words, symbols, or places have any fixed significance: the vocabulary is at once coherent and diversified, for it is used in a way that takes into account the momentary interest of the parties present.

The traditional speech. A traditional speech may be "hurled" at the crowd by a speaker perched on a flimsy platform; the crowd responds with a muffled cry at the end of each "sentence"; it is the quality of accord reached between the crowd and the speaker that comprises the "new content" of the familiar, traditional speech (henceforth becoming part of its tradition), while the accord also conveys what other cultures would call the "message"—the new stage in a rising political career, perhaps, or a declaration denoting peace between combatants or the beginning of a war. Anyone analyzing the speech itself has to search for and consider carefully all the possible interpretations and temporal conditions it might have. The traditional form of the speech is respected all the more because it allows this variation of content—within acceptable limits—to be conveyed by nuances that may easily escape observers from other cultures.

The form of the speech also presents a somewhat thorny problem. On the surface, it is a simple enumeration of the local groups and their symbols (that is, the portion of the land that is theirs, together with its animals, plants, weather conditions, and so on). The recitation of these physical realities is an affirmation of their very society, and they are stated one after the other, linked by some stock connecting phrase. Each listed item is given its precise geographical location; of special importance are the names of special places where authority is exercised or where rites are practiced. The whole constitutes a world vision, or system, in which the individual society and its members have their place. The native audience is perfectly familiar with the spatial affirmation of their society contained in the speech, and from the dry enumeration of its components, they are able to supply for themselves a history of alliances and wars and to remodel the traditional text until it fits the conditions of the present. The orator's delivery, the nuances he maintains or introduces into the speech, are a sign of his success.

The myth. Another kind of literature, less easily defined, is the one commonly understood as myth (a term that smacks of Western classical culture and thus carries overtones that are irksome from the scientific point of view). Involved here are all formalized recitatives, straightforwardly delivered, made up of narrative with pleasing scenes interpolated, whose apparent aim is to give the account of a rite, for example, or to describe the situation of some group of family lineage (perhaps in connection with land ownership or political status) or to tell of the origin of man and his culture. The recitative may involve as little as ten lines, or it may take a man three hours to dictate. In any given local culture, the rules that govern the way in which the texts' content is formalized and those that govern the way in which they are recited are consistent. The symbolic vocabulary, formally identical with that used in public speeches, carries elaborate but acknowledged references. A text may be established on the basis of a single symbol, but in general the symbolic pattern is so complex that other cultures have great difficulty in understanding it. Indeed, it is only possible to decipher its meaning if the cultural significance of every place mentioned in the text is understood, if it is known what creature or being is worshipped where and to the benefit of which group, and often only if the intineraries that are the subject of a majority of the myths are entirely familiar to the interpreter, so that old placenames—and thus their meaning—that are referred to in the text can be identified.

POLYNESIAN AND MICRONESIAN LITERATURES

Similarity of Polynesian and Micronesian literatures

Problems of interpretation. Polynesian and Micronesian literatures are similar in structure, but they vary in detail from island to island. Little is known about the Micronesian texts, however, and work on the interpretation of Polynesian literature has been hindered—first by the excesses of European Romanticism and its flights of fancy, then for nearly two centuries by well-intentioned amateurism. Authentic texts were not collected when the opportunity was at hand, and it is now often too late to do so because the traditions—especially of religion—that maintained them have not survived. A great body of mythological material was gathered in the 19th century, but it has still not been critically edited, and it cannot be overemphasized that nothing like a properly established record of their myths has yet been produced. Polynesian "priests" have been likened to those of ancient Egypt, their recitatives treated as theological monuments—interpretations that distort the situation entirely. Such textual interpretation must, in any case, be demonstrated scientifically; since the commentaries published to date do not attempt to do this, the serious investigator is suspicious also of the soundness of their attendant translations. These translations, on the slightest pretext, indulge in lyrical flights of fancy that seem to have been inspired by old memories of Sir Walter Scott or the poets of the French Parnassian school. Editions of Polynesian texts are published by the Bernice Pauahi Bishop Museum in Honolulu, and, though interesting in the absence of any other methodical work, these, too, are open to criticism. Their translations are never literal but attempt some kind of free translation, as though dating from a period when translations aspired primarily to literary beauty.

As a result, the theogonies that have been published (supplied with diagrams indicating the superposition of the heavens) are engaging, but little more. They are no more than easy summaries for the casual reader, lacking any explanation of the texts on which they are based. Because they are simplicity itself, they enjoy much popularity, even in Polynesian circles, but they are of no scientific help to the serious modern investigator. He, nonetheless, must all too often fall back on such a summary, simply because nothing else exists to help him.

The scholar must be on his guard against an outstanding storyteller. He must take the events narrated, put them into their proper place and space, and compare them with what else is known in as complete a way as possible, so that contradictions in the fabric are revealed. Such contradictions, subtle but evidently present, indicate that an entire population has sided with the storyteller to disguise the real truth, which is in itself an interesting phenomenon for study.

The legendary cycles. There are very few comparative studies of Polynesian and Micronesian literatures. The methods they employ are not always impressive—only authentic texts should be used for comparison, not simplified summaries—but they do give a glimpse of how wide-

Figure 9: Melanesian dance of spirit impersonation, Ambrim Island, New Hebrides, Melanesia.
Kal Muller—Woodfin Camp

The "trickster" figure in legend

spread certain legendary cycles were, the most frequently recurring, from a geographical point of view, being that of the trickster figure, Maui-tiki-tiki, who was a fisherman of the islands and who discovered fire. He can be recognized, on the fringes of the Polynesian area, as the god of the first fruits of the yam harvest on this or that island, sometimes revered under a symbolic manifestation or sometimes as a less abstract figure. Indications of his variety of function over such a large area makes the loss, sustained over two centuries of interpretation willfully blind to the true Polynesian cultural phenomenon, all the more frustrating.

The few scraps of knowledge available about the indigenous literatures of Australia and Micronesia, as well as those of Polynesia, indicate that the figures of their great mythological cycles were simultaneously general symbols and local divinities. The people saw no contradiction in this double manifestation: great cultural heroes were naturally assigned to a specific place when an individual within that culture would establish a reverential dialogue with any one of them. (J.Gt.)

III. The performing arts: music and dance

THE ROLE OF MUSIC AND DANCE

Music and dance in Polynesia and Micronesia are audible and visual extensions of poetry, whereas in Melanesia they are more aimed at spectacular display during times of life crisis and secret-society rituals. The differences between Melanesian and Polynesian music and dance can be related to a basic differentiation in political types that reflects differences in social structure that have been characterized as "bigmen" societies and chiefdoms.

Melanesia. The leader, or bigman, in many Melanesian and New Guinea societies is often a self-made man; he becomes a leader by creating followers, succeeding because he possesses skills that command respect in his society, such as oratory, bravery, gardening prowess, and magical powers. He amasses goods and has great public giveaways, often in connection with the erection of a bigman's dwelling or a men's house, the purchase of higher grades of rank in secret societies, the sponsorship of funeral or other religious ceremonies, or the erection and consecration of slit-gongs (or slit-drums, percussion instruments made from hollowed-out logs or living tree trunks). These ceremonies occasion spectacular displays of the visual and performing arts.

Impersonation of otherworldly beings

There are basically two kinds of dance in these Melanesian ceremonies: dances of impersonation (Figure 9) and dances of participation (Figure 10). In the first type, the dancer impersonates mythical or ancestral beings; the dance-actor becomes someone else, and his attire is usually distinctly unhuman or supernatural—consisting often of huge masks and a full otherworldly costume. The dance movements are dictated by the two considerations that the dancers are not human and that their attire is difficult to move in. Thus the movements are those of legs and swaying bodies; the arms are often covered and frequently used to steady the costume and mask or perhaps hold a drum to accompany the dance. The movements do not interpret recited poetry; however, the voices of these supernatural beings may be heard in the sounds of musical instruments.

The dances of participation are often extensions of these dramatic ceremonies, for individuals who do not impersonate spirits often join in and dance with them, imitating the steps of the supernatural. In dances celebrating headhunting, warfare, funeral rites, or fertility—in which everyone participates—the same movements are used, often to the accompaniment of drumming and communal singing. The dances have a character of spontaneity and do not require long and arduous training. Their aim is not the simultaneous flawless execution of music and intricate movements but, rather, the creation of a mass rhythmic environment that might be characterized as a visual extension of rhythm. If words are associated, they are repetitious and seem not to tell a story; they may even be unintelligible. Although the specific structure of any single dance tradition in Melanesia is not yet known, it seems probable that the isolated units of movement would be primarily those of legs and body.

Polynesia. The entirely different world of Polynesia stands in contrast. Polynesian dance is a visual extension of poetry that uses chant or heightened speech as a vehicle for the praise and honour of high-ranking chiefs or visitors (Figure 11). In Polynesia power resides in chiefly office, and texts tell of a chief's deeds and his descent from the gods. Genealogical rank is a distinctive feature of Polynesian societies, and music and dance pay allegiance to the rank-based sociopolitical structure, reflecting and validating the system of social distinctions and interpersonal relationships. In these societies, where power resides in the office and the regime is long and enduring, specialists compose poetry, add music and movement, and rehearse the performers for many months before a public ceremony. Movements are primarily those of hands and arms, and interpretation is that of a storyteller. The dancer does not become a character in a drama, and his stylized gestures do not correspond to words or ideas as they do in literature-inspired dance traditions of Indonesia and Southeast Asia. In Polynesia the dancer interprets a story orally, usually chanting or reciting metred poetry, and accompanies the words with actions. The presentation is not dramatic in the Western sense of the word, for there is no conflict. Instead, the dancer is storyteller *par excellence*, audibly and visually telling *about* a person, place, event, or emotion, Yet, although Polynesian dance texts are based on traditional stories, legends, or myths, a story is not "told" in the usual sense: traditional literature is referred to in a roundabout way, but the poetry is often the vehicle for

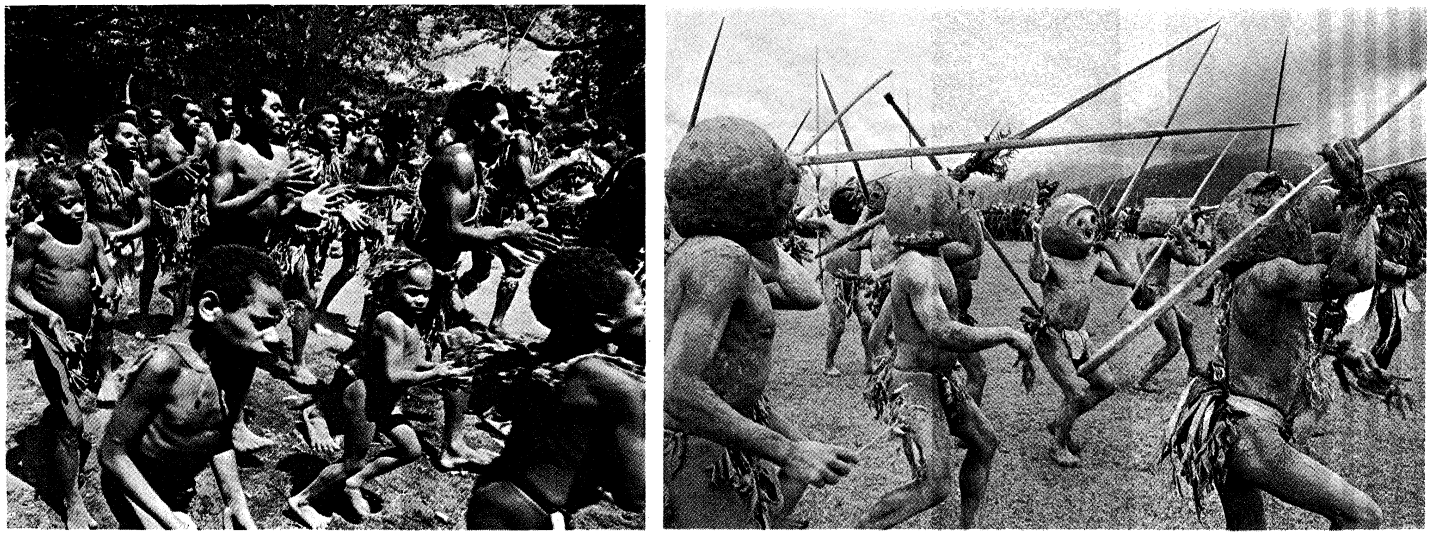


Figure 10: *Melanesian dances of participation.*
(Left) Dancers celebrating a good yam harvest, Tana Island, New Hebrides. (Right) Mudmen of the Asaro River area, New Guinea, performing a dance in which they re-enact their ancestors' defeat of their enemies.
(Left) Kal Muller—Woodfin Camp; (right) © Axel Poignant

saying something else, usually something relevant for the occasion at which it is presented. In addition, the order of the dances and the choice and placement of the dancers often supply further information about the social structure.

The structure is known for at least two Polynesian dance traditions—Tongan, in western Polynesia, and Hawaiian—and the basic units of movement are primarily those of the arms. The only Polynesian dance tradition, however, that has been thoroughly studied is the Tongan. Tongan dance is a visual extension of poetry and is closely intertwined with social organization. This sung poetry is a series of references to mythology, chiefly genealogies, famous scenic places, and contemporary events. The dances, which are performed either standing or sitting, interpret selected words of the text with hand and arm movements. The distinguishing characteristics of Tongan dance are the emphasis on the rotation of the lower arm and the flexion and extension of the wrist, as well as a quick sideward tilt of the head. The legs are used mainly to keep time with sideward stepping movements, and there is a marked absence of hip or torso movement. In pre-European times the most important dance was the *me'etu'upaki*—a paddle dance performed by a large group of men in accompaniment to singing and a slit-gong, which was often played by a high-ranking chief. This dance is still performed today. Group dances called *me'elaufola* were performed by men or women separately in accompaniment to singing, long bamboo stamping tubes, and percussion sticks. The evolved form of this dance, which flourishes today, the *lakalaka*, is performed by men and women together in accompaniment to sung poetry only. Solo and small group dances performed by one, four, or eight women often follow the large group dances and are more concerned with beautiful movements than with interpretation of poetry, although the same movements are used. In the 20th century Polynesian dances are classified into six genres, three of which have survived from pre-European times. The most acculturated dance type, the *tau'olunga*, is a combination of Tongan and Samoan movements and accompanies Western-style singing in conjunction with stringed instruments.

Micronesia. Music and dance in Micronesia, though certainly not the same as their Polynesian counterparts, are closely related to them. With the exception of Truk in the central Carolines, which displays traits of Melanesian and possibly Indonesian influence, the music structure of all parts of Micronesia is predominantly word determined, as is that of Polynesia. Dance movements are mainly of hands and arms in accompaniment to poetry. In some islands, such as Yap (in the western Carolines) and the Gilberts, there is a similar concern for rank in the placement of dancers, as well as the emphasis on rehearsed execution of songs and movements. But, although movements and types of dance have a superficial similarity to Polynesia, there are differences. In the Yap Empire, for example, dances were given as tribute to Yap by Ulithi, Woleai, and other islands, and the dance texts were in languages not intelligible to the Yapese dancers—the function of movements was not to illustrate a story but to decorate it (Figure 12). Instead of acknowledging a chief's deed or genealogy, the Yapese dancers demonstrated the overlordship of Yap to the other islands. Even in Ifalik, where texts were in their own language, the movements did not interpret poetry but were apparently abstractly decorative. The same is true for the Gilbert Islands. Thus Polynesian dance could be characterized as illustration of poetry and Micronesian dance as decoration of poetry, while music in both areas serves as an elevated form of audible performance for poetry.

Dance as ornament to poetry

In many parts of Micronesia, dance and music were associated with tattooing, and with the decline of tattooing has come the virtual demise of these genres. The importance and dependence of the Micronesians on the sea is illustrated in poetry, music, and dance. In some areas dances were performed on the platform of canoes, canoe-like paddles were used in other dances, and in some areas canoe head ornaments were worn by the performers. A study of the influence of the sea on Micronesian music and dance would reveal a great deal about the underlying value systems.

Again, the specific structure of any of the Micronesian dance traditions is not known, but apparently the basic units are primarily those of hands and arms and, if early descriptions are to be believed, the head.

Australia. Music and dance of Australian Aborigines are important elements of sacred ceremonies that re-enact mythological origins of the tribes and insure the continued supply of foods through the propitiation of totemic plants and animals. Little is known about the internal structure or basic movements of the various Aboriginal dance traditions; however, in general terms there are often mimetic movements that involve the entire body to add a visual interpretative extension to the oral literature of the tribe.

MUSICAL INSTRUMENTS

Oceanic cultures have developed a large variety of sound-producing instruments. Some are unique, such as the friction blocks of New Ireland: three to four plaques carved out of a wooden block are rubbed with the hands to produce shrieking or hollow-resonant sounds, depend-

Figure 11: Welcome dance performed as a mark of respect to visitors, Viti Levu Island, Fiji Islands, Polynesia.

ing on size (eight to 80 inches for the entire instrument). Most instruments are not used in musical contexts but for other purposes—for example, to produce the voices of supernatural beings (in Melanesia), as lures (shark rattles), as toys, and for communication.

Musical instruments proper generally lack provisions for musical efficiency and easy handling. Flutes have no or few finger holes and no air ducts. Tuning devices with drums are rare, as are fixed resonators with chordophones (string instruments), which are represented only by simple types of musical bows and zithers. In general, instrumental music is culturally less important than vocal music, and in some areas it is absent altogether. In some other areas, however, such as the Solomon Islands, there are highly developed pan-flute orchestras.

Although some types of instruments—*e.g.*, conch trumpets and slit-drums—can be found in many parts of Oceania (excepting Australia), others occur only locally or are distinguishing features of certain musical areas. Open, hourglass-shaped drums with one membrane are typical of New Guinea and Melanesia, while Polynesians used cylindrical drums. Flutes of various types are usually blown with the mouth in Melanesia, with the nose in Polynesia (nose flutes), and both ways in western Micronesia. In contrast to their simple technical structure, which is the more demanding on the skills of the player, some instruments display elaborate ornamentation related to their function as cult objects in Melanesia, or they may be highly carved and finished befitting their function of honouring gods and chiefs in Polynesia.

REGIONAL STYLES AND TRADITIONS

Melanesia. Melanesia—including New Guinea—houses a multitude of regional musical styles, few of which have been thoroughly investigated. The diversity, which parallels the linguistic situation, is assumed to be a result both of migrations and the relative isolation of ethnic groups due to geographic conditions. Intergroup contacts, including European influences, mostly since the late 19th century, have had a clear influence on present styles.

Generally speaking, Melanesian music tends to be less word determined than Polynesian and Micronesian music. Melody, rhythm, and form appear to be shaped by intrinsically musical principles rather than by text structure and meaning. Song texts often use archaic or foreign languages and are thus frequently unintelligible to all participants in a performance.

Melodic characteristics. Musical scale and melodic movement have been primary criteria in the analysis of Melanesian music. The main types of melodic form are triadic, in which the melody moves exclusively or predominantly on the steps of a triad (three tones, each a third apart, as C–E–G); and pentatonic, which uses five steps within an octave, the melodic structure typically emphasizing seconds, fourths, and fifths. Other types include "narrow," in which melodic movement is restricted to an ambit (range) of a third; and "tiled," in which the melody consists of a sequence of short narrow phrases on different tonal levels, always in a descending order.

The triadic and pentatonic forms

Several attempts have been made to link these types of melodic form with specific cultures within Melanesia. The Dutch scholar Jaap Kunst attributed the tiled type found in the interior of western New Guinea, the Torres Strait, and Australia to "a people who, without doubt, emigrated from Asia to Australia—where the majority of them finally settled—by way of New Guinea and Torres Strait." Triadic melody style has been connected with speakers of non-Austronesian (Papuan) languages in New Guinea and elsewhere in Melanesia, while pentatonic structure is described as an element of Austronesian culture in Melanesia. But, whenever new data become available, previous hypotheses on the distribution of stylistic characteristics and their attribution to ethnic and cultural groupings usually have to be revised.

Also, rarely is the music of an ethnic group found to be based on one single principle of melodic structure. There is usually a mixture of several principles apparent, with one or the other prevailing.

Musical style and cultural context. This mixture of musical structures holds true for the three New Guinea groups whose music has been studied in its cultural context: the Monumbo, the Kate, and the Watut. A more detailed discussion of Kate music will illustrate the stylistic heterogeneity of the Kate, who live in the hinterland of the Huon Peninsula of northeast New Guinea and speak a non-Austronesian (Papuan) language, while some of their neighbours on the coast and on adjacent islands speak Austronesian (Melanesian) languages. A Lutheran mission was established in that area in 1886.

Music of the Kate people of New Guinea

Before the mission terminated their non-Christian religious activities, the Kate shared with their neighbours, specifically the Melanesian-speaking Jabem, Bukawa, and Tami, a secret initiation cult that provided for an exchange of music and dances among participants. The mission introduced "Christian songs" with texts in Kate language and European church tunes; but missionaries also created "Christian" adaptations of traditional Kate melodies, which were more readily acceptable than Lutheran hymns.

By about 1910 the Kate experienced a twofold cultural change resulting from the continuing contact with their Melanesian-speaking neighbours and from the impact of European colonial culture. But many aspects of their precolonial culture were then still functioning or in fresh memory.

Music and dance activities were connected with childbirth, children's games, initiation, hunting, agriculture, ceremonial bartering of pigs, warfare, and death—the

Figure 12: Festival dance, Yap, Caroline Islands, Micronesia.
Burt Glinn—Magnum

latter occasion being the only one to prohibit dancing. Consequently, initiation ceremonies, which usually extended over two years, had to be interrupted whenever a death occurred.

In addition to music and dance in ceremonial context, there were songs for entertainment and expression of individual sentiments or experiences. Most of the common social dances and dance songs were adopted from the offcoast Siassi Islands, including texts that were unintelligible to the Kate.

Structural characteristics of Kate music include triadic and pentatonic melody, both pure and in various degrees of blending; monophony (music having a single voice part); reverting and progressive strophes (stanzas) of varying lengths; basically isorhythmic organization (*i.e.*, using recurrent rhythmic patterns); and a wide range of tempi. Analysis of stylistic characteristics in relation to context and historical data revealed that triadic melody and short progressive strophes are associated with the old, non-Austronesian (Papuan) stratum of Kate culture. Pentatonic melody and longer, reverting strophes seem to represent Melanesian influence, while a specific melody style characterized by the use of several pentatonic modes in succession ("modulating pentatony"), very wide ambit, and extended strophic form can be attributed to European-culture contact.

The Solomon Islands. While the music of New Guinea and western Melanesia—particularly the Bismarck Archipelago—is predominantly vocal and monophonic, the music of the Solomon Islands is largely determined by use of highly developed panpipes. These instruments have three to nine closed tubes, usually doubled by open tubes that sound the higher octave. New instruments are tuned by comparison with old "masters," at the occasion of special ceremonies. Panpipes are used in orchestras alone and in conjunction with song. Vocal melodic style, which is characterized by triadic structure, wide melodic leaps, elaborate polyphony (several simultaneous voice parts), a specific timbre, or tone colour, and frequent change of register, is apparently an imitation of panpipe music. Types of polyphony include parallel melodic movement a third or a sixth apart, a practice most likely a result of overblowing the double panpipes, which may produce such parallels automatically.

Panpipe ensembles

The triadic melody style of the Solomon Islanders seems to have spread into adjacent western Polynesia, where similar melodic types are found in the Ellice Islands and Futuna.

Micronesia. From the Carolines in the west to the Gilberts in the east, most traditional music is accompanied by dancing in standing or sitting posture. Group singing with rhythmic accompaniment by body and ground beats or concussion sticks is the prevailing type of musical performance. Purely instrumental music was performed on nose and mouth flutes in the Carolines and Marianas.

Except in the central Carolines (Truk), where musical influences from Melanesia and eastern Indonesia are prominent, elements of chanting and metred declamation are the most conspicuous characteristics of musical structure, underlining the importance of poetry versus intrinsically musical principles. Vocal polyphony takes the forms of drone (sustained note heard against a melody) and parallel movement in a variety of intervals, with fourths most common.

Micronesia is the least known part of Oceania, as far as music and dance are concerned. On some of the Gilbert atolls, consecrations of assembly halls, races of boat models, and meetings between local groups were or still are connected with performances of dances in which both women and men participated. Active participation and choreographic role are determined by individual proficiency as dancer and singer and by social rank. The *ruoia* is a sequence of standing dances in which movements are slow and mainly those of the arms and hands. In introductory and main dances, up to six leading dancers, male or female, pose as "gliding frigate birds" in front of the other dancers, who are lined up according to status in their patriclans (kinship groups). All dancers participate in chanting long poems that are rehearsed beforehand. Endings of dance songs are frequently shouted, and texts of the final dance in a sequence are recited in heightened speech throughout. There are also sitting dances, with arm and hand movements similar to those of the standing dances, and stick dances. Dance gestures are not illustrative of the song texts, which are not generally understood by performers or audience.

The texts of traditional dance songs were "received" by composers from ancestor spirits (*anti*) in special rituals and probably in trance. Since the early 20th century, multipart singing of European church tunes has spread throughout the area. In consequence of a general culture change, social dances based on traditional movement patterns but accompanied by adaptations of Western music have become dominant.

Polynesia. Since the second half of the 18th century, all parts of Polynesia have undergone a drastic culture change that affected music and dance traditions severely. At present, adaptations of Western musical forms are predominant throughout the area. Both European church singing introduced by various missions and secular styles, ranging from the whalers' chanties to modern international entertainment music, have participated in this process. Yet, remnants or elements of precontact Polynesian music have survived almost everywhere, either alongside of or within acculturated styles—be it because of a marked traditionalism, as in Hawaii and New Zealand, or because of a delayed acculturation process, as in many of the smaller and remoter islands.

Common traits. The first useful descriptions of Polynesian music and dance come from Capt. James

Cook and his companions on his exploration voyages (1768–80). Such reports of early travellers agree with 20th-century research in suggesting that, despite regional variance, the concepts and structural characteristics of music and dance are highly similar throughout Polynesia. Music serves as a vehicle for Polynesian poetry, as dance is its illustration. The central role of the word explains why Polynesian music is primarily vocal. The only noteworthy traditional instruments used independently from song are the nose flute and the musical bow. Accompaniment of song includes body percussions (*e.g.*, slaps, claps), drums, and various idiophones (instruments the bodies of which vibrate to produce sound, such as rattles and slit-drums).

Polynesian songs

The most obvious stylistic characteristics of songs, common to all parts of Polynesia, result from word orientation. Most traditional songs can be classified as chant, as recitation in heightened speech, or as a blending of both. In some areas—for example, the Ellice Islands—the same text may be performed in either style. Chanting uses a limited number of tone levels, mostly a third or a fourth apart, and numerous tone repetitions. Rhythmic organization varies from recitative bound to the accents of the words to strict repetition of rhythmic patterns. Vocal polyphony is widespread throughout Polynesia and indisputedly indigenous. It was noticed and described by 18th-century explorers at a time when foreign influence could hardly have affected the style of music. The most common form of polyphony, reported from almost all Polynesian groups with the marked exception of the Maori (of New Zealand) and Hawaiians, is drone produced by a second part that follows the melody rhythmically while repeating one—usually the basic—tone. All authorities agree that pure drone is a precontact element of Polynesian music. Other forms of polyphony occur locally and are believed to be a result of European influence; especially notable among these is imitational counterpoint (simultaneous, interwoven melodies using melodic imitation among the voice parts).

Beyond these basic common traits, Polynesian peoples have developed musical forms and stylistic characteristics that are distinctive for individual groups of islands.

Hawaii. What is generally known as "Hawaiian music" is the result of the acculturation that began in the early 19th century and that was greatly enhanced by the introduction (*c.* 1820) of Christian hymn tunes. The ukulele, so closely connected with this almost entirely Western style of singing, is a local adaptation of the Portuguese *bragha*, a small guitar imported to Hawaii about 1879. The Hawaiian guitar is a way of playing the European instrument by stopping the strings with a metal bar.

Despite the predominance of Western—and, more recently, Asian—influences, some evolved forms of precontact Hawaiian music and dance have been preserved. Their stylistic characteristics fall well within the limits of what has been described as common Polynesian elements.

The Society Islands. The inhabitants of the Society Islands, whose ritual and profane dances accompanied by polyphonic chanting, nose flute, and drum playing were admired by 18th-century explorers, experienced a particularly rapid and thorough Westernization of their music. A visitor to Tahiti, the largest of the Society Islands, stated in 1838: "If they ever had any native music, it has long been forgotten and no other singing is now heard but hymns and sailors' songs . . ." (Capt. Charles Wilkes, U.S. Navy; see Burrows, 1934). Still, the modern Tahitian *himene,* contrapuntal compositions in as many as six voices, retain some indigenous elements of music structure derived from polyphonic chant. These *himene,* which deal with various subjects besides the Christian sentiments (from which their name is adopted), represent a highly developed form of a hybrid Polynesian–European music style.

The Tahitian *himene*

The Maori. The Maori of New Zealand have lost most of their instrumental music in the process of acculturation but have preserved many of their traditional chants and dances. Traditional chants and dances are classified according to function and contents of the text. Among the more prominent types are the lullabies (*oriori*), the laments (*tangi*), the incantations (*karakia*), the love songs (*waiata aroha*), the historical or genealogical recitations (*patere*), and the dance songs (*haka*). They are either recited in heightened speech or sung on narrow melodic lines undulating around a central tone, *oro.* Rhythm is largely word-bound. Any polyphony is considered a fault of performance. One important aesthetic concept requires a performance to be uninterrupted even by breaks for breathing. Consequently, chants are usually performed by two or more singers who will take breath at different moments. As in all of Polynesia the younger generations favour adaptations of Western music.

Western archipelagoes. Very little is published on the music of the large archipelagoes of western Polynesia such as Tonga and Samoa, whereas for some of the smaller groups—Uvéa, Futuna, and the Ellice Islands—published studies are available.

Before Western contact, music on the Ellice Islands was closely connected with social rank, religion, and magic. There are no detailed descriptions of dances; vocal styles included recitation in heightened speech and chant with drone polyphony (common to most of Polynesia), and triadic melodies resembling those of the Solomons. The Samoan emissaries of the London Missionary Society who converted the Ellice Islanders to Christianity (1861–76) destroyed the traditional social hierarchy and suppressed dances and songs either related to non-Christian beliefs or simply not fitting for their concepts of morality. They introduced pentatonic Christian songs characterized by two-part contrapuntal polyphony resulting from overlapping antiphony (contrasting groups of singers). This "pentatonic antiphony" is believed by some authorities to have developed in Samoa under European influence. By 1900, it seems to have become the predominant musical style on the Ellice Islands for both religious and secular topics. Since 1914, church hymns and school songs in four-part European harmony began to replace "pentatonic antiphony" as the favourite style. By 1960, four-part harmony was the almost exclusive style of church, school, and dance tunes. International "Hawaiian" music is gradually penetrating the islands as mass media and Western musical instruments such as guitars and ukuleles become available. Remnants of the earlier traditions persist only with members of the older generations, although outside interest has stimulated a modest revival movement.

STUDY AND EVALUATION

Music. Although valuable descriptions of music and dance date back to the three voyages of Captain Cook (1768–80) and other early explorers, systematic studies of music structure were hardly possible before introduction of the phonograph as an anthropological and musicological research tool. The Austrian anthropologist Rudolf Pöch was the first to record Oceanic music, during his field research in New Guinea (1904–06). He was followed by many anthropologists, mostly German, who visited Melanesia, western Micronesia, and Samoa during the next ten years. Similar recording activities began in eastern Polynesia only in the 1920s. Proper ethnomusicological evaluation of these early recordings began only much later and is still far from advanced.

The first comprehensive study of one Oceanic music culture (*i.e.,* music, underlying concepts, social role) based on a musicologist's own fieldwork is Helen H. Robert's 1926 study of Hawaiian music, followed in 1945 by Edwin G. Burrows' study of Uvéa and Futuna. Almost all other musicological descriptions and analyses are based on archive materials rather than the authors' own field research. Only in most recent years have ethnomusicologists resumed direct investigations in Oceania, allowing studies to extend in scope beyond solely musical analysis.

Dance. Even fewer scholarly significant descriptions and analyses of Oceanic dance exist. For some island groups, reports from the voyages of Cook and other early explorers furnish largely untapped raw materials for studies in the ethnohistory of dance as well as for culture change. Much of this early journal material on dance has been extracted and included by Johannes C. Andersen in

his 1933 study of Maori music. Some of the Micronesian literature was reviewed by Mary Browning in 1970 (in *Dance Perspectives*). Nothing comparable has been done for Melanesia or Australia. Scattered references can be found in the "classic" ethnographic studies, such as the Thilenius Expedition for Micronesia, John Layard for Malekula, and Margaret Mead in Samoa.

It was only in the 1960s that a few adequate, systematic studies and film documentations began to appear. Several excellent films of Australian Aboriginal dance in its ceremonial context have been produced by the Institute of Aboriginal Studies, Canberra, Australia. The only scholarly study of dance in Melanesia is Allison Jablonko's 1968 study of the Maring people of New Guinea, accompanied by a film, *Maring in Motion.* There are also dance films and descriptions for the Gilbert and Ellice Islands by Gerd Koch. On Polynesian dance, two excellent films of Hawaiian dances have been produced, and the structures of Hawaiian and Tongan dances have been analyzed by Adrienne Kaeppler. On the whole, both systematic documentation and penetrating interpretation of Oceanic music and dance still remain promising fields for future research. (Di.C./A.Ka.)

IV. The visual arts

GENERAL CHARACTERISTICS

Form and technology. It has often been suggested that the arts of Oceania go back to the cut or polished tool phases of the Stone Age. There is every reason to regard this as an untenable approximation. An early hypothesis that the Polynesians, for example, or at least certain components of their population, were descended from a society that had known metals is not at all unlikely. Oceanic solutions to technological problems in no way suggest the residue of a prehistoric age. On the contrary, they appear to be the result of rational choice, aiming at efficiency and economy. The "progressive" methods later brought by Europeans all too often proved to be clumsy, costly, and deceptive, as far as the organization of daily life and agricultural production were concerned. There are many convincing examples of the Oceanians' capacity for scientific reflection and deliberate experimentation, such as their adaptation of traditional fighting techniques in order to resist the Europeans and their organization of efficient methods of cultivation without the slightest assistance from their conquerors.

Materials. Oceanic art was largely governed by what material means were at hand (Figure 13). In a physical environment whose vegetation, though not luxuriant, was thoroughly understood and utilized, the people there developed a variety of working textures such as Western industrial society, for example, cannot match. Locally established Westerners have turned with immense satisfaction to the materials available there, adopted them as their own, and even transformed them into luxury items. Both as working materials and as a source of colour, the vegetable compounds available to the Oceanian artist could satisfy all of his creative requirements (especially if it is borne in mind that his artifact was never intended to stand in a home with central heating). Contemporary Western experiments in new means of artistic expression do not rival the multiplicity and complexity of Oceanic art. Today in the West, for instance, works of art are being produced by a collage of nonmetallic materials—a process long familiar to the Oceanic artist, who knew how to combine a wide variety of materials using special vegetable glues. These artists discovered many techniques for physically tranforming plant materials, including pounding, chewing, and cooking. The soft paste they obtained was used to remedy any deficiencies on the surface of the principal support (which might be wood, stone, tree fern, rattan basketwork), and it also furnished an absorbent surface to receive vegetable pigment (Figure 6). The form of the finished work was thus not determined by the size of the wood or piece of stone; the work could be extended in any or all of its dimensions to whatever proportions were required to harmonize with, for example, the architecture of the great ceremonial houses.

Techniques of transforming plant materials

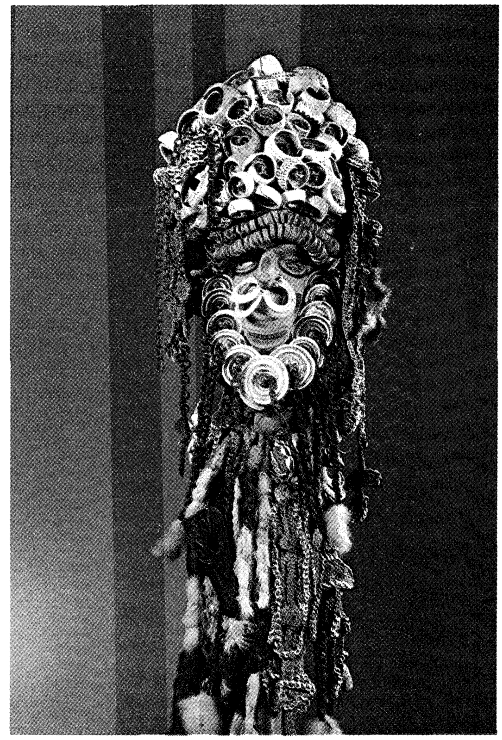


Figure 13: Variety of materials used in Oceanian art exemplified by a Melanesian decorated flute figure, wood, snail and cowrie shells, fibre, feather, opossum fur. From the Sepik area, New Guinea, Melanesia. In the American Museum of Natural History, New York City. Height 68.6 cm.
By courtesy of The American Museum of Natural History, New York

MELANESIAN VISUAL ARTS

Knowledge of metals. The Melanesians have long suffered from a prejudice whereby people who only wish to remain themselves are thought of as barbarians. Their determination, however, kept their culture alive—at least until the mid-20th century, for now it is being swept away by an accelerated process of economic development imposed and controlled by foreigners. This longer duration (outlasting traditional Polynesian art by a century) has allowed observers to discover some perhaps unexpected factors concerning the Melanesians—for example, their adaptation to iron, introduced by Europeans at the end of the 18th century. Hoops from empty barrels, which had contained wine or salted meat, were collected and cut up; the pieces were sharpened and made into carpenters' adzes. These homemade tools were distributed more widely than the "proper" axes sold at a higher price by the crews of Western warships. The technique and form of wood carving was not fundamentally affected, but the results were. First of all, the slow, carefully controlled process of firing green wood to produce a material that could be carved easily was no longer necessary: with the metal axes, wood fibres could now be cut cleanly. Next, there was the new possibility of achieving a right angle—or even an acute one—instead of being obliged to have only obtuse angles (as had previously been the case, at least in the larger works; small pieces, which could be carried by hand, might always be shaped as desired by using stone splinters or animal teeth).

Although this theory of the introduction of metal seems attractive by reason of its neatness and the cultural area in question has enough consistency to lend itself to a linear analysis of its development, it is well to proceed with some caution. The coastal populations adapted with suspicious speed to the relatively massive introduction of metal at this time, and it is as well to remember that the whole area of Serera (Geelvink) Bay, New Guinea, could not have been totally unaware of metals (indeed, bronze blades have occasionally been unearthed there, which indicates a lengthy period of contact with metal). European vessels during the 17th and 18th centuries took on fresh water and provisions there; some of them must

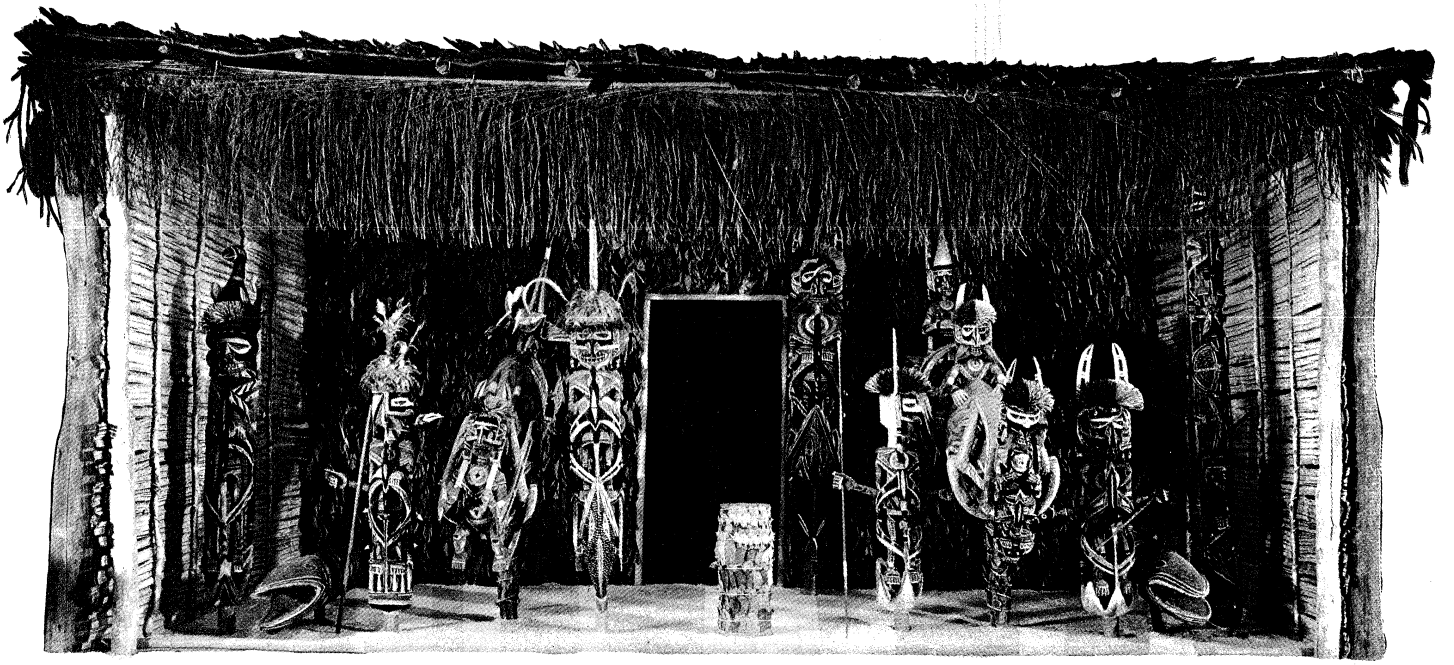


Figure 14: Cult house with *malanggans* from Medina, New Ireland. The house is made of wood, bamboo, small palm leaves, and croton leaves. The figures are carved in wood and painted with earth pigments and yellow oil colour. The house was erected to commemorate 14 deceased persons from the village. In the Museum für Völkerkunde und Schweizerisches Museum für Volkskunde Basel, Switzerland.

By courtesy of the Museum fur Volkerkunde und Schweizerisches Museum fur Volkskunde Basel, Switzerland; photograph, Hans Hinz, Basel

also have been shipwrecked, and the debris washed ashore would undoubtedly contain pieces of metal. Earlier still, Malay *proas* (boats) and even Chinese junks had sailed along the coasts of New Guinea; and how far the trading in objects of Chinese origin was carried on (it occurred principally along the northern coast of western New Guinea, from Indonesia) is not known. The Polynesians, at least, seem to have been already familiar with metal by the time it was "introduced" at the end of the 18th century.

Diversity and range. It is impossible to define "Melanesian art." Melanesia is a geographical concept only, designating an arc of islands—thousands in number—between New Guinea and New Zealand. Its cultural diversity is extraordinary and absolutely defies classification. Its art is no exception. Distinct regions must be treated separately; for local artists, like those of other cultures, have different ways of transforming various components into an original work. Even in the interior of one island there are many quite different plastic traditions, though there are, of course, also elements in common, such as the *uli* and *malanggan* carving traditions of New Ireland (Figure 14). Aesthetic elements are absorbed and diffused so quickly that an alteration of the distribution map of themes and methods would be necessary at least every 50 years. Striking similarities in palette and the use of comparable vegetal materials—as used, for example, in Papua and Arnhem Land (northern Australia)—have not led to a similarity of artifact or style. Even if an influence of one on the other or even an interchange of cultural elements can be supposed as taking place between New Guinea and the area of Australia opposite, the artistic products of each side certainly have no need of this proximity to explain their own existence. Ignorance on the matter is, indeed, considerable, because no one observed the technology behind the production of these objects at the time they were being made, and there is no museum equipped for the scientific study of this essential aspect. The efforts that have been made to understand everything about easel painting have not yet been extended to what is called primitive art.

Cultural diversity of Melanesian art

If, however, the plastic art of Melanesia is to be characterized on the basis of very slim knowledge, attention might first be called to its wide range, which takes in the most durable, finished stone sculptures as well as pieces so delicate that they could be rightly called arachnidean, for they actually use cobwebs as material. Broadly speaking, however, there are two tendencies to be distinguished. One leads to stereotyped art in which the individual variations of each piece lie in the small detail, discernible only on careful examination. The other favours a form of continuous creation, much of it achieved by wholly original plastic innovation, the rest by a double process of first analyzing existing formal elements and then constructing new ensembles by a partial or total regrouping of these same elements.

Stereotyped art. The first tendency can operate best within the framework of a society itself unconcerned with innovation, whose local artists specialize in a way that makes it possible instantly to recognize a piece of their work, regardless of where it may be found. This specialization need not be artistic so much as a matter of craftsmanship. The Huon Gulf and the island of Tami and the Admiralty Islands, especially, have organized specialization by village for the production of useful articles—ladders, house posts, bedsteads, mortars and pestles and spatulas for preparing and spreading lime, and obsidian dagger handles and spear points, wood and coconut-shell spoons, amulets of war or the dance, pirogue (canoe) prows, and platters of various sizes (Figure 15). Each individual member of the given society is a more or less skillful producer of one category of objects.

Innovative art. Sometimes the specialization is a prerogative of individuals because of heredity (as in the case of Big Nambas, the sculptor of tree-fern ridgepieces, or *ponarat*, of North Malekula in the New Hebrides), and sometimes because of a personal vocation (as on Ambrim in the New Hebrides). In these cases, products, though less numerous, become more accessible to classical aesthetic analysis—because of their artistic coherence and style rather than their finish. It also becomes less difficult to attempt tracing a chronology of these works with the help of the oral tradition and, sometimes, of archaeological discovery.

Individual specialization

There are many obstacles to be overcome in analyzing the aesthetics of Melanesian art. There is not even a documentary language adequate to describe the human body, so often the essential theme of plastic representation. A simple description of an object has no value at the level of scientific interpretation unless it is closely linked

Figure 15: *Melanesian articles of everyday use.* (Left) Wooden lime spatula. From the Trobriand Islands, New Guinea. In the Ethnographical Museum, Budapest. Length 30.5 cm. (Right) Spoon fashioned from a coconut shell. From the Admiralty Islands. In the Ethnographical Museum, Budapest. Height 21.9 cm.
By courtesy of the Ethnographical Museum, Budapest

to the most exhaustive iconography possible. The more prominent characteristics of a work cannot meaningfully be isolated from the piece as a whole. The terms of the art historian's vocabulary have proved inadequate to convey the inventiveness and refinement of these artists. The eye, framed by some orbital motif—circular, oval, lozenge-shaped, double- or triple-comma-shaped—is a characteristic of the area that includes the Sepik Valley and the northeast coast of New Guinea (Figure 16). But the

Holle Bildarchiv, Baden-Baden, West Germany

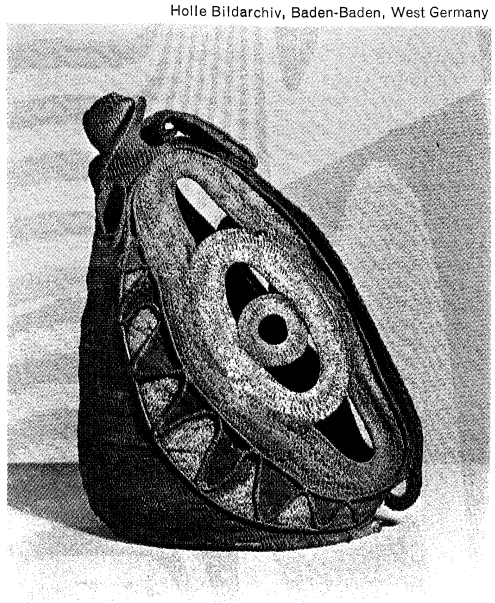


Figure 16: Mask with orbital motif, rattan painted with earth pigments. From the southern Maprik area, New Guinea, Melanesia. In the Museum für Völkerkunde und Schweizerisches Museum für Volkskunde Basel, Switzerland. Height 54.6 cm.

variations are countless; and the only scientifically useful definition of them would be a list of every known case, with statistics of each variation and a comprehensive index. The result would be unreadable and probably useful only in the task of settling the origin of pieces that, stylistically, might belong to more than one region.

Classification by natural materials. It is perhaps the factors of environment that offer the best means of classification. There are, for example, elements common to all the areas forming the immense marshy stretches of New Guinea: palm leaves, interlaced and stitched together, are used as a surface to be painted (Figure 17); trees, whose outer wood at least offers enough resistance to the chisel, are used by sculptors such as those of the Asmat coast (Figure 5, left); beds of coloured clay along the raised banks of the rivers provide a plentiful supply of pigments. There are, of course, local variations of material and technique. In Papua the bark of the mulberry tree is treated and used to make masks, both secular and ceremonial (Figure 6, top left), whereas this material is not seen in works from Sepik and the Asmat coast. The bark, however, is again used for masks made in the mountainous regions toward the northern coast of New Guinea, as well as on the islands between that coast and New Britain, and also in the Walonga (Baining) masks of northern New Britain itself (Figure 17).

Limited knowledge of art in the New Guinea interior

Perishable art of New Guinea. The high plateau and mountain valleys of the New Guinea interior have yielded so little that they might seem to constitute an aesthetic void. Such a judgment does not, however, take into account some fragile, perishable works that are made to be destroyed immediately after use. Beyond the fact that such objects are indeed made, little else is known about them.

The archipelagoes. More information about the archipelagoes—which are, properly speaking, Melanesian—awaits proper research into a number of historical collections, especially that of the British Museum, which have yet to be made public. Representations of human, beast, and fish figures from the Solomon Islands (Figure 7), notable for their austere, stylized simplicity, suggest an area of common influence that includes Merat Island (north of western New Guinea); the Huon Gulf and Tami; the Massim area; and the Admiralty Islands, even extending to western Polynesia (including Fiji) and central Polynesia. The figures also have features in common with some that have been dug from peat bogs in New Zealand.

Southern Melanesia. At the southern end of the Melanesian arc, however, from New Caledonia to the New Hebrides, there exists the familiar contrast between some areas, where there is a stable tradition of sculptors specializing in one or another type of mask or in monumental decorative pieces for the large clan house, such as doorjambs (Figure 18) and lintels, posts, ridge pieces, and other areas where there is an inventive culture whose aesthetic coherence is closely connected with public or partly esoteric rituals and where the art of carving is linked to a right that is sold and bought, but not inherited. Technological ingenuity, using vegetal elements, is greatest in the New Hebrides, especially in south Malekula (Figure 6, right). The Loyalty Islands, however, and the southern New Hebrides (Aneityum, Tana, Eromanga) offer little of interest other than monumental architecture (Loyalty Islands), weapons for close combat, plaited articles, and elements of body decoration.

MICRONESIAN VISUAL ARTS

Micronesia is culturally an area of transition in every way. Known facts and common sense link Micronesia—with the exception of such Polynesian islands as Kapingamarangi, or the southern Gilberts, and the Ellice Islands, where there is Samoan influence—with the aboriginal cultures of Taiwan, the Philippines, and the islands situated off the northern coast of New Guinea, with which there were continuing relations in both peace and war. Micronesian pirogues reached the Solomon Islands (especially the "Polynesian outliers," where Micronesian components form an important part of the basic population) and the northern New Hebrides. Little is known of the countless comings and goings between Micronesia and Indonesia, though Spanish archives give some account of these contacts.

Because this was a warrior civilization, living in fear of powerful enemies likely to come from the east, either

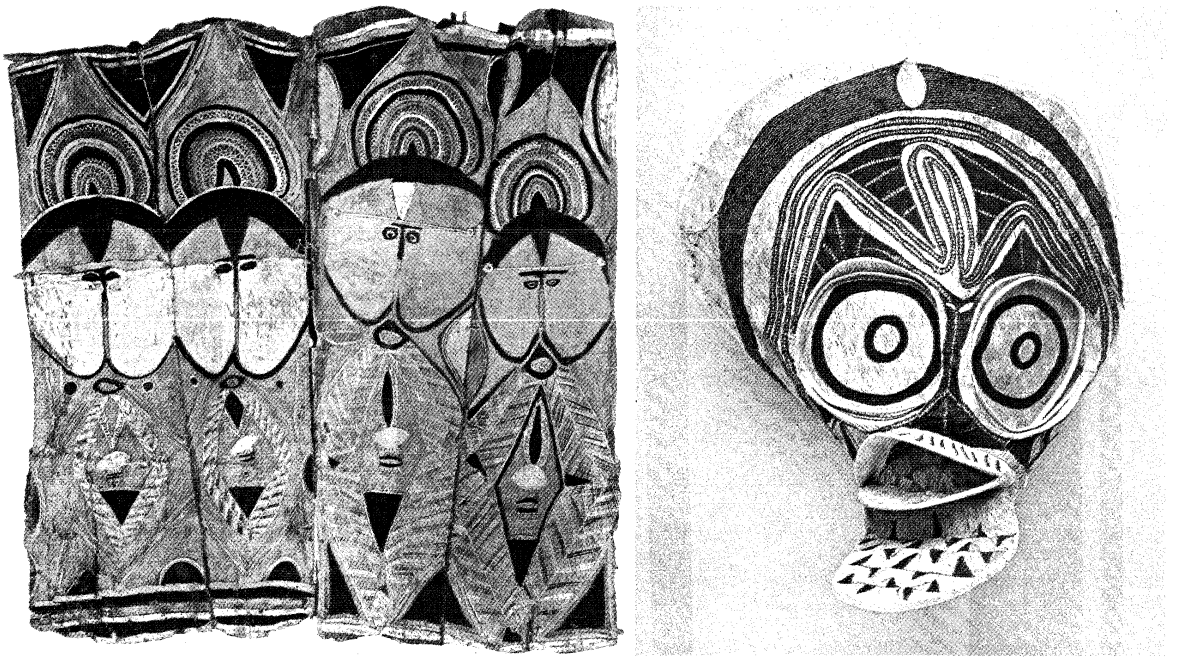


Figure 17: *Natural materials used by Melanesian artists.*
(Left) Painted plaques made of interlaced and stitched palm leaves. From the southern Maprik area, New Guinea. In the Museum für Völkerkunde und Schweizerisches Museum für Volkskunde, Basel, Switzerland. Height 78.1 cm. (Right) Tapa (bark) cloth mask made by the Walonga (Baining) tribe of northern New Britain, New Guinea. In the Museum für Völkerkunde und Schweizerisches Museum für Volkskunde Basel, Switzerland. Height 79.4 cm.
By courtesy of the Museum fur Volkerkunde und Schweizerisches Museum fur Volkskunde Basel, Switzerland; photographs, Hans Hinz, Basel

expecting or actively preparing for warlike expeditions within their perimeter, art tended toward the monumental (raised stones, monumental doorways) rather than the figurative. The technique in wood was, however, that of cabinetmakers, and household articles, such as the platters and stools used when grating coconuts, show modern functional forms. The artists knew how to work with mother-of-pearl inlay (Palau Islands), and their technique of plaiting ribbons cut out of pandanus leaves resulted in a delicacy of work that has something in common with certain Indonesian fabrics. The large gable figures of the Caroline Islands representing the human face (Figure 19, left) and the standing *tino* statuettes (Figure 19, right) seem, by their perfection, to be the result of a very long cultural development. In the final period, the Micronesians devoted their energies and ingenuity to organizing complex political systems, establishing a hierarchy of their islands, so that there was perhaps no longer a place for artistic creativity.

POLYNESIAN VISUAL ARTS

General characteristics. Polynesia is a world that ought to be better understood than it is. The West, for instance, has been aware of it since the 18th-century voyages of Capt. James Cook and Louis-Antoine de Bougainville. But it is difficult to sort out from old documents what is useful information and what stems from the romantic imaginations of 19th-century observers. Later research, such as that undertaken by the renowned Bernice P. Bishop Museum in Honolulu, though seriously intended, too often produces only unrelated items of information, bits and pieces of folklore, which never cast much steady light on how the society under consideration functions.

In addition, much essential information was early destroyed, partly by Western missionaries, some of whom were bigots, and partly by the Polynesians themselves, who gave way with such alacrity before the power and wealth of the West that they became iconoclasts, destroying their own culture. An early-19th-century missionary, the reverend John Williams, managed to preserve some of their religious sculptures by hanging them from the yards of his ship, as "testimony" to the destruction of the devil's reign, so that the Polynesian evangelists would agree to spare the "pagan idols" from the pyre they had already prepared. These sculptures are now in London (Figure 8). But many specimens of plastic art did not escape. There is no way of knowing whether what has been preserved is wholly representative, or, if so, of what.

Making do with what is known and awaiting the publication of scientifically inferred catalogs, it is possible to advance the idea cautiously that Polynesia offers the most stereotyped—and in many ways the least colourful—visual arts culture of the whole Oceanic area. There are striking exceptions: Hawaiian wickerwork (either hard—to make representations of gods—and supple—to provide royal cloaks), with its outer covering of coloured feathers, displays an extraordinary iridescence (Figure 20).

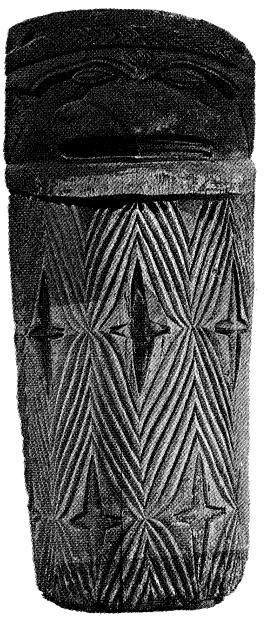


Figure 18: Wooden doorjamb. From Ouebia, New Caledonia, Melanesia. In the Museum für Völkerkunde und Schweizerisches Museum für Volkskunde Basel, Switzerland. Height 1.65 m.
Holle Bildarchiv, Baden-Baden, West Germany

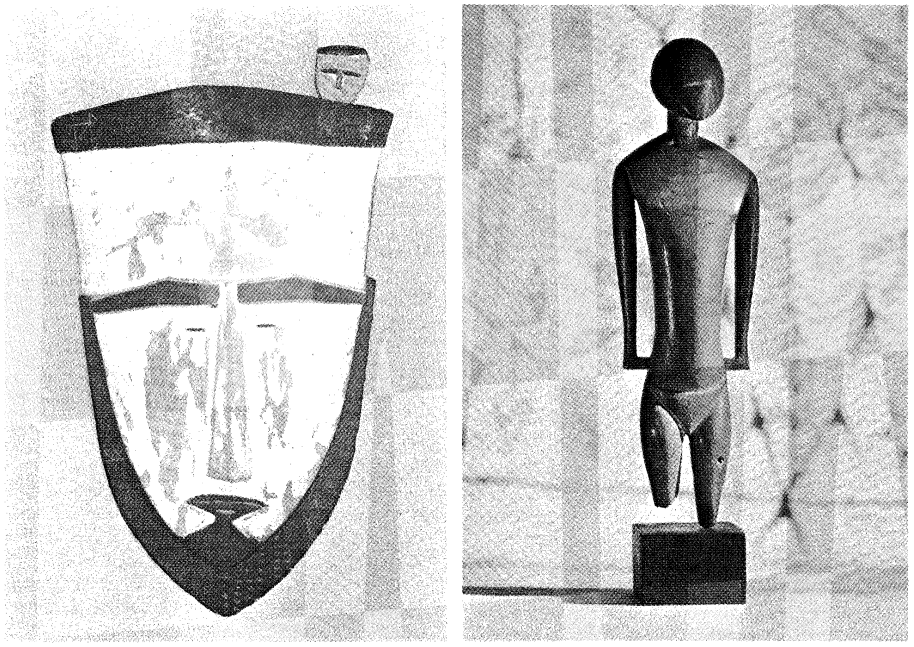


Figure 19: *Micronesian art.*
(Left) Wooden gable-end ornament. From Mortlock Island, Caroline Islands. In the Museum für Völkerkunde, Hamburg. Height 45.1 cm. (Right) Wooden *tino* figure. From the Caroline Islands. In the Musée de l'Homme, Paris. Height 33 cm.
By courtesy of (right) the Musee de l'Homme, Paris; photograph, (left) Holle Bildarchiv, Baden-Baden, West Germany

Protective function of violent colour

New Zealand. Violent colours—red, black, jade green—are also employed in the indigenous art of New Zealand but with little subtlety. Their function is rather to provide a coloured background or even to protect perishable surfaces such as wood or interwoven reeds.

Using hardwoods and volcanic stone—materials that would not so easily be destroyed by time—and patiently carrying out the arduous task of carving with their inadequate tools, the Maori became meticulous sculptors (Figure 3). They showed a preference for squat human figures approximating the fetal position, whose faces, with their globular eyes, recall the people's dread of miscarriage and were indeed a reverent symbol of that to which life has been denied. These works display an astonishing mastery of technique: the Maori sculptor, for example, passing from the representation of the face to that of the profile in the bas-reliefs adorning the great reception houses found a technical solution to the problem more advanced than that of the ancient Egyptian sculptor. Maori art remained alive the longest, though basic changes and innovations were produced by the introduction of iron. Thanks particularly to the results of a half

By courtesy of the trustees of the British Museum; photograph, (left) J.R. Freeman & Co. Ltd.

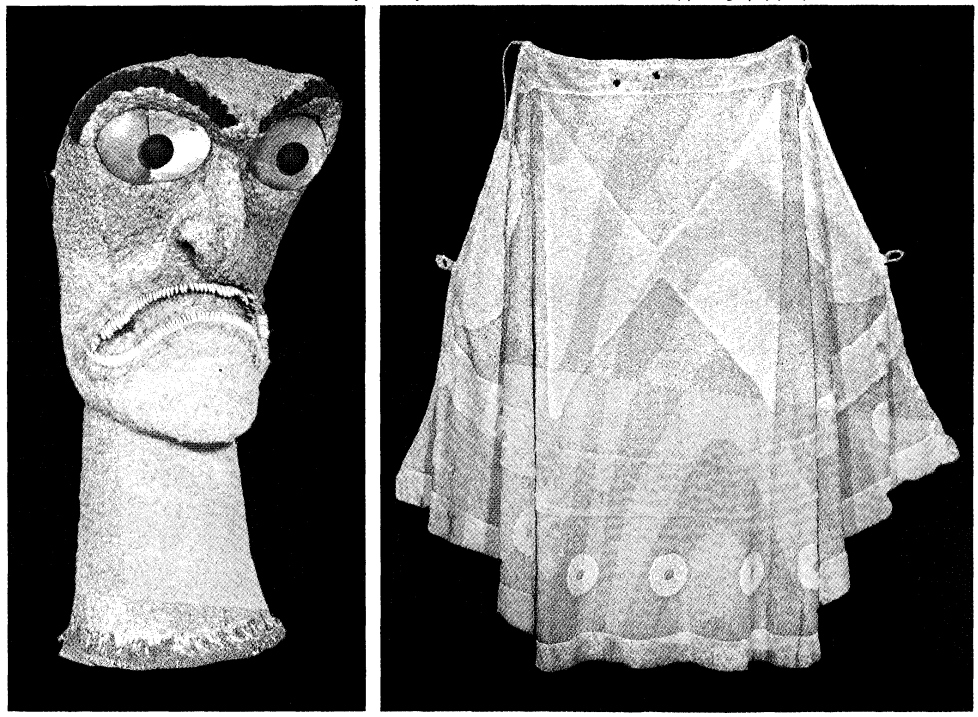


Figure 20: *Hawaiian wickerwork.*
(Left) War god Kukailimoku, wicker framework covered with netting into which feathers are knotted. In the British Museum. Height 81.3 cm. (Right) Royal cloak, wicker covered with netting into which feathers are knotted. In the British Museum. Height 1.75 m, overall diameter 2.89 m.

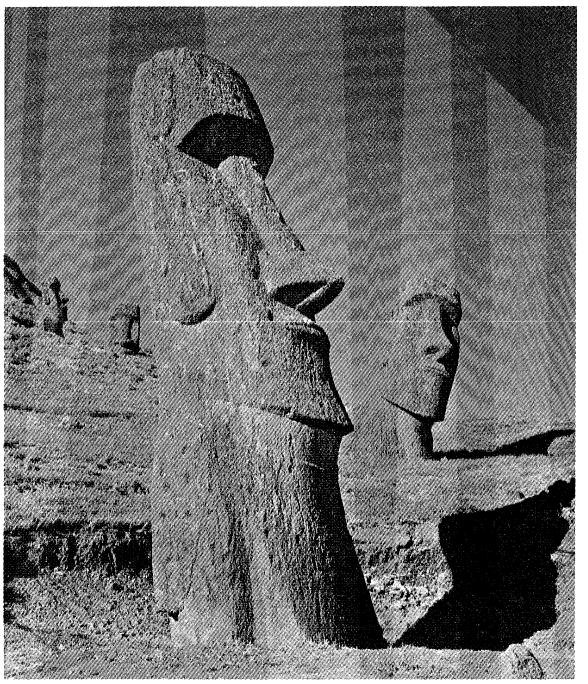


Figure 21: Sculptures cut from volcanic rock, Easter Island, Polynesia.
Shostal–Ernest Manewal

century of local archaeology, Maori art is perhaps the best documented in the Oceanian area.

Marquesas Islands. Like New Zealand, the Marquesas Islands offer both monumental sculptures and small objects, the latter often more valuable for the study of style. Nothing like the exquisite boxes for feathers or the votive statuettes of the Maori are to be found there; but the minute scenes that are represented in filigree on ear ornaments and, with less animation, on Marquesan tortoiseshell plaques or fan handles (Figure 2), are very attractive. Also of great interest is the art of tatooed body decoration practiced there. Decoration may cover the whole body, especially the man's, with contrasting light and dark lines and bands and elaborate designs. Stone sculpture is not used so much to represent gods as to elaborate the main structural points of an open architecture consisting of terraces that are raised one above the other.

Easter Island. Easter Island will always arouse astonishment for the ingenuity of a people on a barren windswept island who, by managing volcanic tufa and utilizing the strength of hundreds of men, erect huge, heavy sculptures (Figure 21). Many of them are adorned with tufa hats, and their angular faces top bodies that were unknown for a long time afterward because they were buried in the Earth. The Easter Islanders also carved twisted forms from the branches of the only available shrubs. The forms are human representations that reflect the harsh conditions of life in a climate in which men are poised between life and death each time a prolonged dry spell scorches all cultivation, leaving only seafood as possible nourishment—and this only on days when rough seas do not make fishing an impossibility. The peoples of this inhospitable island also produced ideographic signs that are still in the process of being deciphered. They seem, as much as anything, to have been devices that were used to aid the memory during religious recitals.

Nonfigurative art of Tahiti

Tahiti and the surrounding islands. In eastern Polynesia, Tahiti experimented in nonfigurative art: a bundle of coconut fibres, with shell pendants, would represent the figure of the god; tapa cloth, obtained from the bark of the mulberry tree, would be given a thin wash of colour obtained from fern leaves and enjoyed simply for its decorative value (Figure 22). Art in the surrounding islands, and also in the Îles Tubuai (Austral) group, is related to that of the Cook Islands, where a considerable part of the population claim to have come from the Society Islands. Sculptures of humans are not common; but the decoration of fan handles (Huahine), of votive axes (Mangaia in the Cook Islands), of ceremonial ladles and vertical drums (Îles Tubuai) displays great subtlety. Designs, both engraved and in relief, are made up of repeated, stylized representations of a figure with outstretched arms and legs. The statuettes of Rarotonga (representing the god Te Oro and his sons) or the representation of Tangaroa (Figure 23) creating mankind from the human forms that burgeon from his own body (Rurutu, Îles, Tubuai) show the mastery of relief that Polynesian sculptors achieved, especially considering the limits that were imposed by the myth itself on its plastic representation. The perfection of form and decoration of ordinary objects—pestles to crush the taro root in preparing poi, or the countless fishing weights bearing a human face or fish form—ought also to be noted.

AUSTRALIAN VISUAL ARTS

In many parts of Australia man has had to make a great effort to survive, especially when faced by the desert conditions of the centre or the semidesert conditions to the north during the dry season. Iron was not known, and the camel was introduced only in recent times. Incomparable walkers, their naked bodies tough and resistant to hardship, the Australian Aborigines strove to maintain their culture against colonial contempt and the repugnance of Christian missionaries, with the result that much more is known today about them than about the Polynesians, who were treated infinitely better by white colonialists.

Since Sir Baldwin Spencer and Francis James Gillen (late-19th–early-20th-century British ethnologists), a host of researchers have carried out a thorough survey of the aboriginal cultures. Although not so much attention was paid to material culture and art as to other aspects of their societies, local museums display rich collections, both old and new, which are somewhat better documented than in other cases. A substantial amount of work remains to be done, but there are qualified informants still living who could give more information. It has been possible, however, thanks to the persevering efforts of Karel Kupka, to discover that colours and motifs were appropriated by certain social groups (the Dua and Yiritya dualities) but that the superposition on this normative framework of systems of relations and interpersonal exchanges, implying, from one bark painter to another, prestations in the form of license to use colours and themes belonging to others, means that no work is strictly what it should be and that to decipher it one must be familiar with the men and groups involved (Figure 1).

Abstract art of the central desert. Similarly, it is known that in central Australia, the engraved decorations to

Holle Bildarchiv, Baden-Baden, West Germany

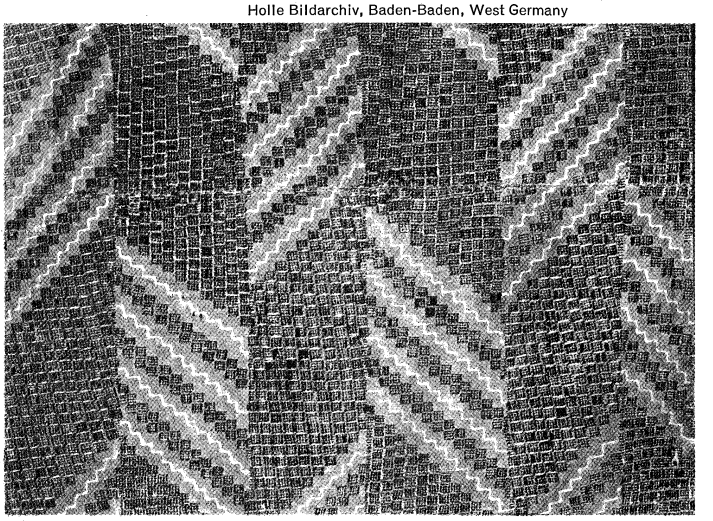


Figure 22: Detail of a tapa cloth printed with stamps and painted. From the Hawaiian Islands, Polynesia. In the Sammlungen des Instituts für Völkerkunde der Universität, Göttingen, West Germany. Size of detail 47 cm x 64.1 cm.

Figure 23: Wooden image of the god Tangaroa creating man and lesser gods. From Rurutu, Iles Tubuai (Austral Islands), Polynesia. In the British Museum. Height 1.11 m.
By courtesy of the trustees of the British Museum

The *tjurunga*

be found on stone plaques (*tjurunga*) served the Aborigines as an aid in reciting their myths of origin. The art is abstract, and the various elements—stippled decoration, half circles, concentric circles, and parallel lines—can be understood to represent specific themes, such as seated man, tree, footprints, water hole, mythological figure, and so forth. It does not, however, represent a coherent ideographic language that can be deciphered without consulting those Aborigines who inherit a knowledge of its significance.

Ritual structures. An important part of Australian Aboriginal art consists of ritual structures, usually planted in the soil, made from a wooden framework on which strands of human hair are strung, stuck on with human blood often drawn from the male sex organ. Other works, using the same materials, were laid out over the surface of especially prepared ground. Both kinds of art were destined to be destroyed immediately after the rite, and only rare examples are known.

Rock art. Australian rock art is certainly the richest in Oceania, perhaps because ecological conditions, including the dryness of the climate, have ensured a better state of preservation for these works (Figure 24). Another factor, this time a ritual one, explains this good state of preservation. It is now known that, in a great many cases, paintings that are to be seen in caves and rock shelters are restored each year at the end of the dry season, on the occasion of what is known as a multiplication rite. To retrace the outlines of the figures, say, of a kangaroo, ensures the multiplication of that species and thus a fruitful hunt during the rainy season—the season of full stomachs. The same is true of other animal figures, including human representations, as an assurance that the group will not be in danger of extinction. (J.Gt.)

Holle Bildarchiv, Baden-Baden, West Germany

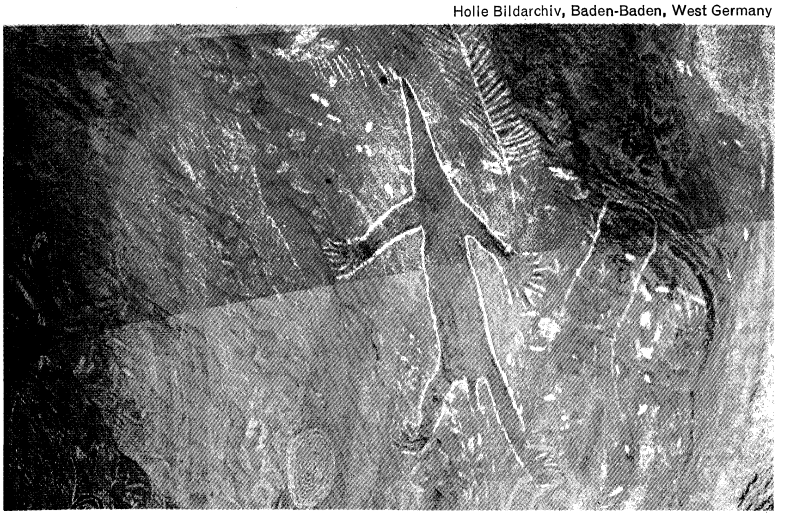


Figure 24: Rock painting of a lizard-like animal, Hawker, South Australia.

BIBLIOGRAPHY

Literature: There are no comprehensive surveys of the literature of the Oceanian peoples. The following are representative of the types of works available: JOHANNES C. ANDERSEN, *Myths and Legends of the Polynesians* (1928, reprinted 1969); CARL STROVEN (ed.), *The Spell of the Pacific: An Anthology of Its Literature* (1949); PHILIP SNOW (comp.), *Best Stories of the South Seas* (1967); JAMES NORMAN HALL, *The Forgotten One, and Other True Tales of the South Seas* (1950); WILLIAM ARMAND LESSA, *Tales from Ulithi Atoll: A Comparative Study in Oceanic Folklore* (1961); JOHN FRANCIS STIMSON, *Songs and Tales of the Sea Kings: Interpretations of the Oral Literature of Polynesia* (1957); BACIL F. KIRTLEY, *A Motif-Index of Traditional Polynesian Narratives* (1971); INEZ HANES, *Folk Tales of the South Pacific* (1969).

Music: WILLIAM P. MALM, *Music Cultures of the Pacific, the Near East and Asia* (1967), a general survey useful as an introduction; HANS FISCHER, *Schallgeräte in Ozeanien* (1958), the only comprehensive study of sound instruments in Oceania. (*Melanesia*): JAAP KUNST, *Music in New Guinea* (1967), three authoritative studies covering western New Guinea (West Irian); DIETER CHRISTENSEN, *Die Musik der Kate and Sialum* (1957), a treatise on the music of two northeastern New Guinea tribes, based on data collected 1908–10; HERBERT HUEBNER, *Die Musik im Bismarck-Archipel* (1938), the only extensive study of Melanesian music styles outside New Guinea, but the approach and conclusions are now obsolete. (*Polynesia*): JOHANNES C. ANDERSEN, *Maori Music with Its Polynesian Background* (1934), covers Polynesia in general, with an emphasis on the Maoris of New Zealand—a useful compilation of relevant excerpts from the travelogs of 18th- and 19th-century witnesses; HELEN H. ROBERTS, *Ancient Hawaiian Music* (1926, reprinted 1967), a thorough, comprehensive study; EDWIN G. BURROWS, *Native Music of the Tuamotus* (1933), a study of Tuamotu chant in its cultural context; and *Songs of Uvea and Futuna* (1945), a study of the music cultures of two small West Polynesian islands, based on the author's field work in 1932; DIETER CHRISTENSEN and GERD KOCH, *Die Musik der Ellice-Inseln*

(1964), a study of the music history of this West Polynesian archipelago, based on field research in 1960–63. See also JAAP KUNST, *Ethnomusicology*, 3rd ed. (1959, with suppl. 1960), the largest and most comprehensive ethnomusicological bibliography, with indexes—covers Oceania very well.

Dance: GERD KOCH, in the *Encyclopaedia Cinematographica* (E-415–418, 915–920, 1962–68), dance films and descriptions of Gilbert and Ellice Islands groups; MARY BROWNING, *Micronesian Heritage* (1970), a review of early descriptions of and writings on Micronesian dance.

Visual arts: GILBERT ARCHEY, "South Sea Folk," *Handbook of Maori and Oceanic Ethnology*, 2nd ed. (1949) and "Sculpture and Design: An Outline of Maori Art," *Handbook of the Auckland War Memorial Museum* (1955); TERRY BARROW, "Maori Decorative Carving: An Outline," *Journal of the Polynesian Society*, 65:305–331 (1956); TIBOR BODROGI, *Oceanian Art* (Eng.. trans. 1959) and *Art in Northeast New Guinea* (1961 Eng. trans.); ALFRED BUHLER, TERRY BARROW, and CHARLES P. MOUNTFORD, *Ozeanien und Australien: Die Kunst der Sudsee* (1961; Eng. trans., *Oceania and Australia*, 1962); STEPHEN CHAUVET, *Les Arts indigènes en nouvelle Guinée* (1930) and *L'Île de Pâques et ses mystères* (1935); B.A.L. CRANSTONE, *Melanesia: A Short Ethnography* (1961); DANIEL SUTHERLAND DAVIDSON, *A Preliminary Consideration of Aboriginal Australian Decorative Art* (1937); ADOLPHUS PETER ELKIN and CATHERINE and RONALD BERHDT, *Art in Arnhem Land* (1950); JEAN GUIART, *Océanie* (1963); KAREL KUPKA, *Un Art à l'état brut: peintures et sculptures des aborigènes d'Australie* (1962; Eng. trans., *Dawn of Art: Painting and Sculpture of Australian Aborigines*, 1965); MAURICE LEENHARDT, *Arts de l'Océanie* (1947; Eng. trans., *Arts of the Oceanic People*, 1950); RALPH LINTON and PAUL S. WINGERT, *Arts of the South Seas* (1946); DOUGLAS NEWTON, *Art Styles of the Papuan Gulf* (1961); GLADYS AMANDA REICHARD, *Melanesian Design: A Study of Style in Wood and Tortoise Shell Carving*, 2 vol. (1933); ADRIAN A. GERBRANDS, *Wow-ipits: Eight Asmat Woodcarvers of New Guinea* (1967); T. P. VAN BAAREN, *Korwars and Korwar Style: Art and Ancestor Worship in North West New Guinea* (1968).

(J.Gt./Di.C./A.Ka.)

Oceanian Peoples and Cultures

The peoples and cultures of Oceania are spread from the Hawaiian Islands in the north to New Zealand in the south and from New Guinea and Australia in the west to Easter Island in the east. Most of the islands of Oceania lie in the central portion of the Pacific Ocean, bounded on the north by the Tropic of Cancer and on the south by the Tropic of Capricorn. There is no accurate count of the number of islands in Oceania; tabulations vary because the definition of what constitutes an island is not consistent. Estimates range from 7,500 to 10,000; many of these islands have never been inhabited by man.

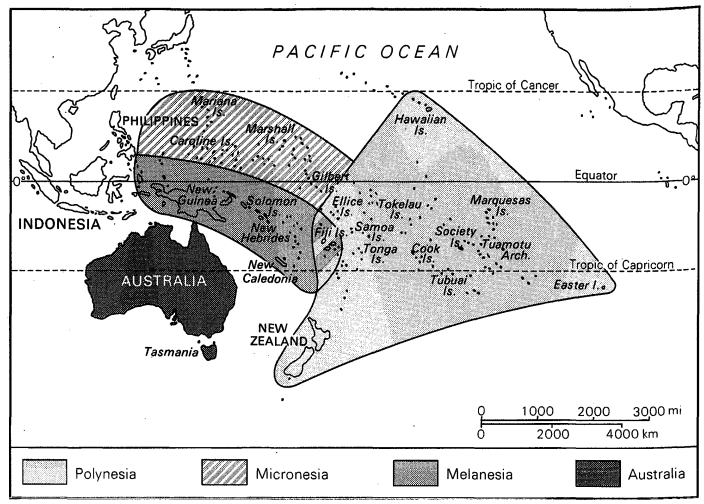
Major subdivisions of Oceania

Oceania is often subdivided into four main divisions or culture areas: Polynesia, Micronesia, Melanesia, and Australia. This classification system developed out of early-19th-century usage. Early European explorers often referred to the native Pacific islanders as Indians, grouping them conceptually with the inhabitants of the New World. Later observers noted that certain Oceanian peoples, although sometimes separated by vast expanses of water, were similar in appearance, language, and custom. An areal typology was devised that recognized both the commonly held characteristics of these peoples and the geographical locations of their island homes. The people of a large area, such as Polynesia, were seen to exhibit considerable variation, some groups even appearing atypical for the region in which they lived. Thus, geographical location alone was not a determining factor in the genetic heritage, language, and customs of its residents. The people living on an island in Melanesia, for example, were not necessarily Melanesian culturally, and Polynesian peoples were found residing on a few islands in Micronesia and Melanesia. These islands later came to be referred to as Polynesian "outlyers." The terms Polynesia, Melanesia, Micronesia were formulated by Europeans using the Greek words *polys* ("many"), *melanos* ("black pigment"), *mikros* ("small"), and *nēsos* ("island"). The term Australia was taken from *terra australis incognita*, the "unknown southern land," a long-sought hypothetical region whose existence was first postulated by early Greeks and Romans.

The islands of Polynesia are contained within a great triangle with apexes at New Zealand, Easter Island, and the Hawaiian Islands. Included are such groups as the Marquesas, Cook, Tubuai, Society, Tuamotu, and Ellice Islands, as well as the Samoa and Tonga Islands. The Fijian islands are situated on the imaginary line separating Polynesia and Melanesia, and Fijians reflect their medial position, sharing physical features and cultural and linguistic ties with both regions.

Melanesian islands include New Caledonia, the New Hebrides, the Solomon Islands, New Britain, New Ireland, and other small islands of the Bismarck Archipelago, stretching in an arc to the Admiralty Islands and including New Guinea—after Greenland, the largest island in the world.

Micronesia includes the widely dispersed islands of the western and eastern Carolines (principally the Palau Islands, Yap, Ponape, Truk, and Kusaie), the Marianas, the Marshalls, and the Gilbert Islands.



The culture areas of Oceania.

THE PEOPLE

Archaeologists have undertaken more research in Oceania since 1960 than ever before and are providing new conclusions about migration and settlement patterns. Yet there is an irregularity in the distribution of field research, and many important key sites are yet to be worked. The area of Oceania first penetrated by human immigrants was limited to New Guinea (and a few adjacent islands), Australia, and Tasmania. The earliest known occupation dates are all from Australian sites and are late Pleistocene: several excavations have produced materials that date from more than 30,000 years ago, and sites 40,000 years old are expected to be found. Some prehistorians believe that man reached Australia via New Guinea as early as 50,000 years ago. It is even possible that certain hominid forms in Australia were coeval with these early in-migrating *Homo sapiens* groups.

Earliest occupation dates

Racial groups. The reconstruction of the racial history of Oceanian peoples has been productive of many theories and explanations. The opportunities for breeding groups to develop in isolation and evolve adaptive responses to special conditions were numerous, and just as important to the racial diversity of the area is the hybridization produced by centuries of ebb and flow of migrating peoples. Such diversity is certainly the combined result of isolation or its lack, adaptation to differing environments, mating patterns, and natural selection.

Authorities still disagree over whether early hominid forms known from fossil finds to have inhabited Indonesian, Melanesian, and Australian areas were progenitors of peoples found there in recent time. There is similarly a great diversity of opinion on just which groups may have been antecedent—the Australoid, Veddoid, or Negritoid peoples found in the insular projection of lands south of Asia. Some specialists once proposed commonality of origin or at least some link between the Negroid peoples

of Africa and the darkly pigmented, sometimes frizzly haired groups found in Melanesia. Recent blood group research studies, however, show no evidence for such claims. In general, notions of a series of massive migrations, of relatively "pure" racial types from the Asian continent, and their miscegenation with antecedent peoples, seem to most specialists today to be simplistic. A quarter century of blood-group genetical studies produced no definitive conclusions about Pacific migration theory and little definitive information about the racial composition of Oceanian peoples.

Main racial elements

The racial composition of the peoples of Oceania is a composite of Caucasoid, Mongoloid, and Negroid elements; the Negroid peoples have survived or been reported in remote areas of Melanesia, the Philippines, Australia, and Tasmania. The Australian Aborigines may be described as a subcontinental isolate. Generalizations about the normative racial type of a particular culture area may be misleading. There are general tendencies, of course, but there are also such ranges of characteristics within virtually all areas that a representative of one area could easily be assimilated in one or more other areas.

Hair form among certain Melanesians is kinky, and many Polynesians have wavy to straight hair; but kinky hair is also to be found in Polynesia, as is curly to straight hair in Melanesia. The epicanthic eye fold, which distinguishes Mongoloids, is found in abundance in Micronesia, but it occurs elsewhere as well. Recognizing the difficulties of generalization, then, it may still be noted that Tongans, Fijians, Samoans, and Hawaiians are often tall and inclined to corpulence. Pigmentation generally tends to be darker among Melanesian groups than is normal in Polynesia or Micronesia. In some of the Melanesian islands, such as the Solomons and New Britain, colour is virtually blue black in some individuals. The ruggedness of Australian Aborigine facial characters exceeds that of other regions of Oceania. The people of the New Guinea highlands are of relatively short stature and have facial features of distinctive quality.

Migration sequences. The influxes of population into the area of New Guinea, Australia, and Tasmania continued until at least the end of the Pleistocene (10,000 years ago). Migration to Micronesia and Polynesia, however, appears not to have begun until 3,000 or 4,000 years ago: the islands of the Pacific were thus the last habitable portions of the earth to be populated.

Chronological sequences based on pottery styles, types of fishhooks, and other dating techniques suggest that man's presence in Micronesia was earlier than the oldest radioactive-carbon date yet found there of approximately 3,500 years ago. The evidence points to the entry of people into central Polynesia from eastern Melanesia. In due course, migrations radiating from both western and eastern central Polynesian island groups are thought to have settled the more distant reaches of Polynesia. Current opinion is that Hawaii was first reached midway to late in the 1st millennium AD by voyagers from the Marquesas, who were followed several centuries later by those from the Societies. The Societies also appear to have been the point of origin for the Polynesians who settled in New Zealand prior to AD 1000. Although the validity of the radioactive-carbon dates that have been recorded for Easter Island, indicating human presence by 400, has been questioned, it is in accord with dates of 100 to 300 in the Marquesas.

In Pleistocene times, when man first made his way out of Asia, the polar regions of the earth are believed to have held so much ice that the surface of the seas was several hundred feet lower than is the case today. Many areas that became submerged with the melting of the ice caps were then dry land and rendered migration easier.

The watercraft used by the ancient migrants from Asia were probably crude rafts or coracles: it is likely that the highly sophisticated sailing vessels so distinctive of the cultures of Oceania developed over a long period. However simple the early craft, those that men used thousands of years later to explore and settle the far reaches of Micronesia and Polynesia were impressive indeed. These were great double-hulled craft, capable of carrying a hundred or more voyagers and their provisions for a voyage of a month or longer, as well as domesticated animals and plants for propagation in a new homeland.

The issue of deliberate versus accidental voyages of discovery in the Pacific has prompted an impressive array of arguments from specialists on both sides. There is considerable evidence that some islands were, indeed, populated by seafarers who, because of a storm or mere random sailing, came unexpectedly upon their new home. There is also a great deal of evidence, however, of the navigational skills and abilities of Oceanian peoples and of certain deliberate voyages of discovery. It is also clear that return voyages were made to points of origin.

Language groups. The study of Oceanian languages sometimes requires generalizations that are no more than statements of probabilities. Many Oceanian languages have yet to be studied, and some are beyond study because they are extinct. Nevertheless, linguists offer the following characterization.

Two major language classifications

The languages of Polynesia, Micronesia, Melanesia, and Indonesia may be divided into two major groups: the Austronesian (sometimes called Malayo-Polynesian) and the Non-Austronesian (also called Papuan). Virtually all of the latter group are found on New Guinea and islands to the east and west of it. It is estimated that there are at least 500 Non-Austronesian languages spoken today. The number of Austronesian languages probably exceeds 500. Their geographical distribution extends from Easter Island to Madagascar and from Hawaii and Taiwan to New Zealand, and they are spoken in New Guinea, except for the south coast area. There are even some Austronesian languages found in southeast Asian mainland communities.

Most of the more than 200 languages of Australia, although in some instances quite different from each other, are thought to be related to one another—more than 85 percent being closely related. Common structural features strongly suggest a single original linguistic stock. Attempts to trace the Australian languages to those of other areas, even to contiguous ones, have proved fruitless; this probably indicates the great antiquity of the languages of Australia and the long isolation of the Aboriginal population (see also AUSTRONESIAN LANGUAGES, PAPUAN LANGUAGES; and AUSTRALIAN ABORIGINAL LANGUAGES).

Population trends. The nature and number of the various Oceanian peoples at the time of their discovery by Europeans are uncertain. The period of discovery was spread over several hundred years, beginning in the 16th century and becoming intense by the 18th and 19th. It was 1788, for example, before colonists from Europe settled in Australia. The number of Aboriginals at that time is not known, and their widely dispersed and seminomadic hunting and gathering groups are estimated to have been organized into from several hundred to nearly a thousand tribes. Tribal groups were generally associated with a given territory in which they eked out an existence from a relatively inhospitable land. Theirs was a simple yet highly functional material culture, which contrasted with a complex social system and a rich ceremonial life, featuring ancestor worship and totemic beliefs. They had rudimentary, impermanent shelters, went unclothed but for personal adornment, and foraged, hunted, and fought with simple but efficient implements and weapons. Some have estimated an Australian Aborigine population of 300,000 at time of European contact, but this is at best a guess. It is thought that the population had over time stabilized in relation to the ecology of the subcontinent and that, as a food-gathering and hunting culture, it was in balance with the environment. Like many other Oceanian peoples, the Australian Aborigines suffered severe population loss as a result of contact with the Western world. Only recently has the population of Aborigines begun to expand significantly. There are about 50,000 Australian Aborigines today, as well as a similar number of individuals who are part Aboriginal.

The island of New Guinea dominates the rest of Mela-

Populations of the major culture areas

nesia; 1,300 miles (2,100 kilometres) in length and more than 300,000 square miles (800,000 square kilometres) in extent, it has a mountainous spine rising to cloud-shrouded heights of 15,000 feet (5,000 metres). It was named in the 16th century by European explorers who were reminded by its dark-skinned inhabitants of those they knew from the Guinea Coast of Africa. Among the least altered portions of Oceania, it dominates Melanesia and Oceania not only in geographical size but also in population: estimates range from 2,500,000 to 3,000,000. The native population of the rest of the Melanesian islands, together with that of Micronesia and Polynesia, is a third to a half of that of New Guinea.

The native population of Micronesia (Guam, Trust Territory of the Pacific Islands, and the Gilbert Islands) is close to 200,000 or 225,000; that for Polynesia 1,000,000 to 1,200,000; and for Melanesia (including all of New Guinea) 2,700,000 to 3,200,000. Australian Aborigines number from 50,000 to 100,000. Therefore, all of Oceania probably contains somewhere between 3,750,000 and 4,725,000 indigenes. It must be pointed out, however, that reliable census data (or data for comparable years) is far from adequate.

CULTURE PATTERNS

The people and cultures of present-day Oceania differ greatly from those of the days when the areal classification into Polynesia, Melanesia, Micronesia, and Australia was devised. The Tasmanians have been extinct since 1876. Others, such as those living in the Mariana Islands and on Easter Island, often are more Spanish and Chilean, respectively, than they are Micronesian or Polynesian. With few exceptions, such as in remote areas of New Guinea, the peoples of Oceania have been exposed intensively to Western civilization and have adapted their ways to it either by choice or by necessity. Only certain features of their traditional cultures remain. There are, in greatly varying numbers, native inhabitants of all regions who manifest the physical types commonly associated with their culture areas. In some places, however, the native population is in the minority. It is, therefore, no longer accurate to say that the people of the Hawaiian Islands or of New Zealand possess a Polynesian culture or that those of New Caledonia are Melanesian culturally. Nevertheless, the fourfold typology of Oceanian peoples and their cultures continues to be used with reference to the contemporary inhabitants as well as to the traditional people of Oceania and their ways of life.

Stratification. Highly stratified social structures are typical of the aboriginal cultures of Polynesia and Micronesia. Hereditary chieftainship was closely related to supernatural beliefs that incorporated concepts of power (*mana*) and avoidance (*tapu*). Features of these two culture areas include an elaborate mythology; specialist craftsmen; distinctive artistic styles produced in part by isolation; pandemic and sometimes savage warfare; strong bonds of kinship and a related emphasis on genealogies; and, in places, strong trading or tributary relationships between island communities.

Melanesian societies are, because of their great diversity, less easily characterized. They were in general less concerned with social rank based on birth than with prestige gained through manipulation of resources. Root-crop agriculture (principally yams, sweet potatoes, and taro) was practiced as in many parts of Micronesia and Polynesia, and the resources of the sea were widely exploited. Ancestor worship and a rich ceremonial life were combined to produce a wealth of religious practices; they marked life crisis situations and incorporated both venerative and propitiative behaviour. In many areas complicated and highly ritualistic trading relationships were developed, and there were usually distinct differences between coastal and inland societies. Personal adornment was often elaborate; warring raids, in which heads were taken and cannibalism was sometimes practiced, were common; stone tools were polished; and animism was featured in supernatural beliefs. Its domesticated animals were those found elsewhere in Oceania, but pigs were vastly more important in the pursuit of personal prestige.

Technology. Most of the people of Oceania, prior to contact with the Western world, were at the Neolithic stage of technological development; that is, they worked stone by grinding rather than by chipping and flaking as in the Paleolithic stage. Some areas also had various other cultural features of the Neolithic such as agriculture, a few domesticated animals, pottery, and watercraft.

Utilization of materials

Seashells were used generally both as ornaments and as cutting tools. Shark, snake, and lizard skins were used as heads for drums. Bird, whale, and human bone, as well as fish vertebrae, were used for ornamental purposes and for tools, implements and weapons. Basalt, obsidian, jadite, chalk, coral stone, and other lithic materials are irregularly distributed throughout Oceania. Where these materials were available craftsmen produced a bounty of objects—adz blades, ornaments, images, and weapons —by skillfully chipping and grinding. Native woods of both hard and soft varieties were used to produce finely executed weapons, drums, masks, images, and building ornamentation. Great logs of pine were felled in New Zealand to produce flotillas of war vessels that could carry more than 100 men. Other islanders, without suitable timber for such canoes, scoured the beaches for driftwood logs or devised canoes from smaller trees, fastening planks together with cordage and caulking the seams with gum. Polynesian ironwood provided a source for some of the most intricately carved and heaviest of weapons. Ebony, bamboo, tree fern, and palm all added to the lumber supply of the Oceanian carver. Feathers of colourful birds were woven into garments and effigies in some places. Headdresses of feathers appeared in many parts of Oceania, but only the Hawaiians devised beautiful feather helmets. The kiwi, the tropic bird, the parrot, and the cassowary were but a few of the birds that added their plumage to the artisan's kit. Fibres and barks were used abundantly, particularly Orchid fibre, hibiscus, banana, flax, pandanus and ti leaves, and coconut-husk fibre. Pigments such as turmeric were obtained from plants. Earth and burned coral provided others. Among the countless other items that were utilized by the Oceanian craftsman were resins, gums, seaweed, spider webs, flying-fox fur, human hair, dog skin, kangaroo fur, boar tusks, vegetable seeds, shells, and teeth from humans, dogs, fish, and fruit bats.

The environment in which the traditional Oceanian native lived provided the setting for his culture. If there was little suitable land for agriculture, he had only tree crops such as coconut or pandanus; yam growing could thus not become a factor in social prestige nor tapa making a fine art. Although environment did not absolutely determine the nature of the culture that evolved in a given location, it served as a limiting factor; the technology of a people was in direct consonance with the surrounding region. The peoples who lived on continental islands (that is, on islands geomorphologically connected with a continent) had at their disposal a variety of minerals, a diversified flora and fauna, and, usually, a good freshwater supply. By contrast, the resident of a small coral atoll was preoccupied with securing an adequate supply of potable water and had fewer species of plants and animals. The type of island, its age, composition, and elevation, as well as its location and climate were vital factors in the development of a particular culture. Oceanian peoples had to adapt to a broad range of extremely diversified environments—tiny Micronesian atoll, humid New Guinea forest, or arid Australian desert. An artisan used the materials at hand for the most part. At times primitive trading systems allowed him to secure some materials he did not already possess, but he was mainly dependent upon his surroundings.

MODERN DEVELOPMENTS

Oceanian cultures are today blended of many elements. The number and strength of residual Aboriginal cultural traits vary greatly with the particular history of contact with non-Oceanic cultures. Much of the Polynesian cultural heritage of present-day Society Islanders has been supplanted by French influences. At the other extreme, certain remote natives of the New Guinea highlands,

whose exposure to alien influences has been negligible and whose very existence was unsuspected until a few decades ago, live much as their stone-age ancestors did for thousands of years. Even there, however, modern society is penetrating. Everywhere the time-honoured usages succumb to newly introduced supernatural beliefs, to contemporary social, economic, and political practices, and to modern technology.

The world of Oceania has been a world in flux since the great voyages of discovery by Europeans began in the 16th century to expose the peoples to new customs, values, ideas, diseases, and genes. The particular blend found in any corner of Oceania today is a unique amalgamation of surviving ancient cultural roots and modern influences.

Variety of foreign cultural imprints

Thus, an Oceanic community that has been primarily in contact with representatives of French culture (the Society or Marquesas Islands or New Caledonia, for example) differs greatly from that of a community largely exposed to British cultural influences (Gilbert or Solomon Islands, New Zealand, or the Cook Islands). The educated Fijian civil servant speaks in cultured Oxford accents. The Papuan presiding officer of the House of Assembly in Port Moresby is attired in black robe and white wig. The American Samoan policeman is uniformed in a manner like his counterpart in the United States, but the gendarme in Papeete is attired in the French tradition. Micronesians schooled under the Japanese mandate of their islands between World Wars I and II often speak and read Japanese, have a preference for Japanese cuisine, and may name a child Yoshio or Hiroshi. Spanish influence in the Mariana Islands has made Catholicism strong, whereas Protestant missionaries brought Calvinist beliefs and traditions to the Samoa and Tonga Islands of Polynesia.

Modern Oceania has many of the same concerns, problems, and responsibilities that face other areas of the world. There is concern over the environment, population control, urbanization and its attendant ills, and the need to develop viable economies. People must sometimes cope with severe problems in transportation, communications, food production, public health, and education. Their responsibilities involve the training of indigenous persons for specialized professional, technical, and administrative positions once held by representatives of governing powers or foreign entrepreneurs. They seek more self-government and political autonomy, higher levels of general education, and economic self-sufficiency.

Problems of health and welfare

Some commonly shared achievements that transcend cultural and national boundaries in the Pacific relate to public health and human welfare. Generally, people are healthier and live longer than formerly because of improved health services and preventive and remedial treatment. Controls have been effective—sometimes dramatically so—in the treatment of yaws, tuberculosis, malaria, filariasis, and leprosy. Infant mortality is greatly reduced, and immunization programs as well as village sanitation procedures have had a positive effect upon most island communities. Despite these gains, physical facilities often tend to be substandard, and trained medical practitioners are in short supply. The very success of health programs and the reversal of depopulation trends that began with epidemics started by early European voyagers, kidnapping of natives for slaves, and warfare with firearms have brought about one of the most severe problems of modern times. In most island communities, population doubled in the two decades following World War II. Today, great numbers of people cluster in port cities and towns, seeking employment, captivated by the colour and novelty of 20th-century technology. There are slums in Papeete and in Port Moresby. Sometimes, in a quest for the accoutrements of latter-day society, a people emigrates to a metropolitan centre en masse. The people of the Tokelau Islands of Polynesia, for example, have virtually transplanted themselves to New Zealand.

Human crowding occurs elsewhere, as in Fiji, from the immigration of "outsiders." The Indian population of Fiji, which settled in the islands first as labourers on sugar plantations, today dominates much of the commercial activity in Fiji and seems destined to vastly outnumber the indigenous population. Settlers from Europe clearly are predominant in New Zealand and Australia, and native Hawaiians represent a small proportion of the population of Hawaii. French business interests control the mining industry in New Caledonia, to which Europeans and islanders from Wallis and Futuna and from French Polynesia have been attracted in great numbers. Urban crowding is attended by the social ills found in cities elsewhere in the world—delinquency, crime, poverty, fragmented family life, mental illness, and ennui.

Social change brings mixed blessings. Literacy rates are higher today than ever before, but industrial development pollutes the environment dramatically in places such as Nouméa. The building of a jet airport can cause tourism to quadruple in two or three years. Tourism produces income, but many island leaders are concerned with the impact of such repetitive invasions upon their societies. More secondary school graduates are produced each year, yet there are too few opportunities for employing educated individuals when they return to their villages.

Economic issues and political aspirations

The economies of most island communities are highly dependent on the activities of external administering authorities, whose stewardship most islanders are striving to escape. The key to independence is undeniably in a greater degree of economic self-sufficiency than many Oceanian societies can manage. Most seek the development of fisheries and agricultural projects. They strive to improve livestock, crop, and forestry production. Mineral resources also are looked to for income. Markets for copra, sugar, cocoa, coffee, and bananas tend to vary, and there are some severe pests, such as rats and the coconut rhinoceros beetle, as well as devastating typhoons, which make crop production undependable. Nauru, which is rich in phosphate, has mined this resource so assiduously that its people will soon face a soil shortage. Similarly, mining exploitation in the Solomon Islands and lumbering and petroleum operations in New Guinea threaten a profound disruption of natural resources.

In attempting to cope with the problems of the modern world, the Oceanian peoples have worked with international organizations such as the Food and Agriculture Organization, the World Health Organization, the United Nations Development Programme, the United Nations Educational, Scientific, and Cultural Organization, and the South Pacific Commission. The nations that for so long dominated the Pacific world are withdrawing and are imposing their controls and initiatives less and less, and the Oceanian cultures are returning to autonomous self-government. The difference is that they cannot now live in a world of relatively splendid isolation as before. They face an earth grown smaller through technological change, and in which new dimensions have been added to old problems.

BIBLIOGRAPHY. F.R. FOSBERG (ed.), *Man's Place in the Island Ecosystem* (1963), is the most valuable source of its kind to date. SHERWIN CARLQUIST, *Island Life* (1965), treats the natural history of islands in great detail. The most complete compilation of information on the geographical exploration of the Pacific is provided by HERMAN R. FRIIS (ed.), *The Pacific Basin* (1967). Two recent readers offer a comprehensive sample of cultural diversity in Oceania: ANDREW P. VAYDA (ed.), *Peoples and Cultures of the Pacific* (1968); and THOMAS G. HARDING and BEN J. WALLACE (eds.), *Cultures of the Pacific* (1970). A number of important original contributions on Pacific migration theory, archaeology, linguistics, and ethnography are found in GENEVIEVE A. HIGHLAND *et al.* (eds.), *Polynesian Culture History* (1967). Australia's past is expertly covered by D.J. MULVANEY, *The Prehistory of Australia* (1969). Summaries of recent research in linguistics and archaeology are included in R.C. GREEN and M. KELLY (eds.), *Studies in Oceanic History*, vol. 1 and 2 (1970–71). No work on the coming of Europeans to the Pacific surpasses that of J.C. BEAGLEHOLE, *The Exploration of the Pacific*, 3rd ed. (1966). Two exceptionally well-illustrated books on Oceanic art are JEAN GUIART, *The Arts of the South Pacific* (1963); and ROLAND W. and MARYANNE FORCE, *The Fuller Collection of Pacific Artifacts* (1971). For a series of treatments of contemporary Oceania, see ALEXANDER SPOEHR (ed.), *Pacific Port Towns and Cities* (1963); and ROLAND W. FORCE (ed.), *Induced Political Change in the Pacific* (1965). Helpful statistical and

other summaries may be found in the *Pacific Islands Year Book and Who's Who* and *The Far East and Australia* (both issued annually).

(R.W.Fo.)

Oceanic Ridges

Nothing was known of the depth or shape of Earth's ocean floors until 1840, when Sir James Ross of the British Royal Navy made the first deep soundings. By 1855 Commo. M.F. Maury had discovered a broad zone of shoal water in the North Atlantic, indicating the presence of an oceanic ridge, and evidence slowly accumulated thereafter that confirms the existence of other submerged ridges in all the oceans. But soundings made by lowering a weight on a rope or wire took so long that only a few were made, and the ocean floor was considered to be generally smooth. After 1920, however, the use of echo sounding—in which sound signals are beamed toward the ocean bottom and the time each takes to return is recorded—by British, Danish, German, and United States ships demonstrated that a great ridge extends along the axis of the North and South Atlantic oceans. They showed that another lies in the centre of the Indian Ocean, and that some parts of the Eastern Pacific Ocean were shallower than others.

Since 1954, refined techniques have permitted the recording of continuous profiles of the ocean floor beneath a ship's track, with an error of only one or two metres. Large uncertainties in the precise location of the ships while making the surveys remained a problem and were the chief source of error until after 1960, when methods of satellite navigation were introduced. In recent years the number of expeditions has increased to such an extent that no large parts of the ocean remain unexplored, and the pattern of major ridges has been established, although many details have yet to be deciphered. It is recognized today that oceanic ridges comprise one of the major features on the Earth's surface, rivalling the subaerial mountain systems in magnitude.

This article treats the classification, properties, origin, occurrence, and distribution of the oceanic ridges. For further information on the landforms and general topography beneath the oceans see OCEAN BASINS; CONTINENTAL SHELF AND SLOPE; and EARTH, PHYSIOGRAPHY OF. Additional treatment of the mechanisms that produce oceanic ridges and the significance of both the mechanisms and the ridges in Earth history is provided in the articles SEA-FLOOR SPREADING; CONTINENTAL DRIFT; ROCK MAGNETISM; ISLAND ARCS; EARTH, STRUCTURE AND COMPOSITION OF; and MOUNTAIN-BUILDING PROCESSES.

CLASSIFICATION OF RIDGES

The charting of the ocean floors has shown that they are not smooth and that many long and prominent ridges divide them into separate basins. These ridges can be classified into a few types on the basis of their properties and their relationships to one another and to the adjacent continents. The most important question is whether or not a ridge is active; that is, whether earthquakes occur along its axis. This distinction, combined with bathymetry, or the delineation of submarine topography, suffices to provide a classification that is confirmed by other geophysical and geological properties. Only the midocean ridge system and the island arcs of the world are active; all others are aseismic or quiescent.

The midocean ridge system. In 1956 the U.S. scientists Maurice W. Ewing and Bruce C. Heezen proposed that the several great sections of the known active ridges might form a continuous system. In support of this proposition it was pointed out that earthquake focuses in the southern Atlantic, Pacific, and Indian oceans join together and might mark the course of connections that had not been detected because of the paucity of soundings in Antarctic waters.

Continuity and dimensions

Expeditions sent for the purpose soon confirmed that the midocean ridge is indeed continuous and that it winds for 60,000 kilometres (40,000 miles) through all the world's oceans. It extends down the axis of the entire Atlantic Ocean, passes midway between Africa and Antarctica, and turns north to the centre of the Indian Ocean, where it branches, the main ridge continuing midway between Australia, New Zealand, and Antarctica to cross the east side of the Pacific Basin, running all the way to the mouth of the Gulf of California.

Because the system is so large and the behaviour and properties of different sections vary, several sections have separate names, such as the Mid-Atlantic Ridge, the Carlsberg Ridge, or the East Pacific Rise (Albatross Cordillera). Nevertheless, the system has an essential unity and is called the midocean ridge because of its generally axial position. Seismically active branches of the ridge system join the Azores to Gibraltar, the centre of the Indian Ocean to the Gulf of Aden, and other points on the crest to New Zealand and to Chile. A few short sections occur independently, for example, off the coast of Oregon.

The essential characteristic of the midocean ridge is that it is a broad and enormous uplift along the axis of the ocean floors. The floors slope up from depths of over 6,000 metres (20,000 feet) in marginal basins to the crest of the ridge, which is almost everywhere less than 4,000 metres (13,000 feet) in elevation, with crestal mountains commonly rising to depths of 2,000 to 3,000 metres (7,000 to 10,000 feet) and sometimes reaching above sea level.

Profiles of the East Pacific Rise show that the central portion is fairly smooth. In contrast, in the North Atlantic Ocean the crest is marked by a rift valley averaging 30 kilometres (20 miles) in width and up to 1,500 metres (5,000 feet) deep, and the inner flanks are rough and block faulted. The crest is, of course, the highest part of the ridge, but it does not lie at a uniform depth; rather, it rises towards a number of well-separated peaks or domes. In the Atlantic the crest rises above the sea to form seven groups of islands, of which Iceland and the Azores are conspicuous examples.

At the crest the basalt floor is everywhere exposed, but sediments increase steadily in thickness down the flanks and fill depressions until only abyssal (deep-sea) hills project, and finally all the bedrock is hidden beneath flat abyssal plains. But the essential structure of the ridge continues beneath the sediments to continental margins or to boundary ridges.

Fractures and earthquakes

At intervals the ridge is crossed, usually at right angles, by large topographic disturbances called fracture zones that offset, or displace, the crest. These offsets are called transform faults. Major ones, many of the largest of which cross the ridge close to islands on the crest, show offsets of as much as several hundred kilometres, but small ones cross the Mid-Atlantic Ridges every ten to 30 kilometres (five to 20 miles).

Shallow earthquake focuses follow the crest, and it has been demonstrated that those along the axis of the ridge are caused by normal faulting (one block moves downward, relative to the other, along a steeply inclined fault plane), while those on fracture zones are due to shearing (involving horizontal movements). Further, all the earthquakes have directions of motion compatible with tension and the spreading of the crest.

The midocean ridge system is the largest feature of the Earth's surface after continents and ocean basins. Its volume exceeds 100,000,000 cubic kilometres (25,000,000 cubic miles), and, if it did not exist, sea level would occur at more than 250 metres (800 feet) lower than it does.

Other ridges. *Lateral ridges.* Aseismic ridges that extend wholly or partly from uplifted islands on the axis of the midocean ridge to coasts of adjacent continents are called lateral ridges. Usually they occur in pairs and are perpendicular ("normal") to or make large angles with the midocean ridge. For example, broad ridges whose crests are everywhere less than 2,000 metres (6,500 feet) deep extend in opposite directions from Iceland to the coast of Greenland and to the continental shelf of Europe near the Faeroes and Scotland. Also, the Walvis Ridge and Rio Grande Rise make a nearly complete chevron-shaped pattern about the Tristan da Cunha group, in the South Atlantic Ocean.

Linear chains. Nearly straight ridges that terminate at one end in a young or active volcanic island are called

linear chains. The other islands of the chain appear to get steadily older away from the volcanic end. Except for local earthquakes connected with volcanism, these chains are aseismic. The Hawaiian Islands are an excellent example.

Boundary ridges. In the Atlantic Ocean the midocean ridge system lies near the centre of the ocean, except in the Arctic Ocean, where its continuation is much closer to the Siberian coast than to North America. On the other hand, the ridge does occupy a symmetrical position halfway between the Siberia coast and the Lomonosov Ridge. The latter is a great submarine mountain system that cuts right across the Arctic Ocean from north of Greenland and Ellesmere Island to the delta of the Lena River. It is generally supposed that this aseismic ridge marks the boundary of the present Atlantic system with an older section of sea floor north of Alaska.

Microcontinents. The Seychelles archipelago in the Indian Ocean resembles a continent in everything but size. The rocks of these islands are Precambrian (older than 570,000,000 years) gneisses (metamorphic rocks, formed under high temperatures and pressures); the crust beneath them is thick and continental, and the flora and fauna are of continental rather than insular types. There is complete agreement that this archipelago constitutes a microcontinent. There is also good evidence from deep-sea drilling and geophysical investigations that some submarine rises, notably the Rockall Bank, off the British Isles, are also submerged microcontinents, and many other banks and islands, including the Lomonosov and Walvis ridges and Kerguelen Island, have also been considered to be continental, but the evidence in many cases is still uncertain.

ORIGIN AND GROWTH OF RIDGES

Gravity, seismic, and magnetic data

General geophysical properties. Gravity measurements show that the midocean ridge is in isostatic equilibrium —*i.e.*, its height above the seabed must be held up by lighter material below. Seismic refraction studies show that the upper mantle (zone immediately beneath the Earth's crust), to depths of 100 or more kilometres beneath the crest, has low seismic velocities (speed of travel of earthquake waves) of about 7.6 kilometres (4.7 miles) per second, compared with normal values of about 8.2 kilometres (5.1 miles) per second. Velocities measured perpendicular to the rise are as much as 8 percent higher than those measured parallel to it. Other measurements show that the heat flow along the crest is higher than normal. To explain these observations it has been proposed that hot matter from the deeper mantle is upwelling toward the surface (the ocean floor) along the axis of the ridge.

Another remarkable geophysical feature of the ridge is that everywhere along it a pattern of magnetic anomalies occurs as long strips, each a few kilometres wide and parallel to the axis. This can be explained if the lava in successive strips on the sea floor is alternately normally and reversely magnetized. It has been proposed that along the locus of the crest of the midocean ridge the ocean floor has been spreading apart and that the natural reversals that occur in the Earth's magnetic field cause the pattern of alternating strips of magnetic anomalies. This proposal fits the assumption that hot material is welling up under the ridge.

The theory of plate tectonics. The pattern of ridges is so regular and their properties are so constant that these features cannot be caused by chance. The only theory advanced that explains their characteristics satisfactorily is that of plate tectonics. Because this hypothesis has been accepted almost unanimously by marine geologists and geophysicists, it will be used here as a framework upon which to base descriptions of the ridges.

According to the theory of plate tectonics, the outer 50 kilometres (30 miles) of the Earth constitutes a brittle shell, or lithosphere, that is broken into six large and perhaps a dozen smaller plates. Convection currents, or upwellings from the deep interior, cause these plates to move slowly about relative to one another and to the interior, and these motions generate the ridges. A typical large plate includes either one or two continents and the surrounding ocean floor.

The midocean ridge system and its branches mark boundaries along which upwelling is splitting the crest and creating new ocean floor along the trailing edges of both plates. This growth, which occurs equally on both sides, causes the plates to move apart. Particularly active upwellings at some places under the midocean ridge are marked by active volcanoes, which form such islands as Iceland.

Lateral ridges are thought to have formed where this localized activity has been of long duration. They have been formed by the outpouring of excess lava throughout the time of spreading of an ocean basin and thus mark the locus of movement of the two plates away from the hot spot.

Linear ridges are considered to have a similar origin, except that in their case the source of the lava does not lie on the axis of a spreading ridge. Instead, if one plate is driven over a hot spot from which lavas can penetrate to the surface, then those lavas generate a single chain of progressively older volcanoes, which mark the direction of past motion of the plate over that hot spot.

Boundary ridges are believed to have formed where an excessive upwelling of lavas occurred where an old ocean floor was split apart by the creation of a new section of midocean ridge crossing it.

Island arcs, young mountains and their accompanying trenches, mark subduction zones—zones where two lithospheric plates have been forced together so that one has overridden the other and has pushed it down to be reabsorbed into the mantle.

The concept is now widely accepted that some ocean basins, such as the Atlantic, grow by the spreading of the midocean ridges, whereas others, such as the Pacific, shrink by the overlap and absorption of one plate by another. Many authorities incline to the view that a system of convection currents rising under the midocean ridge forms the driving mechanism. Although the concept is clear enough, great difficulties have been encountered by those who have attempted, by either theory or experiment, to find precise models of rotating systems of currents.

Role of upwelling plumes

As a simpler alternative system of flow, U.S. geologist W. Jason Morgan proposed in 1971 that upwellings such as those under Iceland and Hawaii are the tops of plumes of hot material rising from deep within the mantle and that these are the currents that drive the plates. He noted that the plumes lift up the lithosphere above them in broad domes (as much as 1,000 kilometres [600 miles] across and two or three kilometres high) that are capped with volcanoes generated by the rising hot material. Where several domes lift the surface, the lithosphere may break between them to form plates that slide off the sides under the influence of gravity. The lavas that rise from the breaks in the lithosphere linking the domes come from much shallower depths than the lava in the plumes; this could explain the difference in composition between the axial islands and the rest of the ridge and the ocean floors generated by the ridge.

At the present time the plumes of hot matter rising under seven islands along the Mid-Atlantic Ridge appear to be spreading the floor of the Atlantic Ocean. Four other plumes, rising beneath islands on the crest of the East Pacific Rise, also appear to be actively driving that ridge apart. These two systems can both operate independently because the subduction zones around the rim of the Pacific separate them. Some other plumes, in particular those beneath islands in the Indian Ocean, are only partly successful. They have broken the lithosphere into plates and caused that ocean to open, but, because they are also being overridden, the sections of midocean ridge between them have to change position from time to time, which has complicated the pattern in that ocean. The active islands, including Kerguelen and Réunion, no longer lie on the crest.

Other plumes, such as those beneath Hawaii, have not yet joined with others to break the lithosphere and only leave trails of islands.

OCCURRENCE AND DISTRIBUTION

Ridges of the Atlantic Ocean. The Mid-Atlantic Ridge lies down the axis of the Atlantic Ocean. Topographically it occupies the central third of the basin between a series of flat abyssal plains, but structurally it continues beneath these to the continental coasts. Along the crest are seven groups of islands, of which Jan Mayen, Iceland, the Azores, Tristan da Cunha, and Bouvet are volcanoes that have been active in historical times. Ascension Island is also a recent volcano, but St. Peter and St. Paul Rocks are built of ultrabasic rocks (rich in iron and magnesium) believed to be uplifted mantle, although basalt lava has been dredged from the sea floor nearby. It is considered that rising plumes beneath these islands are widening the Atlantic Ocean by spreading the continents that border it apart.

Movements of the African and American plates

In the south the African and American plates are sliding off the ridge by rotation about a pole near Iceland. In the north the Eurasian and American plates are sliding off the ridge by rotation about a pole near the New Siberian Islands. An active ridge that links the Azores to Gibraltar marks the junction of the Eurasian and African plates. From the four axial islands in the South Atlantic the Sierra Leone Rise, the Guinea Rise, the Walvis Ridge, and the Cape Rise are lateral ridges parallel to one another that join these islands to the African coast. Although it has been suggested that the Walvis Ridge is a microcontinent or a block of upfaulted ocean floor, the Cape Rise has been mapped as a chain of volcanic seamounts, and many believe that all these lateral ridges are trails of former volcanoes left by hot spots. On the west side, with the exception of the well-marked Rio Grande Rise from Tristan, the lateral ridges are less clear, but there are some indications of them in islands and seamounts off Brazil. This herringbone, or chevron-shaped, pattern in the South Atlantic contrasts with the orthogonal (right-angle) pattern of the ridges that extend away from Iceland in directly opposite directions to either coast. This difference is believed to have arisen because the southern continents have been moving northward as well as east and west over the mantle since the breakup of Gondwana (a protocontinent), whereas the northern continents have not.

A remarkable feature of the equatorial Atlantic is the series of great fracture zones that offset the ridge to accommodate it to the bend marked by the Gulf of Guinea and the corresponding angle in the northeast coast of Brazil. These are all transform faults formed as the two plates were torn apart, and they do not extend into the continents of Africa and South America.

To the north the Atlantic structure gets progressively more complicated. This may be due to a shift in the position of the Mid-Atlantic Ridge, and it has been postulated that a predecessor of the present ridge lay up Baffin Bay and across the Arctic Basin, where the Alpha Ridge may be its inactive remnant. If so, the Lomonosov Ridge formed as a boundary ridge when the Mid-Atlantic Ridge jumped to the position in which it is now active. The exact nature of the Lomonosov Ridge has been debated, but it is generally believed that it is a sliver of continental shelf separated from the adjacent coast of Siberia, which has the same shape.

In the Norwegian Sea and Northern Atlantic off the British Isles are several other small, irregular ridges. Immediately south of Iceland the actively spreading Reykjanes Ridge is not in the centre of the ocean but halfway between Greenland and the Rockall Bank, which lies from 200 to 400 kilometres (125 to 250 miles) off the British coast and is separated from it by deep water. The nature of the volcanic rocks constituting the islet and bank of Rockall and the results of deep-sea drilling have confirmed that the bank is a sunken continental block, or microcontinent. It remains to be explained why it should have sunk. It is believed that until a few tens of millions of years ago the active ridge lay between Great Britain and Rockall Bank and separated them by spreading. This section of midocean ridge then jumped to its present location.

North of Iceland the active ridge again does not lie along the axis of the Norwegian Sea. It is much closer to Greenland and midway between that island and the submerged Voring Ridge, which is again separated from Norway by deep water. Voring Ridge is therefore regarded as another sunken microcontinent like Rockall Bank and Lomonosov Ridge. Between Norway and the Voring Ridge a minor ridge regarded as the former spreading crest has been identified. It is now inactive.

Ridge movements in the Arctic

This contrast between the regular pattern of steady spreading that has proceeded without interruption in the South Atlantic for over 80,000,000 years is in striking contrast with the behaviour in the Arctic regions, where the location of the midocean ridge has jumped to three locations in the past 60,000,000 years. Yet this difference is believed to be a natural consequence if the plates are driven by random, long-lived plumes rising at fixed locations in the mantle. If by some chance one group of plumes has become dominant—for example, those in the South Atlantic—then the growth and interaction of rigid lithosphere plates are such that other spreading ridges cannot also remain fixed but must change their positions from time to time.

In considering these rather complex interactions, the details of which have not yet been fully worked out, it is essential to distinguish between different kinds of poles of rotation. Any pair of plates have one pole about which they mutually rotate, but it is important to appreciate that the pole of rotation between two plates may itself be moving over the mantle. If so, there exist two other poles, about one of which each of the plates rotates relative to the mantle. To state that the spreading of the South Atlantic is a dominant motion means that all three of these poles between the African and American plates and the mantle are coincident and near Iceland.

The Eurasian plate is moving relative to both of these plates and therefore has different sets of poles with them. Thus, the mutual pole between the American and Eurasian plates is near the New Siberian islands, but, because the pole of the American and African plates is fixed in the mantle near Iceland, the pole of the American and Eurasian plates near the New Siberian Islands cannot also be fixed in the mantle. The Arctic section of the midocean ridge is therefore also moving over the mantle and hence across its driving plumes. When a section of ridge is displaced too far from its driving plumes it is liable to become abandoned, and a new section is created over the plumes again. This could explain the ridges of the Arctic.

Ridges of the Pacific Ocean. The dominant feature of the Pacific Ocean Basin is the East Pacific Rise, which is that part of the midocean ridge system that extends from south of New Zealand to the mouth of the Gulf of California. It is broad and much smoother than the Atlantic section. It also lacks a crestal rift. Several great fracture zones cross it, notably at 55° south latitude and near Easter Island.

Off the west coast of North America, north of Mexico, the presence of several inactive fracture zones and the pattern of magnetic anomalies and short isolated sections of midocean ridge seem to represent the western half of a continuation of the East Pacific Rise and suggest that it once extended farther north until it was overridden by North America.

Several great fracture zones, including the Mendocino, off California, cross this part of the ocean floor. Two active branches of the Rise join Easter Island to the coast of Southern Chile and cross the Panama Basin from near Galápagos Islands to the coast of Columbia.

Spreading of the East Pacific Rise

Any chart of the Pacific shows a striking difference between the southeastern half of the Pacific Basin, which is underlain by the East Pacific Rise, and the northwestern part, which contains many more islands. The topographical boundary of the rise has been placed close to latitude 150° west, except in the extreme south, but many consider that the whole Pacific Basin is progressively older from the rise to the trenches off Japan, and that it is all part of the same system. Others, noting the change in the abundance of islands, consider that a boundary ridge extends down the Pacific from Kamchatka to the Campbell Plateau. Chief elements are the Emperor Seamount Chain and the great ridge of the Line Islands.

It has been suggested that the driving force that is spreading the East Pacific Rise is the upwelling of a series of plumes. The tops of some of these form the Islas de Revillagigedo, off Mexico, the Galápagos Islands, Easter Island, and Macquarie Island, south of New Zealand. It has been suggested that several other plumes that do not lie on the crest have been overridden to form linear chains of islands.

The most conspicuous group of parallel linear chains form the Hawaiian, Marquesas, Tuamotu, Society, and Tubuai islands. Two other parallel lines of seamounts cross the Gulf of Alaska. Where evidence exists, as it does for the Hawaiian and Tubuai islands, it shows that the active and younger islands are to the southeast and the islands get progressively older to the northwest, which suggests that the Pacific plate is moving northwesterly, as confirmed by other evidence. Some authors have regarded the Emperor Seamount Chain, Line Islands, and Cook Islands not as a boundary ridge, but as older continuations of the Hawaiian, Tuamotu, and Tubuai islands. They consider that the angle between the two sets marks a change in direction of motion of the Pacific plate.

Another interesting observation is that the Nazca Ridge and another smaller, parallel ridge of seamounts off the coast of Peru and northern Chile are in a mirror-image position across the East Pacific Rise relative to the Tuamotu and Tubuai islands.

The full significance of these observations is not yet clear. There are not enough geophysical data from the vast Pacific Basin, and only a few drill holes have been obtained by oceanographic vessels. Observation is handicapped because the ridges are largely submarine, and most islands in the tropics are mantled by young coral. Further, only high volcanic islands, of which there are few, can be radiometrically dated.

Between Australia and New Zealand, several submarine ridges, including Lord Howe Rise and Norfolk Ridge, are generally regarded as sunken microcontinents because they are connected with New Zealand and New Caledonia and have continental geophysical properties. If this is so, the ocean floor has spread between them, and sections of the midocean ridge have jumped about, probably for the same reason as that given for the Arctic regions. It is not known why such continental ridges would have sunk, but the same phenomena have also been observed in the Rockall Bank and Agulhas Plateau, both of which have continental properties.

Ridges of the Indian Ocean. The Indian Ocean Basin is the most complex, and its ridges are the least understood. As elsewhere, the ridges may first be divided into active and aseismic varieties. The active midocean ridge system has here been divided into five segments, each with local names. Along the whole length of the Gulf of Aden an active median ridge called the Sheba Ridge can be traced.

Occurrence of fracture zones

It has a median rift and is much offset by fracture zones trending northeast–southwest that displace it by as much as 180 kilometres (110 miles). It shows all the characteristic features of high heat flow, magnetic lineations, low gravity, and a spreading motion indicated by recent earthquakes along the axis.

At the entrance to the gulf, the Owen Fracture Zone displaces the axis 300 kilometres (200 miles) to the southwest and marks the beginnings of the Carlsberg Ridge, which lies halfway between the west coast of India and the Seychelles portion of the Mascarene Plateau (Seychelles-Mauritius Plateau), which is presumably a boundary ridge. Spreading of the Carlsberg Ridge, which is still active, is assumed by authorities to be connected with the rapid northward movement of India during the last 65,000,000 years, but the details of this occurrence have yet to be established.

The Mid-Indian Ridge, which extends south for 25° of latitude, is connected to the Carlsberg Ridge at the Equator. It is broken and offset by many large fracture zones trending northeast–southwest. Recent work suggests that the Mid-Indian Ridge was opened by the spreading of an earlier fracture zone with a north–south strike and whose remnants are the Chagos-Laccadive Plateau, off the coast of India, and the central part of the Mascarene Plateau. The Rodriguez Seamount Chain projects east as a spur from the plateau and appears to be a later addition.

The southern end of the Mid-Indian Ridge is at a triple point and junction with the Southwest Indian Ridge and Southeast Indian Rise. The former passes south of Africa to join the Mid-Atlantic Ridge near Bouvet Island. It differs considerably from most other segments for reasons that are imperfectly understood but are generally attributed to a very slow spreading rate. This is indicated by lack of a well-defined magnetic pattern and the narrow crestal zone free of sediments. Most authorities also consider that it has been cut by many north–south fracture zones.

The Southeast Indian Rise is a normal one. It is offset 1,100 kilometres (700 miles) to the south by a fracture zone through Amsterdam Island and by others south of Australia and New Zealand, where it joins the East Pacific Rise. In its northwestern part its features are similar to those of the Mid-Atlantic Ridge, while in the portion south of Australia they are like those of the East Pacific Ocean in such aspects as a fast spreading rate, the absence of a median rift, and relatively smooth topography. The properties of such active ridges are thus compatible with the view that the principal motion at the present time is a rotation of the northeastern part of the Indian Ocean Basin away from the southwestern part about a pole in the vicinity of Arabia. It is also clear, however, that other motions of unknown nature operated in the past.

The Indian Ocean is also marked by a large number of aseismic ridges of a considerable variety. One striking group is elongated in a north–south direction, and it seems probable that at least some groups originated as large fracture zones or linear chains during the northward movement of India. Of these, the Chagos-Laccadive Plateau has already been mentioned. Three groups of atolls rise from a narrow ridge 2,700 kilometres (1,700 miles) long that apparently terminates to the north in the Deccan basalts near Bombay. The Ninety East Ridge is 4,800 kilometres (3,000 miles) long and about 200 kilometres (125 miles) wide. Its top lies at depths of 1,800 to 3,000 metres (5,900 to 10,000 feet), and its relief is up to 3,500 metres (11,500 feet). It is apparently composed of basalt and has also been held to be an upthrust horst, or a large block bounded by normal faults.

The Mascarene Plateau is a curved ridge 2,300 kilometres (1,400 miles) long. Its northern end, the Seychelles Bank, is of a composite nature and is a true microcontinent of Late Precambrian (older than 570,000,000 years) granite-gneiss. The southern end is the Tertiary volcanic island of Mauritius. In between lie a series of coral banks resting upon basalt foundations. Separated from Mauritius by a small gap is the active volcanic island of Réunion.

The Madagascar and Mozambique plateaus are other elongated ridges extending south from Madagascar and the coast of Africa. They are believed to be associated with fracture zones, but little is known about them. Off the southern tip of Africa the submarine Agulhas Plateau has been shown to be continental, whereas the Crozet Plateau is of unknown origin and is capped with the extinct basaltic volcanoes of the Îles Crozet.

In the southeastern part of the Indian Ocean Basin several ridges are arranged in a mirror-image fashion. These include the Broken Ridge and Naturaliste Plateau, extending west from Australia, and the Kerguelen Plateau, off Antarctica. Before the Southeast Indian Rise opened they would have been together and are thus considered to be boundary ridges.

Since 1956 the principal ridges of the Indian Ocean have been located and named, but they are so large and complex that they are still incompletely known.

BIBLIOGRAPHY. B.C. HEEZEN, *The Face of the Deep* (1971); and *The Sea*, vol. 3, ed. by M.N. HILL (1963) and vol. 4, ed. by A.E. MAXWELL (1970), are the most recent general accounts of ocean floors including the ridges. Shorter and more elementary are D. FLANAGAN (ed.), *The Ocean* (1969); and M.J. KEEN, *An Introduction to Marine Geology* (1968).

Descriptions of the ridges and their geophysical character-

istics are included in: H.W. MENARD, *Marine Geology of the Pacific* (1964); M. KAY (ed.), *North Atlantic: Geology and Continental Drift: A Symposium* (1969), especially the paper by J.E. NAFE and C.L. DRAKE, "Floor of the North Atlantic: Summary of Geophysical Data"; G.O. DICKSON, W.C. PITMAN, and J.R. HEIRTZLER, "Magnetic Anomalies in the South Atlantic and Ocean Floor Spreading," *J. Geophys. Res.* 73:2087–2100 (1968); and R.L. FISHER, J.G. SCLATER, and D.P. MCKENZIE "Evolution of the Central Indian Ridge, Western Indian Ocean," *Bull. Geol. Soc. Am.*, 82:553–562 (1971).

The interested reader also should consult: R.S. DIETZ and J.C. HOLDEN, "The Breakup of Pangaea," *Scient. Am.*, 223: 30–41 (1970); H. JOHNSON and B.L. SMITH, *The Megatectonics of Continents and Oceans* (1970); W.J. MORGAN, "Convection Plumes in the Lower Mantle," *Nature*, 230:42–43 (1971); and J.T. WILSON (ed.), *Continents Adrift* (1972), which discuss possible mechanisms for forming oceanic ridges.

(J.T.W.)

Oceans, Development of

Man's knowledge of the origin and evolution of the oceans has expanded greatly within the last two decades. To provide a unique explanation of the history of seawater requires answers to many other problems of earth history. Although some of these problems are at the forefront of scientific investigation in the earth sciences today and are far from being solved, it is possible to present a model for the development of the oceans that is reasonably consistent with available data and present concepts of earth history. It is, however, likely that the model will be revised and parts of it rejected as knowledge of the earth increases.

The concept of the oceans as a chemical system has changed from one in which the oceans have been thought of as a continuous accumulator of the salts brought down by rivers to a view in which the oceans are chiefly a mechanism of transfer of material from the continents to the sea floor. Today it is generally agreed that the ocean is a steady-state system, in which the continuous influx of dissolved species (material in solution) approximately equals their efflux. Current controversy concerns the degree to which seawater composition can vary through time, as a result of changing rates of influx or efflux or changing chemical composition.

Information about material being added to the oceans is fairly good as to both the minerals in the detrital (solid particles) load and the concentrations of species in the dissolved load of streams. Similar information, except for carbonate minerals (*q.v.*), is not available for the substances that are lost, however. Although estimates of the minerals formed and their relative amounts have been made, these have been neither proved nor disproved by observation.

The chemical history of the oceans has been divided into three stages. The first is an early stage in which the earth's crust was cooling and reacting with volatile or highly reactive gases of an acid, reducing nature to produce the oceans and an initial sedimentary rock mass. This stage lasted until about 3,500,000,000 years ago. The second stage was a period of transition from the first to essentially modern conditions, and it is estimated to have ended 1,500,000,000 to 2,000,000,000 years ago. Since that time it is likely that there has been little change in ocean-water composition.

THE EARLY OCEANS

The initial accretion of the earth by agglomeration of solid particles occurred about 5,000,000,000 years ago. Heating of this initially cool, unsorted conglomerate by the decay of radioactive elements and the conversion of kinetic and potential energy to heat resulted in the development of a liquid iron core and the gross internal zonation of the earth (see EARTH, STRUCTURE AND COMPOSITION OF). It has been concluded that formation of the earth's core took about 500,000,000 years. It is likely that core formation resulted in the escape of an original primitive atmosphere and its replacement by one derived from loss of volatile substances from the earth's interior (see ATMOSPHERE, DEVELOPMENT OF). Whether most of this degassing took place during core formation or soon afterward or whether there has been significant degassing of the earth's interior throughout geologic time is uncertain. Recent models of earth formation, however, suggest early differentiation of the earth into three major zones (core, mantle, and crust) and attendant early loss of volatile substances from the interior. It is also likely that the earth, after initial cold agglomeration, reached temperatures such that the whole earth approached the molten state. As the initial crust of the earth solidified, volatile gases would be released to form an atmosphere that would contain water (H_2O), later to become the hydrosphere; carbon gases, such as carbon dioxide (CO_2), methane (CH_4), and carbon monoxide (CO); sulfur gases, mostly hydrogen sulfide (H_2S); and halogen compounds, such as hydrochloric acid (HCl). Nitrogen also may have been present plus minor amounts of other gases. Gases of low atomic number, such as hydrogen and helium, would escape the earth's gravitational field. Substances degassed from the earth's interior have been called excess volatiles because their masses cannot be accounted for simply by rock weathering. An estimate of the masses of the various volatiles degassed throughout geologic time is given in Table 1.

Earth formation and gases released

Table 1: Estimate of "Excess Volatiles"
(units of 10^{20} grams)

Water	16,600
Total carbon as carbon dioxide	910
Sulfur	22
Nitrogen	42
Chlorine	300
Hydrogen	10
Boron, bromine, argon, fluorine, etc.	4

Source: W.W. Rubey (1951).

At an initial crustal temperature of about 600° C (1,100° F), almost all of these compounds, including H_2O, would be in the atmosphere. The sequence of events that occurred as the crust cooled is difficult to construct. Below 100° C (212° F) all of the H_2O would have condensed, and the acid gases would have reacted with the original igneous crustal minerals to form sediments and an initial ocean. There are at least two possible pathways by which these initial steps could have been accomplished.

One pathway assumes that the 600° C (1,100° F) atmosphere contains, together with other compounds, H_2O, CO_2, and HCl in the ratio of 20:3:1 and would cool to the critical temperature of H_2O. The H_2O therefore would have condensed into an early hot ocean. At this stage, the hydrochloric acid would be dissolved in the ocean ($\sim$ 1 mole per litre), but most of the CO_2 would still be in the atmosphere with about 0.5 mole per litre in the ocean water. This early acid ocean would react vigorously with crustal minerals, dissolving out silica and cations and creating a residue that consisted principally of aluminous clay minerals (*q.v.*) that would form the sediments of the early ocean basins. This pathway of reaction assumes that reaction rates are slow relative to cooling. A second pathway of reaction, which assumes that cooling is slow, is also possible. In this case, at a temperature of about 400° C most of the H_2O would be removed from the atmosphere by hydration reactions with pyroxenes and olivines (*q.v.*). Under these conditions H_2O would not condense until some unknown temperature was reached, and the earth might have had at an early stage in its history an atmosphere rich in CO_2 and no ocean: the surface would have been much like that of Venus today.

The pathways described are two of several possibilities for the early surface environment of the earth. In either case, after the earth's surface had cooled to 100° C (212° F), it would have taken only a short time geologically for the acid gases to be used up in reactions involving igneous rock minerals. The presence of bacteria and possibly algae in the fossil record (*q.v.*) of rocks greater than 3,000,000,000 years old attests to the fact that the earth's

surface had cooled to temperatures lower than 100° C (212° F) by this time and that the neutralization of the original acid gases had taken place. If most of the degassing of primary volatile substances from the earth's interior occurred early, the chloride released by reaction of HCl with rock minerals would be found in the oceans and seas or in evaporite deposits, and the oceans would have a salinity and volume comparable to those that they have today.

Early salinity and volume

This conclusion is based on the assumption that there has been no drastic change in the ratios of volatiles released through geologic time. The overall generalized reaction indicative of the chemistry leading to formation of the early oceans can be written in the form: Primary igneous rock minerals + acid volatiles + H_2O → sedimentary rocks + oceans + atmosphere. Notice from this equation that if all the acid volatiles and H_2O were released early in the history of the earth and in the proportions found today, then the total original sedimentary rock mass produced would be equal to that of today, and ocean salinity and volume would be near that of today. If, on the other hand, degassing were linear with time, then the sedimentary rock mass would have accumulated at a linear rate, as would oceanic volume. However, the salinity of the oceans would remain nearly the same if the ratios of volatiles degassed did not change with time. The most likely situation is that presented here, namely, that major degassing occurred early in earth's history, after which minor amounts of volatiles were released episodically or continuously for the remainder of geologic time. The salt content of the oceans based on the constant proportions of volatiles released would depend primarily on the ratio of sodium chloride (NaCl) locked up in evaporites to that dissolved in the oceans. If all the NaCl in evaporites were added to the oceans today, the salinity would be approximately doubled. This value gives a sense of the maximum salinity the oceans could have attained throughout geologic time.

Production of oxygen

One component missing from the early earth's surface was free oxygen, because it would not have been a constituent released from the cooling crust. Early production of oxygen was by photodissociation (separation due to the energy of light) of water in the earth's atmosphere as a result of adsorption of ultraviolet light. The reaction is $2H_2O + h\nu \rightarrow O_2 + 2H_2$ in which $h\nu$ represents photon of ultraviolet light. The hydrogen produced would escape into space, and the O_2 would react with the early reduced gases by reactions such as $2H_2S + 3O_2 \rightarrow 2SO_2 + 2H_2O$. Oxygen production by photodissociation gave the early reduced atmosphere a start toward present-day conditions, but it was not until the appearance of photosynthetic organisms approximately 3,000,000,000 years ago that it was possible for the accumulation of oxygen in the earth's atmosphere to proceed at a rate sufficient to lead to today's oxygenated environment. The photosynthetic reaction leading to oxygen production may be written $6CO_2 + 6H_2O + h\nu \rightarrow C_6H_{12}O_6 + 6O_2$, in which $C_6H_{12}O_6$ represents sugar.

THE TRANSITION STAGE: 3,500,000,000–1,500,000,000 YEARS AGO

The nature of the rock record from the time of the first sedimentary rocks (~3,500,000,000 years ago) to about 1,500,000,000 to 2,000,000,000 years ago suggests that the amount of oxygen in the earth's atmosphere was significantly lower than today and that there were continuous chemical trends in the sedimentary rocks formed and, more subtly, in oceanic composition. The source rocks of sediments during this time were likely to be more basaltic than would later ones; sedimentary detritus was formed by the alteration of these rocks in an oxygen-deficient atmosphere and accumulated primarily under anaerobic marine conditions. The chief difference between reactions involving mineral-ocean equilibria at this time and today was the role played by ferrous iron. The concentration of dissolved iron in today's oceans is low because of the insolubility of oxidized iron oxides. During the period 1,500,000,000–3,500,000,000 years ago, oxygen-deficient environments were prevalent; these favoured the formation of minerals containing ferrous iron (reduced state of iron) from the alteration of basaltic rocks. Indeed, the iron carbonate siderite and the iron silicate greenalite, in close association with chert and the iron sulfide pyrite, are characteristic minerals that occur in middle Precambrian iron formations. The chert originally was deposited as amorphous silica; equilibrium between amorphous silica, siderite, and greenalite at 25° C (77° F) and one atmosphere total pressure requires a CO_2 pressure of about $10^{-2.5}$ atmospheres, or ten times today's value.

The oceans at this time can be thought of as the solution resulting from an acid leach of basaltic rocks, and because the neutralization of the volatile acid gases was not restricted primarily to land areas as it is today, much of this alteration may have occurred by submarine processes. The atmosphere at the time was oxygen deficient; anaerobic depositional environments with internal CO_2 pressures of about $10^{-2.5}$ atmospheres were prevalent, and the atmosphere itself may have had a CO_2 pressure near $10^{-2.5}$ atmospheres. If so, the pH of early ocean water was lower than that of modern seawater, the calcium concentration was higher, and the early ocean water was probably saturated with respect to amorphous silica (~ 120 parts per million [ppm]).

The pH of seawater

To simulate what might have occurred, it is helpful to imagine emptying the Pacific Basin, throwing in great masses of broken basaltic material, filling it with HCl so that the acid becomes neutralized, and then carbonating the solution by bubbling CO_2 through it. Oxygen would not be permitted into the system. The HCl would leach the rocks, resulting in the release and precipitation of silica and the production of a chloride ocean containing sodium (Na), potassium (K), calcium (Ca), magnesium (Mg), aluminum (Al), iron (Fe), and reduced sulfur species in the proportions present in the rocks. As complete neutralization was approached, Al could begin to precipitate as hydroxides and then combine with precipitated silica to form cation-deficient aluminosilicates. The aluminosilicates, as the end of the neutralization process was reached, would combine with more silica and with cations to form minerals like chlorite, and ferrous iron would combine with silica and sulfur to make greenalite and pyrite. In the final solution, chlorine (Cl) would be balanced by Na and Ca in roughly equal proportions, with subordinate K and Mg; Al would be quantitatively removed, and silicon (Si) would be at saturation with amorphous silica. If this solution were then carbonated, Ca would be removed as calcium carbonate ($CaCO_3$), and the Cl balance would be maintained by abstraction of more Na from the primary rock. The sediments produced in this system would contain chiefly silica, ferrous iron silicates, chloritic minerals, calcium carbonate, calcium-magnesium carbonates, and minor pyrite.

If the HCl added were in excess of the CO_2, the resultant ocean would have a high content of calcium chloride ($CaCl_2$); but the pH would still be near neutrality. If the CO_2 added were in excess of the Cl, Ca would be precipitated as the carbonate until it reached a level approximately that of the oceans today, namely, a few hundred parts per million.

If this newly created ocean were left undisturbed for a few hundred million years, its waters would evaporate and be transported onto the continents (in the form of precipitation); streams would transport their loads into it. The sediment created in this ocean would be uplifted and incorporated into the continents. Gradually, the influence of the continental debris would be felt and the pH might shift a little. Iron would be oxidized out of the ferrous silicates to make iron oxides, but the water composition would not vary a great deal.

The primary minerals of igneous rocks are all mildly basic compounds. When they react in excess with acids such as HCl and CO_2, they produce neutral or mildly alkaline solutions plus a set of altered aluminosilicate and carbonate reaction products. It is improbable that ocean water has changed through time from a solution approximately in equilibrium with these reaction products, which are clay minerals and carbonates.

CHEMICAL VIEW OF MODERN OCEANS

It is likely that the oceans achieved their modern characteristics 1,500,000,000 to 2,000,000,000 years ago. The chemical and mineralogical compositions and the relative proportions of sedimentary rocks of this age differ little from their Paleozoic counterparts. Calcium sulfate deposits of late Precambrian age testify to the fact that the acid sulfur gases had been neutralized to sulfate by this time. Chemically precipitated ferric oxides in late Precambrian sedimentary rocks indicate available free oxygen, whatever its percentage. The chemistry and mineralogy of middle and late Precambrian shales is similar to that of Paleozoic shales. Thus, it appears that continuous cycling of sediments like those of today has occurred for 1,500,000,000 to 2,000,000,000 years and that these sediments have controlled oceanic composition.

Salt content of the oceans

It was once thought that the saltiness of the modern oceans simply represents the storage of salts derived from rock weathering and transported to the oceans by fluvial processes. With increasing knowledge of the age of the earth, however, it was soon realized that, at today's rate of delivery of salts to the ocean or even at much reduced rates, the total salt content and the mass of individual salts in the oceans could be attained in geologically short-time intervals compared to earth's age. The total mass of salt in the ocean can be accounted for at today's rates of stream delivery in about 12,000,000 years. The mass of dissolved silica in ocean water can be doubled in only 20,000 years by addition of stream-derived silica; to double sodium would take 70,000,000 years. It then became apparent that the oceans were not simply an accumulator of salts, but as water evaporated from the oceans, along with some salt, the introduced salts must be removed in the form of minerals. Thus, the concept of the oceans as a chemical system changed from that of a simple accumulator to that of a steady-state system, in which rates of inflow of materials into the oceans equal rates of outflow. The steady-state concept permits influx to vary with time, but it would be matched by nearly simultaneous and equal variation of efflux. Calculations of rates of addition of elements to the oceanic system and removal from the system show that for at least 100,000,000 years the oceanic system has been in a steady state with approximately fixed rates of major element inflow and outflow and, thus, fixed chemical composition.

Mineral–seawater equilibria. In recent years it has been shown that not only is the oceanic system steady state but that its composition is probably controlled by chemical equilibria involving seawater and minerals found in marine sediments. In 1961, L.G. Sillén published a paper on *The Physical Chemistry of Sea Water*, in which he showed that the chemical composition of seawater is approximately that of a theoretical solution brought to chemical equilibrium with the minerals quartz, illite, montmorillonite, chlorite, kaolinite, calcite, and phillipsite, and the atmosphere. He pointed out that in his model the ratios of the major chemical species would be fixed but that their absolute concentrations could increase or decrease with changes in the concentration of NaCl. Sillén's model further implies that even if the rates of addition of feed material to the oceanic system are changed or the proportions of minerals in the feed, the chemical composition of seawater will be invariant.

Equilibrium with silicate minerals

If seawater maintains near equilibrium with the detrital silicates that have been falling through it for billions of years, its composition should correspond to that predicted by thermodynamic calculations of the composition of an aqueous solution in equilibrium with those phases. It is impossible, of course, for ocean water to have one equilibrium composition, a balance that depends upon temperature and pressure, among other factors, because its temperature ranges from 0° to 35° C (32° to 95° F) and its pressure from 1 to 1,000 atmospheres. Moreover, it cannot be in equilibrium simultaneously with all the minerals being delivered. Gibbsite and feldspar, for example, are incompatible because if placed together in a solution they would react: one would disappear and the other would be left with the final solution and an intermediate phase.

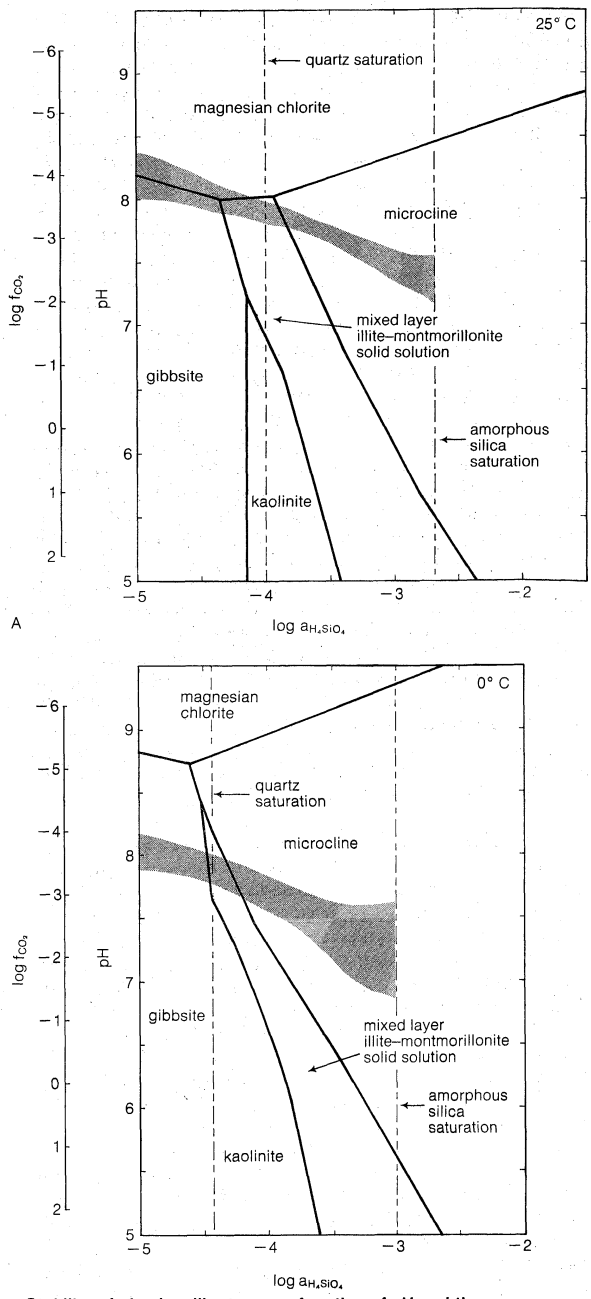
The Figure shows the phases that would be in equilibrium with present-day seawater at 25° C and 0° C (77° F and 32° F), respectively. The stability fields of the phases are shown as functions of pH and dissolved silicon dioxide (SiO_2). The shaded areas denote the range of seawater compositions. Gibbsite, kaolinite, montmorillonite, K-feldspar, and illite all can be stable species within the commonly observed range of seawater composition. At a P_{CO_2} near that of today's atmosphere, calcite is also a stable phase. This relation does not prove that these minerals are the cause of seawater composition, but it does show that if a mixture of these minerals were placed in a NaCl solution with the same Cl as seawater, maintained in equilibrium with the atmosphere, and allowed to come to equilibrium with the final solution, no matter which of the various coexisting mineral assemblages remained, it would fall within the typical range of composition of seawater.

The chemical equilibrium relations predict that present-day seawater has a chemical composition that is compatible with that expected for a solution continuously reacting with the great variety of minerals being added to it. Of these minerals, it is nearly in equilibrium with the typically detrital clay minerals and is out of equilibrium with most of the primary igneous rock minerals such as plagioclase feldspar, pyroxenes, and olivines. This situation may be transient and but an instant in a long-term chemical trend. On the other hand, it is compatible with continuous control of seawater by the synthesis and dissolution of a few silicate phases, resulting in seawater that never deviates far from saturation.

Mass balance of the oceans. The various equilibrium models of the oceanic system imply that the major chemical components of seawater are held near saturation values (value at which addition of the component will result in precipitation); oversaturation may occur, but the departure from saturation should not be great. For calcium carbonate, this situation has been shown to be a distinct possibility within the limits of measurement; the influx of calcium carbonate is equal to its efflux, and ocean water is slightly oversaturated with respect to calcium carbonate ($CaCO_3$) in the upper few hundred metres and slightly undersaturated at depth. For the other dissolved constituents brought to the oceans by streams, the mechanisms of removal are not well known. Furthermore, if these constituents do enter newly formed minerals, the question remains whether the formation of the minerals can hold their chemical components near saturation values in seawater.

Calcium carbonate equilibrium

Current models of seawater composition imply that stream-derived dissolved species must precipitate from ocean water or be cycled back through the atmosphere to fall on the continents. Balance sheets have been drawn up to account for the removal of these constituents. The most complete balance to date provides for removal of the various constituents of stream water in minerals commonly found in modern and ancient sediments and sedimentary rocks. In this balance, it was found necessary to remove part of the stream-derived Na, Mg, K, and dissolved silica by reaction with poorly crystalline, alumina- and silica-rich solids of the suspended load of streams, when these materials reach the oceans. The minerals chosen as the most likely candidates for removal of constituents consisted of 28 percent evaporite minerals (salt and gypsum), 26 percent carbonate rock minerals (calcite, dolomite, etc.), 41 percent clay minerals (kaolinite, montmorillonite, illite, and chlorite), and minor percentages of pyrite and silica. Several of the proposed mineral species have not been conclusively demonstrated as forming in seawater. The amounts of new minerals formed each year that are necessary to remove dissolved constituents and maintain a steady-state oceanic system are very small. In this balance the point was made that to maintain a constant composition ocean requires reverse weathering. If the CO_2 consumed in the weathering process by reaction with carbonate and silicate rock minerals is not returned to the atmosphere by depositional processes, decrease in the CO_2 content of the atmosphere will result; and ocean water, because of the accumulation of bicarbonate ion, will tend to become alkaline. Because the weathering of



Stability of aluminosilicates as a function of pH and the activity of dissolved silica drawn for seawater of present composition, one atmosphere total pressure, unit activity of H_2O, at (A) 25° C and (B) 0° C (see text).

calcium carbonate is known to be balanced by an equal rate of calcium carbonate deposition in the oceans, any net drain on the CO_2 of the atmosphere will come from failure of reversal of silicate weathering reactions. About 20 percent of the bicarbonate ion (HCO_3^-) in stream waters is derived from the neutralization of CO_2 involved in silicate weathering; the reaction is siliceous silicate + CO_2 + H_2O = less siliceous silicate + cations + dissolved silica + HCO_3^-. The mass of CO_2 in the present atmosphere is about 1.6×10^{18} grams, that neutralized by weathering of rock minerals on the continents is 1.3×10^{14} grams per year. Therefore, removal of all CO_2 by weathering of carbonate and silicate minerals at the present rate of consumption would require only about 12,000 years. This value provides an estimate of the rate required for restorative processes of CO_2 addition to the atmosphere if the atmospheric CO_2 content is to remain at today's level and the approximate time scale on which the atmosphere might change if the restorative processes failed to operate. No reversal of silicate weathering in the oceanic system during deposition or on a time scale of a few tens of thousands of years of burial would mean that the atmosphere could be nearly completely drained of CO_2; this situation is not likely to have obtained during much of geologic time.

Experimental evidence. Another approach to the question of what mechanisms remove dissolved materials added to the oceans is laboratory investigation of the reactions between clay minerals and aqueous solutions containing cations and dissolved silica. If it can be demonstrated that clay minerals react rapidly when the new phases formed are like those found in marine deposits and if the resultant solution has a chemical composition like that of seawater, then the feasibility of the maintenance of a steady-state oceanic system with a small range of composition will have been shown.

Clay minerals and silica equilibrium

Clay minerals have been added to seawater solutions initially containing little dissolved silica (~0.03 ppm) or an excess of silica (~25 ppm). In both cases the pH and silica concentration of the solutions were monitored with time; constant values of pH and silica concentration were reached in short time intervals compared to the time of exposure to seawater of clay minerals deposited in the marine environment. In the silica-deficient experiments, dissolved silica was added to the seawater by reaction with the clay minerals; constant values of 2–6 ppm were reached. For the silica-enriched solutions, dissolved silica was removed by reaction with clay minerals; the constant concentration values obtained were higher and more variable, ranging from 6 to 18 ppm. Release of silica from the minerals into silica-deficient seawater was accompanied by formation of a second compound, less siliceous than the parent clay, whereas uptake of silica from the enriched solutions took place by adsorption of molecular silica or by the formation of a new phase more siliceous than the one that was reacted. Further experimental work has shown that the new phases produced are probably not simple nonaluminous silicates, whereas theoretical thermodynamic calculations are consistent with the phases, being similar in composition to hydrated aluminum oxides and aluminosilicates (clay minerals).

The experiments show that the clay minerals of the detrital load of streams, when they enter the oceans, should be able to influence the dissolved silica content of the oceans, and they can influence the cation concentrations and the pH if the new phases formed are aluminosilicates. Unfortunately, the silica-uptake experiments did not demonstrate that a new phase is formed in accord with the reverse weathering reaction; if silica uptake were by adsorption of molecular silica, there would be no effect on the bicarbonate balance and the return of CO_2 to the atmosphere. Field tests of the reversible nature of silicate weathering today may serve to demonstrate the validity of the reversal of the reaction by finding an appropriate amount of silicate regeneration taking place in the marine environment. Alternatively, such tests may show that the dissolved silica brought to the oceans is precipitated in molecular form as the tests of siliceous organisms, thus proving indirectly that silicate synthesis over a short period of time does not take place. The results of attempts to apply both tests are inconclusive to date. No clear-cut examples of regeneration reactions of the magnitude required have been discovered, and a controversy rages over the rate at which siliceous organisms are abstracting silica from seawater and accumulating in marine sediments (*q.v.*). Furthermore, because the silicate reverse weathering reactions can involve intervals of time of a few tens of thousands of years before the CO_2 content of the atmosphere is affected, it is possible that these reactions take place as the sediment is being buried and involve the return of materials through the sediment column to the hydrosphere–atmosphere. Our knowledge of the quantitative importance of this path is even less than that of the significance of early reverse weathering reactions.

SEDIMENTARY ROCK MASS AND COMPOSITION

Sufficient data now have accumulated to predict to a first approximation the total mass of sedimentary rocks (*q.v.*)

that has accumulated throughout geologic time and the distribution of the mass as a function of time. The actual distribution and mass of the various rock types also can be estimated. Because of the chemistry and mineralogy of the sedimentary rocks that have been used to determine the nature of the hydrosphere–atmosphere through time, however, it is worthwhile at this point to evaluate what information can be gained from these rocks and the concepts of differential sedimentary cycling and chemical uniformitarianism, that is, that present chemical processes are similar in kind to those which were operative in the past.

Cycling of sedimentary rocks through time

A major problem faced by students of sedimentary rocks is whether the compositions of these rocks and the distribution of rock types through time are representative of the compositions and ratios at the time of deposition or whether they have been modified by processes that were operative after burial. If the features are primary, then they can be used directly to evaluate the history of the atmosphere–hydrosphere; if they are secondary, then conclusions need to be tempered by inclusion of the processes that modify the compositions and ratios through time. For example, the relatively high proportion of iron formations in the sedimentary rock mass of middle Precambrian age previously was taken as representative of a primary feature of the rocks accumulated during this time. Thus, the iron formations are indicative of a hydrosphere–atmosphere different from today. Also, the appearance of calcium sulfate deposits in rocks of late Precambrian age has been used as evidence to suggest that by this time most of the reduced sulfur gases degassed from the earth's interior had been oxidized to sulfate. However, there is evidence that suggests that some features of the sedimentary rock mass are secondary.

Recent analysis of the mass and distribution of sedimentary rock types as a function of time has shown that the mass of sedimentary rocks remaining is about $32{,}000 \times 10^{20}$ grams and that this mass is distributed in such a way that 56 percent of it consists of rocks accumulated during the last 600,000,000 years of geologic time whereas the remainder is of Precambrian age. The centre of mass is at 500,000,000 to 600,000,000 years. Simple mathematical models of this distribution show that the best fit is obtained if five times as much sediment has been deposited during geologic time as accumulated; that is, the whole sedimentary mass has been turned over five times. These models also demonstrate that limestones cycle at a rate equal to two times the sedimentary rock mass as a whole, and evaporites three times. Thus the centres of mass of limestones and evaporites are at about 300,000,000 and 200,000,000 years, respectively. One of the important points of these models is that, because of differential cycling rates of the various sedimentary rock types, the limestones and evaporites form a progressively larger proportion of sedimentary rocks accumulated per unit time with decreasing rock age. This relationship is a secondary feature of the sedimentary rock mass and is a result of the selective, more rapid destruction by erosion of carbonates and evaporites than shales and sandstones and redeposition of these types farther ahead in time. Thus, the ratios of sedimentary rock types of a particular age that remain today do not necessarily represent the ratios of the same types at the time of deposition. Furthermore, the use of these ratios to explain hydrosphere–atmosphere evolution must be considered in light of differential sedimentary cycling.

Two further aspects of modelling of the sedimentary rock mass that are of importance in considering the development of the oceans are: (1) The proportions of rock types deposited at any particular time since 1,500,000,000–2,000,000,000 years ago have been nearly constant; that is, as a first approximation, the types and quantities of sedimentary materials passing through the oceans have been nearly the same for this interval of time. As a consequence the oceanic system has been in a steady state, with a nearly fixed seawater composition governed by reactions between the materials entering the oceans and the water. (2) The nature of the detrital and dissolved materials entering the oceans throughout this time interval has not changed significantly. These materials represent a slow, selective, postdepositional water leaching of substances such as $CaCO_3$ in shales and selective, more rapid destruction of easily eroded materials such as limestones and evaporites. The predominance of the clay mineral illite with minor chlorite in pre-Mesozoic/Cenozoic shales and the abundance of the clay minerals montmorillonite and kaolinite in younger shales is thought to be a postdepositional feature resulting from the reaction: (high silica) montmorillonite + (low silica) kaolinite + potassium → (intermediate silica) illite + SiO_2, Na, Mg (lost from rock).

The materials lost from shales through time by the above process, plus materials gained by leaching and selective destruction of $CaCO_3$ and evaporite minerals enter streams and are redeposited in the oceans. The rock mass remaining is changed through time in mineralogy and chemistry and in proportions of rock types, but the materials brought to the oceans in the last 1,500,000,000 to 2,000,000,000 years to a first approximation have remained nearly the same, a statement of the concept of chemical uniformitarianism. Differential sedimentary cycling creates problems in the interpretation of hydrosphere–atmosphere evolution because the rock ratios of a particular age may not represent the ratios at the time of deposition; however, chemical uniformitarianism implies that the nature of the materials transported to the oceans since 1,500,000,000 to 2,000,000,000 years ago has been nearly constant. This constancy of composition of materials passing through the oceans suggests that the oceanic system, when viewed over time intervals of 200,000,000 or 300,000,000 years, is in a steady state and that seawater has had a composition since middle Precambrian time much like that existing today.

PERIODIC COMPOSITIONAL EXCURSIONS

Alterations of the model

It is premature to try to evaluate the fluctuations in seawater composition that would have resulted from time to time, as coincidences of various sets of physical and biological conditions changed the composition of the material going into or coming out of the oceans. If there were times when almost no erosion of the land was occurring, then seawater might have had a chance to react more completely with detrital primary rock materials, as opposed to being continuously exposed to the altered products of weathering. At best a few limits that have been established can be cited. In general, they constitute negative evidence, and the arguments are based on the absence of a mineral that would form if seawater composition were changed in a given way. Knowledge of such limitations is scant; little experimentation on changing the composition of seawater (with the exception of evaporating it) has been done.

The silica content of seawater cannot increase above 25 ppm at a pH of eight without precipitating a magnesium silicate such as sepiolite ($Mg_2Si_3O_6[OH]_4$). Sepiolite is found as a mineral in oceanic sediments, but its occurrences are rare. It is common in the deposits of saline lakes but relatively unimportant in sedimentary rocks.

If the pH of seawater is increased to about ten, brucite ($Mg[OH]_2$) would precipitate. Brucite, like sepiolite, occurs in sedimentary rocks, but it does not exhibit the extensive distribution that would result from precipitation throughout the whole mass of seawater.

It can be calculated that instantaneous addition and dissolution of about 7 percent of the evaporite deposits into the oceans would saturate them with respect to gypsum. On the other hand, the total salinity would be increased only a few parts per thousand. There is no record of gypsum in sedimentary rocks interpreted as open-ocean deposits; all the evaporites have formed in restricted portions of the sea. Gypsum seems to be a mineral whose component concentrations in the open oceans always have been less than the saturation value. Furthermore, the absence of evidence in the rocks for open-ocean gypsum requires the continuous storage of most of the sulfate in evaporites and, hence, uniformitarianism.

These are a few of the kinds of restrictions that can now be suggested for the limits of seawater chemical

variation; eventually a much greater number will be investigated by experimentation.

THE PRESENT HYDROSPHERE

Ocean waters and sediment pore waters form most of the present hydrosphere (Table 2). About 80 percent of the water in the hydrosphere is contained in the oceans and seas. The pores of sediments and sedimentary rocks hold nearly all the rest: about 24 percent of that in the oceans or 20 percent of the total. Ice now locks up a little more than 1 percent of the total and may have accounted for 3 percent or so during the peak of the last Ice Age, but water storage in rivers, lakes, or the atmosphere represents a trivial percentage of the total.

Table 2: Mass of the Present Hydrosphere

	total mass (units of 10^{20} grams)	percentage of total hydrosphere
Oceans	13,700	80.0
Pore waters in sediments	3,300	18.8
Ice	200	1.2
Rivers, lakes	0.3	0.002
Atmosphere	0.13	0.0008
Total hydrosphere	17,200	100.0

Relationship of freshwater to seawater

On the other hand, the rate of water circulation through the rain–river–ocean–atmosphere system or hydrologic cycle (*q.v.*) is relatively fast; the amount of water discharged into the oceans each year from the land is 0.32×10^{20} grams and is approximately equal to the total mass of water stored at any instant in rivers and lakes. The average dissolved solid concentration of stream water is 130 ppm; thus about 42×10^{14} grams of dissolved constituents are carried to the oceans yearly by streams, the main source of materials for the ocean.

Seawater has an average salinity of about 35,000 ppm; the proportions of the various dissolved solids in seawater and river water are markedly different. River water is a bicarbonate solution containing significant proportions of calcium, dissolved silica, and sulfate, whereas seawater is a sodium chloride solution (Table 3).

Table 3: Major Constituents of River Water and Seawater

constituents	river water		seawater	
	ppm	millimoles per liter	ppm	millimoles per liter
Cl^-	7.8	0.220	19,000	535.2
Na^+	6.3	0.270	10,500	456.2
Mg^{2+}	4.1	0.171	1,300	54.2
SO_4^{2-}	11.2	0.117	2,650	27.6
K^+	2.3	0.059	380	9.7
Ca^{2+}	15	0.375	400	10.0
HCO_3^-	58.4	0.958	140	2.3
SiO_2	13.1	0.218	6	0.1
NO_3^-	1	0.016	—	—
Fe^{2+}	0.67	0.012	—	—
Al	0.01	—	0.001	—
Br^-	—	—	65	0.8
CO_3^{2-}	—	—	18	0.3
Sr^{2+}	—	—	8	0.1
Dissolved organic C	9.6	—	0.5	—
Sum	129.5	—	34,467	—

The mass balance between streams and oceans previously discussed indicated an attempt to demonstrate how seawater composition could not result from simple evaporation and concentration of river-derived dissolved constituents but involves processes of chemical differentiation and evaporation of river water that affect each element differently and thereby result in waters of different compositions.

The oceans were once commonly thought of as a huge storage tank of water and dissolved solids; however, some comparisons of masses of material and rates of material transfer in the crust–ocean system are of use in negating this idea in favour of the steady-state concept. The mass of the oceans is only about 20 to 30 percent of the mass of sedimentary rocks now in existence and about 5 percent of the earth's crust. The mass of materials brought to the oceans by streams, small as they seem on a yearly basis, over geologic time would overwhelm the ocean basins. It would take only 12,000,000 years to account for the mass of dissolved constituents in the oceans at today's rate of stream delivery. Enormous tonnages of materials have passed through the oceans throughout geologic time, perhaps more than a million trillion metric tons. Dissolved solids constitute only 3.5 percent of the oceanic mass; thus storage by the oceans accounts for roughly 1 percent of the materials currently involved in the sedimentary cycle. Little wonder that the idea of the oceans as a simple accumulator of materials has given way to the steady-state concept.

The dissolved materials transported by streams to the oceans remain as dissolved constituents in the oceans or interstitial waters of marine sediments and are chemically or biochemically precipitated or are released into the atmosphere as sea-spray particles generated at the ocean–atmosphere interface. These tiny salt particles fall back on the sea surface or are transported over the continents by winds and fall as rain or dry fallout on the land surface. Thus recycled salts are added to the stream-borne dissolved constituents derived from rock weathering. About 50 percent of the sodium and chloride in stream waters is derived from the atmosphere. For other constituents, the percentages are less.

More than 99 percent of the dissolved solids in river water can be accounted for by the constituents chloride, sodium, sulfate, magnesium, calcium, potassium, bicar-

Table 4: Major Dissolved Constituents Delivered to Oceans Annually and Time Required for Them to Reach Oceanic Amounts, with and without Correction for Atmospherically Cycled Salts

constituent	mass delivered by rivers to ocean annually (units of 10^{14} grams)	mass in ocean (units of 10^{20} grams)	storage time—time for river fluxes to attain oceanic amounts (units of 10^6 year)	time for river fluxes to attain oceanic amounts correcting for atmospherically cycled salts (units of 10^6 year)	residence time* (units of 10^6 year)	
					river input	sedimentation
Fe^{2+}	0.223	0.0000137	0.00006	0.00006	—	0.00001
Al	0.003	0.0000137	0.0046	0.0046	0.004	—
SiO_2	4.26	0.08	0.02	0.02	0.04	0.01
HCO_3^-	19.02	1.9	0.1	0.1	—	—
Ca^{2+}	4.88	6	1.23	1.24	1	8
K^+	0.74	5	6.8	8	10	11
SO_4^{2-}	3.67	37	10.1	10.7	—	—
Mg^{2+}	1.33	19	14.3	15.4	22	45
Na^+	2.07	144	69.7	108	210	260
Cl^-	2.54	261	103	23	—	—
Dissolved organic C	3.2	0.007	0.0002	—	—	—
H_2O	325,000	13,550	0.042	—	—	—

*Residence time is expressible in terms of the same variables as storage time; however, residence time calculations may or may not contain corrections for atmospherically and rock-derived cyclic materials and for eolian particulate matter. Also, the amount of the constituent in suspension in the ocean or streams may enter the residence time calculation.

bonate, and silica; the same constituents make up more than 99 percent of the dissolved materials in ocean water. A feeling for the huge amount of H_2O and dissolved solids carried to the oceans by streams can be attained simply by calculating the time it would take for river fluxes alone to attain oceanic amounts. Table 4 shows that these times vary considerably but that only chloride should require more than 100,000,000 years. A correction for atmospherically cycled marine salts can be applied to river fluxes, and the same calculation made. The time necessary to attain oceanic amounts for river-derived sodium and chloride is about doubled when the corrections for cyclic sodium and chloride are made. This further illustrates the large amount of sodium and chloride that is cycled through the atmosphere. The other major constituents are predominantly derived from the weathering of rocks, and the correction for atmospherically cycled salts affects only slightly the time required for them to reach oceanic amounts.

These times, uncorrected for cyclic salts, may be called storage times; storage time can be expressed by the equation:

$$\lambda = \frac{A}{dA/dt}$$

in which A is the total amount of the constituent in the ocean, and dA/dt is the amount in solution introduced by streams per unit time. The storage time of a particular constituent in the ocean, if it is considered that the ocean is a well-mixed, steady-state system, is indirectly a measure of the rate at which stream-derived dissolved constituents leave the ocean.

BIBLIOGRAPHY. R.M. GARRELS and F.T. MACKENZIE, *Evolution of Sedimentary Rocks* (1971), a modern text dealing with global aspects of the origin and history of sedimentary rocks, continents, and oceans; A.B. RONOV, "Probable Changes in the Composition of Sea Water During the Course of Geological Time," *Sedimentology*, 10:25–43 (1968), a review and discussion of the possible changes in the chemical composition of seawater during geologic time; W.W. RUBEY, "Geologic History of Sea Water: An Attempt to State the Problem," *Bull. Geol. Soc. Am.*, 62:1111–1147 (1951), a classic, well-written paper on the history of seawater and the concept of "excess volatiles"; L.G. SILLEN, "The Physical Chemistry of Sea Water," in *Oceanography*, ed. by M. SEARS (1961), quantitative treatment of ocean water as a solution in equilibrium with the solids within it.

(F.T.M.)

Oceans and Seas

The oceans and seas cover about 71 percent of the Earth's surface and constitute its most conspicuous feature. These waters, together with the relatively small amount that occurs in the form of rivers, lakes, ice, and groundwater, are called the Earth's hydrosphere. The other physical spheres of the Earth are the atmosphere and the lithosphere (the rock sphere of the Earth).

The oceans and seas form an integrated unit and together may properly be called the World Ocean. The Caspian Sea and the Dead Sea, however, are generally considered to be salty lakes. The exact boundaries between the various seas and oceans are arbitrarily defined and have been fixed by convention (see Figure 1).

For many years, five oceans were accepted, namely the Atlantic, Pacific, Indian, Arctic, and Antarctic oceans. After the work of Otto Krümmel (*Handbuch der Ozeanographie* 1897), however, it became common practice to recognize only three oceans, the Atlantic, Pacific, and Indian. The Arctic Ocean is now regarded as belonging to the Atlantic Ocean; this is a not unreasonable view, because it is a marginal sea of the Atlantic. The Bering Strait, which divides the Arctic from the Pacific Ocean, is only 58 kilometres (36 miles) wide and 58 metres (190 feet) deep.

The great Southern Ocean, as it is sometimes called, is one continuous stretch of water encircling the Antarctic continent. By convention, it has been divided into three portions, one for each of the three principal oceans. The dividing line between the Atlantic and the Indian oceans is the meridian through Cape Agulhas, South Africa (20° E). The Atlantic is divided from the Pacific Ocean by a line extending from Cape Horn at the southern tip of South America to the South Shetland Islands, off the tip of the Antarctic continent, in the Falkland Islands dependency; in the north, the separation consists of the narrowest part of the Bering Strait, separating Alaska and western Siberia. The dividing line between the Indian Ocean and the Pacific extends from the Malay Peninsula through Sumatra, Java, Timor, and Cape Londonderry in Australia to Tasmania and thence continuing along the meridian of 147° E, directly south to Antarctica. Notwithstanding these conventional divisions, the great marine area in the far south, the Southern Ocean, exhibits global continuity and marked oceanographic and meteorological features that distinguish it as a physical entity. One outstanding feature is the predominance of west winds, which give rise to the powerful West Wind Drift, a broad, closed west–east current in the upper water layers.

The oceans predominate over land areas in the Southern Hemisphere far more than they do in the Northern Hemisphere; the ratio of water to land area is roughly 4:1, 81:19 in the Southern Hemisphere and roughly 3:2, 61:39 in the Northern. Considering zones, or belts, on the earth's surface at intervals of five degrees latitude, land predominates only between 45° and 70° N, where the Eurasian continent lies, and between 70° and 90° S, which is the location of Antarctica. Everywhere else the oceans predominate; indeed, between about 84° and 90° N there is no known land at all, and from 45° to 66° S only a very small fraction of the surface is land. The areas of the Atlantic, the Indian, and the Pacific oceans, including their marginal seas, are roughly in the proportion of 10:7:17, respectively. More accurate figures are given in Table 1, which also shows the areas of various marginal seas.

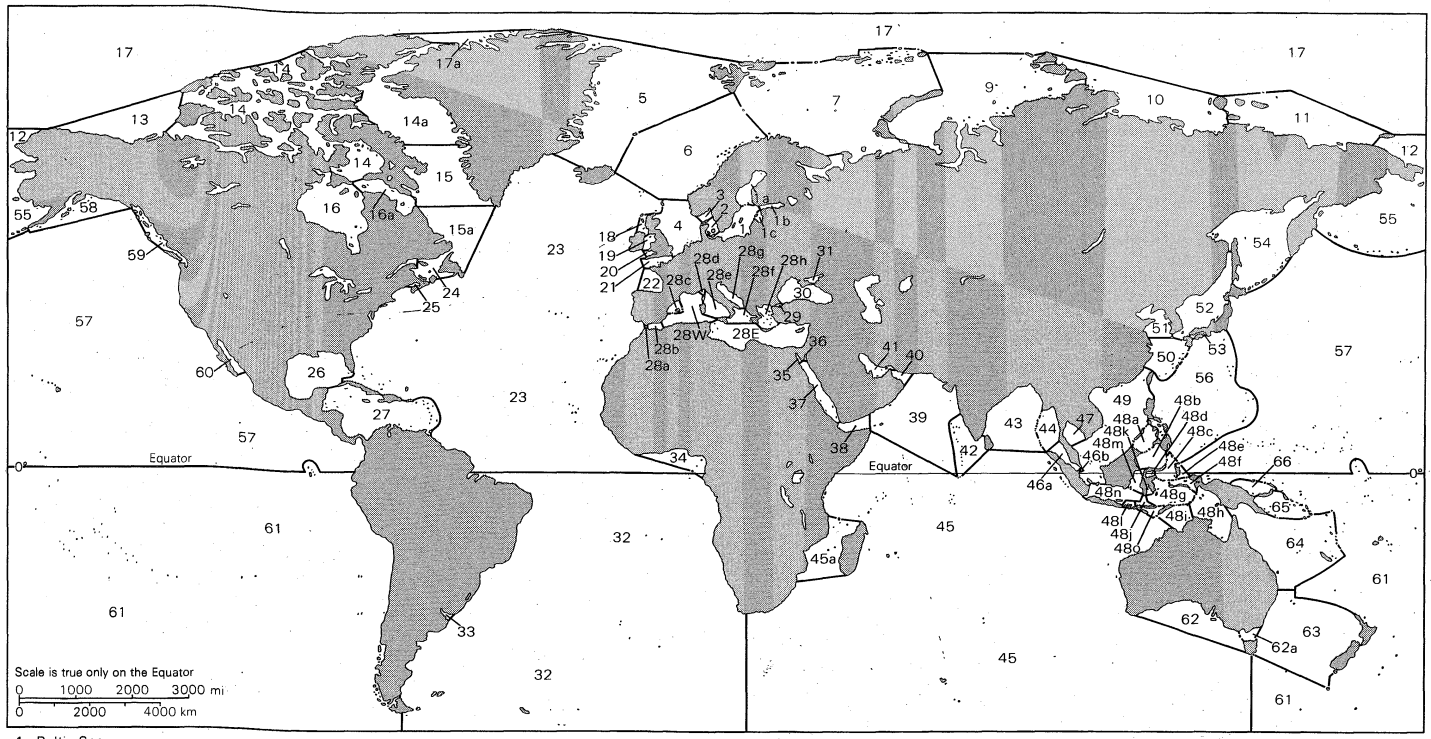
In shape, the Atlantic Ocean is distinctly oblong in a north–south direction and is much more irregular than the other two oceans. It narrows between the eastern tip of Brazil and the western bulge of Africa, and this aspect of its shape is striking. On both sides of the North Atlantic but especially on the eastern side, large marginal seas and bays occur; the Mediterranean and Black Sea, the Bay of Biscay, the North Sea, the Baltic Sea, all on the eastern side, Baffin Bay, Hudson Bay, the Gulf of Mexico, and the Carribbean Sea on the western side are examples (see GULFS AND BAYS). By contrast, the configuration of the South Atlantic is much smoother, as is the eastern littoral (coastal region) of the Pacific Ocean from north to south. The western border of the Pacific, however, is indented considerably by adjacent seas such as the Sea of Okhotsk, the Sea of Japan, the East China Sea, the Yellow Sea, the South China Sea, and the seas of the Indonesian archipelago. Finally, in the Indian Ocean, only a few seas of significance occur along its northern margin, namely the Red Sea, the Persian Gulf, the Arabian Sea, and the Bay of Bengal.

The Atlantic Ocean has the greatest length of coastline because of its irregular shape; the length is greater than that of the Indian and Pacific oceans combined. Another distinctive feature of the Atlantic Ocean is that the majority of major continental rivers discharge into it. Because navigation has played an important part in the history of civilization, the presence of irregular coastlines with many marginal seas, shielded bays, and river mouths, which provide excellent natural harbours, has contributed to the development and spread of culture. The lands bordering the Mediterranean Sea and western Europe and the coastal regions of India, China, and Japan have been so favoured.

Depths of water

Table 1 shows the horizontal dimensions of the oceans and seas and, in addition, gives their average depths and volumes.

The average depth of all the seas has been estimated at 3,790 metres (12,430 feet), a figure considerably larger than that of the average elevation of the land above the sea level, which is 840 metres (2,760 feet). If the average depth is multiplied by its respective surface area, the volume of the World Ocean is 11 times the volume of the



1. Baltic Sea
a. Gulf of Bothnia
b. Gulf of Finland
c. Gulf of Riga
2. Kattegat, Sound and Belts
3. Skagerrak
4. North Sea
5. Greenland Sea
6. Norwegian Sea
7. Barents Sea
8. White Sea
9. Kara Sea
10. Laptev Sea
11. East Siberian Sea
12. Chukchi Sea
13. Beaufort Sea
14. Northwestern Passages
a. Baffin Bay
15. Davis Strait
a. Labrador Sea
16. Hudson Bay
a. Hudson Strait
17. Arctic Ocean
a. Lincoln Sea
18. Inner Seas (off the west coast of Scotland)
19. Irish Sea and St. George's Channel
20. Bristol Channel
21. English Channel
22. Bay of Biscay
23. North Atlantic Ocean
24. Gulf of St. Lawrence
25. Bay of Fundy
26. Gulf of Mexico
27. Caribbean Sea
28W. Mediterranean Sea (Western Basin)
28E. Mediterranean Sea (Eastern Basin)
a. Strait of Gibraltar
b. Alboran Sea
c. Balearic Sea
d. Ligurian Sea
e. Tyrrhenian Sea
f. Ionian Sea
g. Adriatic Sea
h. Aegean Sea
29. Sea of Marmara
30. Black Sea
31. Sea of Azov
32. South Atlantic Ocean
33. Río de la Plata
34. Gulf of Guinea
35. Gulf of Suez
36. Gulf of Aqaba
37. Red Sea
38. Gulf of Aden
39. Arabian Sea
40. Gulf of Oman
41. Persian Gulf
42. Laccadive Sea
43. Bay of Bengal
44. Andaman Sea
45. Indian Ocean
a. Mozambique Channel
46. Malacca and Singapore straits
a. Strait of Malacca
b. Singapore Strait
47. Gulf of Thailand
48. East Indian Archipelago (Indonesia)
a. Sulu Sea
b. Celebes Sea
c. Molucca Sea
d. Gulf of Tomini
e. Halmahera Sea
f. Ceram Sea
g. Banda Sea
h. Arafura Sea
i. Timor Sea
j. Flores Sea
k. Gulf of Boni
l. Bali Sea
m. Makassar Strait
n. Java Sea
o. Savu Sea
49. South China Sea
50. East China Sea
51. Yellow Sea
52. Sea of Japan
53. Inland Sea
54. Sea of Okhotsk
55. Bering Sea
56. Philippine Sea
57. North Pacific Ocean
58. Gulf of Alaska
59. Coastal waters of Southeast Alaska and British Columbia
60. Gulf of California
61. South Pacific Ocean
62. Great Australian Bight
a. Bass Strait
63. Tasman Sea
64. Coral Sea
65. Solomon Sea
66. Bismarck Sea

Figure 1: Boundaries of the world's oceans and seas.
By courtesy of the International Hydrographic Organization

land above sea level. The maximum depth of the ocean, 10,850 metres (35,597 feet), occurs in the Mariana Trench, halfway between the islands of Guam and Yap in the Pacific Ocean. This depth exceeds the height of Mt. Everest, which is 8,848 metres (29,028 feet).

Table 2 shows the areas that are occupied by various ocean depth zones in relation to total area of the sea surface, as a percentage. Because the 0–200-metre (660-foot) depth zone corresponds to the continental shelf, it covers almost as large an area as the zone with depths of 200–2,000 metres, which is the region of occurrence of the far steeper continental slope; on the other hand, depths of more than 6,000 metres cover only a very small part of the ocean bottom, in contrast to 3,000–6,000-metre depths. The proportions of the various depth zones are shown in Figure 2, along with the names given to the various depth regimes that constitute marine environments.

The relatively shallow, submerged platform bordering the continents, called the continental shelf, slopes gently seaward to the shelf break, where an increase in gradient leads to the continental slope. Conventionally, the edge of the shelf has been placed at the 100-fathom (180-metre; about 600-foot) depth line, although the water depth at the shelf break is more nearly 85 fathoms. The width of the continental shelf varies enormously, from nearly zero along parts of the west coast of North and South America to more than 1,000 kilometres (620 miles) off the north coast of Siberia. The average width is 75 kilometres (47 miles), and the average slope is 1.7 metres per kilometre (0.1°). The continental slope extends downward to a depth of about 4,000 metres (13,000 feet). Its average slope near the shelf is some 70 metres per kilometre (4°) over a width of 20–100 kilometres, and farther out to sea it gradually becomes gentler. The third zone, which may vary in width from 0 to 600 kilometres (400 miles), is called the continental rise. It merges with the deep-sea abyssal plain at an average depth of about 4,000 metres (13,100 feet).

Features of the sea floor

An extraordinary feature of the continental slope is that at many places it is intersected by chasms with steep irregular sides, called submarine canyons, which extend from the continental margin to the ocean floors. Certain other sea-bottom forms shown by bathymetric (depth) charts of the oceans are referred to in established terms.

A rise is a long, broad elevation coming gently and smoothly from the sea floor. A ridge is a long, narrow elevation of the sea floor with steep sides and topography more irregular than that of a rise. Sometimes, the highest parts of a ridge project as islands above the sea, as is the case with the Azores, on the Mid-Atlantic Ridge. A sill is a ridge or rise separating a partially closed basin, trough, or trench from another basin or from the adjacent sea floor. The greatest water depth over a sill is called the sill depth. A plateau is the upper surface of a comparatively flat-topped, extensive elevation of the sea floor, normally rising more than 100 fathoms on all sides. A bank is an elevation of the sea floor that is located on a continental shelf or an island shelf and over which the

Table 1: Surface Area, Volume, and Average Depth of Oceans and Seas

	area		volume		average depth	
	000,000 sq km	000,000 sq mi	000,000 cu km	000,000 cu mi	m	ft
Atlantic Ocean						
without marginal seas	82.440	31.830	324.600	77.900	3,930	12,890
with marginal seas	106.460	41.100	354.700	85.200	3,330	10,922
Pacific Ocean						
without marginal seas	165.250	63.800	707.600	169.900	4,280	14,038
with marginal seas	179.680	69.370	723.700	173.700	4,030	13,218
Indian Ocean						
without marginal seas	73.440	28.360	291.000	69.900	3,960	10,037
with marginal seas	74.920	28.930	291.900	70.100	3,900	12,792
Arctic Ocean	14.090	5.440	17.000	4.100	1,205	3,952
Mediterranean Sea and Black Sea	2.970	1.150	4.200	1.000	1,430	4,690
Gulf of Mexico and Caribbean Sea	4.320	1.670	9.600	2.300	2,220	7,282
Australasian Central Sea	8.140	3.140	9.900	2.400	1,210	3,969
Hudson Bay	1.230	0.470	0.160	0.040	128	420
Baltic Sea	0.420	0.160	0.020	0.005	55	180
North Sea	0.570	0.220	0.050	0.010	94	308
English Channel	0.075	0.029	0.004	0.001	54	177
Irish Sea	0.100	0.040	0.006	0.001	60	197
Sea of Okhotsk	1.530	0.590	1.300	0.300	838	2,749
Bering Sea	2.270	0.880	3.300	0.800	1,440	4,720
The World Ocean	361.100	139.400	1,370	329	3,790	12,430

depth of water is relatively shallow but sufficient for surface navigation. A basin is a large depression of the sea floor that is more or less equidimensional in form. When the length of a depression is considerably greater than the width, and the slope of the sides is fairly gentle, the feature is called a trough. A trench is a long, narrow depression of the deep-sea floor having relatively steep sides and generally greater depths than those occurring in troughs.

Table 2: Depth Zones of the World Ocean and Their Extent

depth of zones		extent of zones		percentage of total surface of the sea
m	ft	000,000 sq km	000,000 sq mi	
0–200	660	2.74	1.06	7.6
200–1,000	660–3,300	15.5	6.0	4.3
1,000–2,000	3,300–6,600	15.2	5.9	4.2
2,000–3,000	6,600–9,800	24.5	9.5	6.8
3,000–4,000	9,800–13,100	70.8	27.3	19.6
4,000–5,000	13,100–16,400	119.1	46.0	33.0
5,000–6,000	16,400–19,700	84.1	32.5	23.5
6,000–7,000	19,700–23,000	4.0	1.5	1.1
over 7,000	23,000	0.4	0.2	0.1

Dealt with in this article are the physical and chemical properties of seawater, the general interaction of the atmosphere and the oceans that produces oceanic circulation and other water motions, and the life and economic potential of oceans and seas. For further information on the configuration of the ocean basins and their development through time, see OCEAN BASINS; CONTINENTAL SHELF AND SLOPE; OCEANIC RIDGES; and SEA-FLOOR SPREADING. See also MARINE SEDIMENTS; OCEANS, DEVELOPMENT OF for discussions of oceanic deposits and the evolution of the oceans, respectively, and WATER WAVES; TIDES; and OCEAN CURRENTS for additional details on water motions. The nature, scope, and methods of oceanography are found in HYDROLOGIC SCIENCES; methods of bathymetry are in HYDROGRAPHIC CHARTING.

This article is divided into the following sections:

I. Physical and chemical properties

COMPOSITION OF SEAWATER

The constituents of seawater include dissolved inorganic substances (such as salts), dissolved gases, and dissolved organic substances. Apart from these dissolved substances, seawater contains widely varying concentrations of particulate matter in suspension. This particulate matter, nonliving and living (such as plankton, floating forms of marine organisms), can influence certain properties of seawater, but the discussion at this point will be confined primarily to inorganic components. See below *Life in the open sea* for a discussion of the role of organisms; and see below *Dissolved organic substances* for a discussion of organic constituents.

Dissolved inorganic substances. Table 3 shows the principal inorganic constituents of seawater other than gases. Excluding water, the constituents are listed not as chemical compounds but rather as ions, the electrically charged parts of compound molecules that dissociate because of the electrolytic action of the water in which they are dissolved. The last column, in addition, gives the percentage occurrence of constituents relative to their total mass.

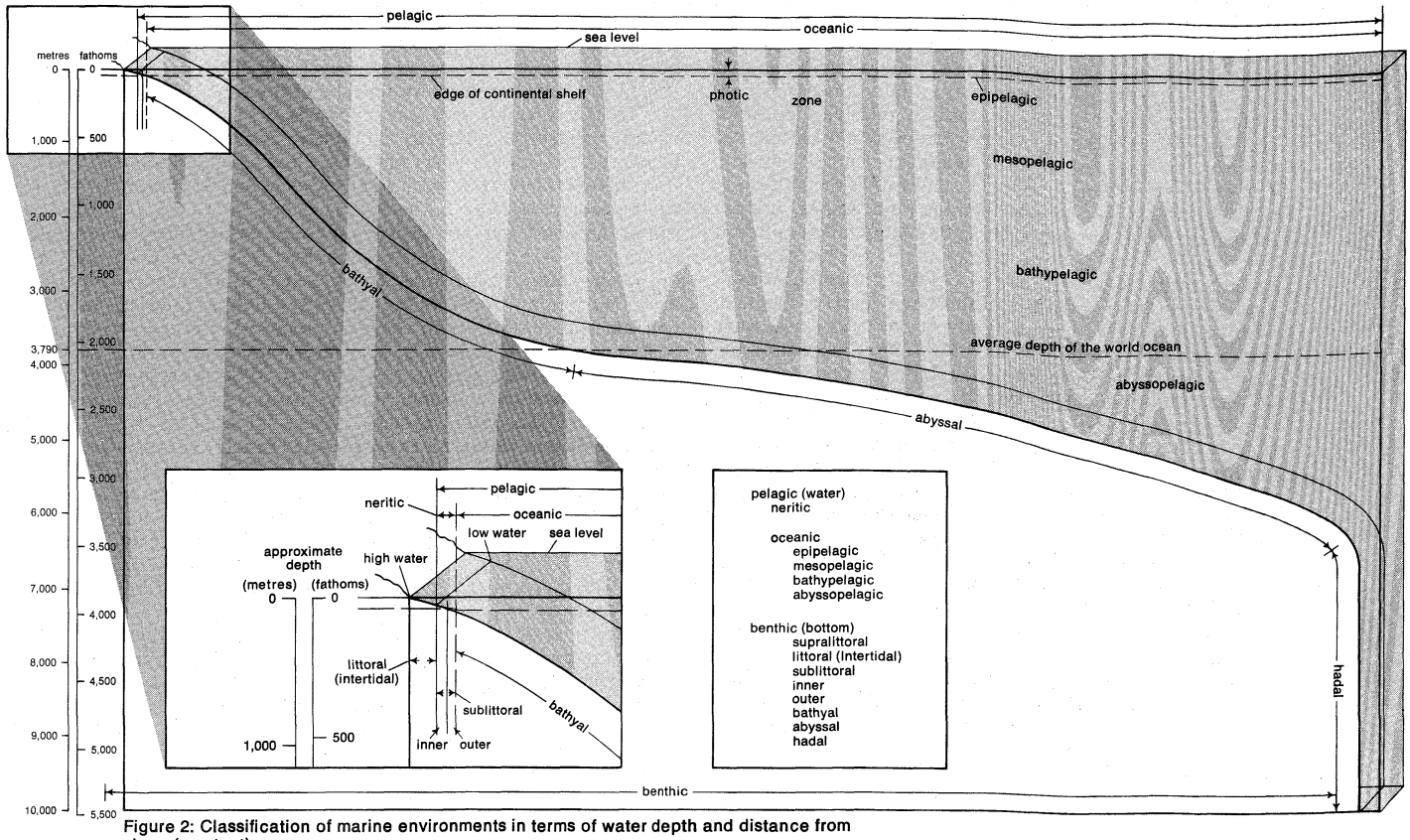


Figure 2: Classification of marine environments in terms of water depth and distance from shore (see text).
From J.W. Hedgpeth (ed.), *Treatise on Marine Ecology and Paleoecology* (1967); The Geological Society of America

Sodium and chloride ions predominate in seawater; together they form more than 85 percent by weight of the total amount of dissolved salts. The last column of the table indicates that, although the total amount of these dissolved substances may vary from place to place and from time to time, the relative proportions of the components are remarkably constant. The total salt content may vary, because seawater can be diluted by additions of precipitation in the form of rain or snow, fresh river water, or meltwater from icebergs. At other times or places, seawater may become more saline because of evaporation of water. In both cases, the proportions of the salt constituents remain the same. The middle column of Table 3 shows the amounts of the various components relative to the amount of seawater, as parts per thousand by weight, or pro mille (grams per kilogram of seawater). These figures are only illustrative, because the total salt content may vary.

Table 3: Principal Constituents of Seawater
(of 19 parts per thousand chlorinity)

constituent	g/kg of seawater	proportion of total salt content (percent)
Chloride (Cl^-)	18.980	55.044
Sulfate (SO_4^{2-})	2.649	7.682
Bicarbonate (HCO_3^-)	0.140	0.406
Bromide (Br^-)	0.065	0.189
Fluoride (F^-)	0.001	0.003
Boric Acid (H_3BO_3)	0.026	0.075
Sodium (Na^+)	10.556	30.613
Magnesium (Mg^{2+})	1.272	3,689
Calcium (Ca^{2+})	0.400	1.160
Potassium (K^+)	0.380	1.102
Strontium (Sr^{2+})	0.013	0.038
Total	34.482	100.000
Water (with traces of other substances)	965.518	
Total	1,000.000	

Because the total volume of the oceans and seas is 1,370,000,000 cubic kilometres (329,000,000 cubic miles), and the average total concentration of salt in the seawater amounts to about 3.6 kilograms per litre, the total amount of salt in the sea may be computed to be roughly 5×10^{18} kilograms, or 5,000,000,000,000 tons. Dried and spread over the whole Earth, this would produce a layer 45 metres (150 feet) thick.

The list of components given here is not complete. Although they account for virtually the entire weight of the dissolved substances, there are other trace elements (minute quantities of matter) in seawater that may be of essential importance to the economy of the sea and to life for certain organisms. Chemical analyses of seawater have revealed traces of iodine, which is an important constituent of some seaweeds, and of copper, which occurs in the blood of crabs. Several chemical elements were found in marine organisms before their identification in seawater. Because rivers continuously carry products of the Earth's crust from the land to the sea, the sea should contain all the elements that are found on the lands. Radioactive matter also occurs naturally in the sea, albeit in exceedingly minute concentrations. Uranium, in concentrations of about three milligrams per cubic metre, and radium, in concentrations of 0.03–0.15 milligram per 1,000,000 cubic metres, generally increase in concentration with water depth. In addition, there are such isotopic constituents as the radioactive isotope of potassium (potassium-40), which has an abundance of 0.0118 percent of the total potassium content and which constitutes by far the largest single source of radioactivity in the ocean. The radioactive isotopes of carbon (carbon-14) and hydrogen (tritium, hydrogen-3) play an important role in estimating the "age" of deep-water masses in the oceans (*i.e.*, the time elapsed since these waters were in contact with the atmosphere).

Salinity and chlorinity of seawater

Salinity is used by oceanographers to characterize the total salt content of a sample of seawater. It is defined as the total amount of solid material (in grams) that is in solution in one kilogram of seawater, when the bromine and iodine content have been replaced theoretically

by an equivalent amount of chlorine, when all the carbonate has been converted to oxide, and when all the organic matter has been completely oxidized. This definition arises partly from certain considerations in analytical chemistry. With respect to the major dissolved constituents, seawater has a composition relatively constant enough so that any single major element can be selected for use as a measure of other elements and of salinity. The standard method of analysis determines the chlorinity of the sample; in effect, bromine and iodine present in the seawater are replaced by an equivalent amount of chlorine, because those two elements are precipitated by silver nitrate as well as the chlorine. The chlorinity, expressed in grams per kilogram of seawater, is identical with the number of grams of silver that is just necessary to precipitate the halogens in 0.3285233 kilogram of the seawater sample. The empirical relation between salinity (S) and chlorinity (Cl), which has been used since the standardization of the method of chlorinity determination, is that salinity equals 0.03 more than 1.805 times the chlorinity, written $S = 1.805\ \mathrm{Cl} + 0.03$. Most salinity determinations currently are made from measurements of electrical conductivity. The conductivity–chlorinity relationship has been internationally established on the basis of analyses of samples from all the seas of the world, and the chlorinity–salinity relation adopted is written $S = 1.80655\ \mathrm{Cl}$.

Both salinity and chlorinity are written as pro mille; that is, as parts per thousand (‰, or grams per kilogram).

In the vast majority of places in the oceans, the salinity lies between 34 and 37 parts per thousand. The horizontal and vertical distribution of salinity will be discussed later (see below *Physical properties of seawater: Salinity distribution*).

Dissolved gases

Because the sea is continually in contact with the atmosphere, the gases that occur in the atmosphere are also found in seawater, in concentrations depending upon their solubilities and on the chemical and biochemical reactions in which they are involved.

The solubilities of nitrogen (N_2), oxygen (O_2), and carbon dioxide (CO_2) differ greatly. They are present in air in proportions of 78 percent, 21 percent, and 0.03 percent by volume, respectively, but their saturation concentrations in seawater of 19 parts per thousand chlorinity at a temperature of 12° C (54° F), in contact with air of one atmosphere pressure, are 11.1, 6.2, and 0.3 millilitres of gas per litre of seawater. The solubility of oxygen considerably exceeds that of nitrogen, and carbon dioxide is much more soluble than either of these gases. The value given for carbon dioxide here includes carbonic acid, the compound formed by reaction of carbon dioxide with water. The solubility of gases decreases with increasing temperature.

Nitrogen is of little importance to life in seawater, with the exception of certain bacteria living on or near the bottom that manufacture ammonium salts and nitrates from nitrogen.

Oxygen is derived from the atmosphere and from marine plants that use sunlight to manufacture carbohydrates from water and carbon dioxide by photosynthesis, by which process oxygen is released. Both sources exist near the surface of the sea; at greater depth it is too dark for photosynthesis by green plants to take place. On the other hand, oxygen is consumed everywhere, even at the greatest depths, because organisms that use oxygen live there. Moreover, oxygen is consumed by combination with organic waste products when the dead remains of organisms sink to the bottom and decompose. Hence, near the surface there is overproduction of oxygen, and at greater depths there is overconsumption. In between there is a level at which production and consumption are just balanced (called the compensation depth), which may vary between one metre and 100 metres (about three to 330 feet), depending on the amount of sunlight available at the sea surface, the transparency of the water, and the abundance of plant life—which in its turn depends on available nutrients. Because oxygen is subject to consumption at all depths, it might be thought that no oxygen exists at great depths, where the production of oxygen is essentially zero, but water refreshment is brought about by the slow, continual, large-scale, water circulation within the World Ocean. At high latitudes the very heavy cold-water masses sink toward the bottom and spread to lower latitudes through the entire deep-sea layers. In the area between the upper layers of the sea, where oxygen is naturally abundant, and the deep layers, where oxygen is supplied by the oceanic circulation, a layer having a minimum oxygen concentration often occurs.

Oxygen-deficient environments

In some seas that are nearly enclosed, ventilation is lacking, and oxygen is completely absent in the stagnating deep water. Part of the Black Sea bottom waters, some Norwegian fjords with high sills (shallow inlets), and the Kau Bay of Halmahera (Indonesia) are examples. In such cases, there is a lack of horizontal communication with outside deep waters, but, in addition, stable water stratification prevents mixing of the upper and lower layers. Anaerobic bacteria (those in reducing, or oxygen-deficient, environments) produce much hydrogen sulfide (H_2S) in such water.

Apart from the fact that carbon dioxide (CO_2) is highly soluble in seawater and combines with water to give carbonic acid, the latter dissociates partly to give bicarbonate (HCO_3^-) and carbonate (CO_3^{2-}) ions, and it also reacts with calcium and magnesium to form calcium and magnesium carbonate and bicarbonate. All these reactions are reversible and may proceed in either direction, depending on the amount of carbon dioxide that is available. Because of this reversibility, the sea has a large buffering capacity (storing and regulating capacity) with respect to processes involving carbon dioxide in the atmosphere and in the sea, including those pertinent to plant life and photosynthesis. The whole system of reactions in which carbon dioxide is involved, however, is extremely complex.

Dissolved organic substances. Seawater contains a diversity of dissolved organic compounds that originate from the decomposition of organisms after their death. The amount of total dissolved organic carbon in the open ocean varies from 0.2 to 2.5 milligrams per litre. Higher values are found in landlocked areas, such as the Black Sea and the Baltic; the shallow water of the Dutch Waddenzee may contain as much as eight milligrams per litre. The highest organic-carbon concentrations are in waters of high phytoplankton (floating forms of small marine plants) productivity. In deeper water, the products of decomposition of organisms that sink toward the bottom tend to be released in the lower layers, beneath the zone where they might serve as nutrients for phytoplankton. From those depths, they may again be brought up by the water-circulation systems and by upwelling along certain coastal regions.

CHEMICAL EVOLUTION

The early hydrosphere may have formed in part by condensation from the early atmosphere, and speculations about the early history of seawater therefore centre on the history of the early atmosphere. Only a very small fraction (less than 1 percent) of the water now present in the oceans may have been derived from the Earth's crust by rock decomposition or weathering; the major part is assumed to have had its source in the release of gases (degassing) from the Earth's core and mantle. Volcanic vents or hot springs often serve as channels for this degassing. A number of other volatile constituents, such as carbon (including carbon dioxide), chlorine, sulfur, and nitrogen, are thought to have arisen from this source because of their abundances. Early accumulations by condensation probably took place in a number of isolated, closed basins. Because of the large amount of carbon dioxide present in the atmosphere at that time, these original oceans, or proto-oceans, must have been quite acid. Exposed rocks would be subject to rapid weathering, and large amounts of dissolved material would be carried into the basins in a short time. Negatively charged ions in the early ocean were probably carbonate and bicarbonate ions predominantly, rather than chloride ions. Sedimentary deposits that are 1,800,000,000 years old suggest that

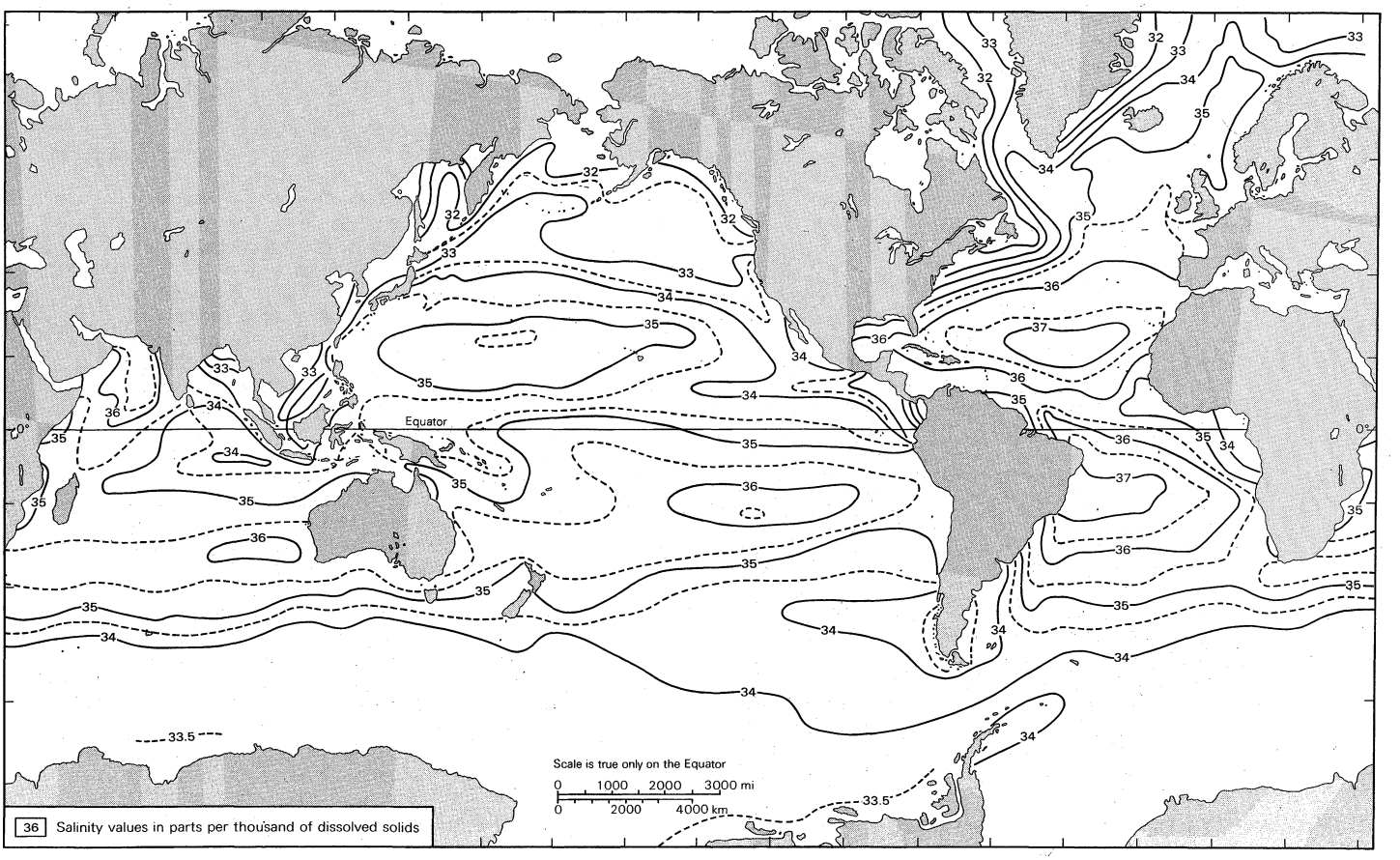


Figure 3: Salinity distribution in surface waters of the World Ocean.
From H.U. Sverdrup, Martin W. Johnson, and Richard H. Fleming, *The Oceans: Their Physics, Chemistry, and General Biology*, © 1942, renewed 1970; Prentice-Hall, Inc., Englewood Cliffs, New Jersey

Characteristics of the initial ocean

the oxygen content of the atmosphere at that time was still very low. Photodissociation (breakdown involving the energy of light) of water (H_2O) into hydrogen (H_2) and oxygen (O_2) was responsible for an increase of oxygen; hydrogen escaped from the gravitational field, but much of the oxygen initially was consumed in the oxidation of such reducing constituents of the primeval atmosphere as methane (CH_4), ammonia (NH_3), graphite, and iron. In the time interval between 2,500,000,000 and 1,000,000,000 years ago, however, the atmospheric oxygen pressure must have risen sufficiently to permit the beginning of life. As biological activity expanded, carbon dioxide was consumed, and oxygen was produced much more rapidly. Throughout this time, condensation of water continued, and the proto-oceans merged to become larger oceans. As the carbon dioxide content of air and seawater decreased, saturation of the seawater with respect to calcium carbonate ($CaCO_3$) required much lower concentrations of calcium than were originally present, and large-scale precipitation of limestones (rocks consisting principally of calcium carbonate) was initiated, aided by biological processes. From that time on, a considerable portion of the Precambrian sediments (older than 570,000,000 years) consists of carbonate rocks. Partly as a consequence of this development, chlorine became the dominant negative ion, and the composition of seawater probably had attained a composition very near that of today's oceans and seas. For further details of the several lines of evidence and of the chemical models involved, see ATMOSPHERE, DEVELOPMENT OF; and OCEANS, DEVELOPMENT OF.

PHYSICAL PROPERTIES OF SEAWATER

Salinity distribution. In the example given in Table 3, the calculated salinity is 34.48 parts per thousand. The salinity generally varies in open seas within rather narrow limits; it is between 34 and 37 parts per thousand in most places (Figure 3). It is lowest wherever there is much rainfall or where there is an influx of freshwater from many large rivers that empty into the sea. In regions receiving much freshwater in this way, particularly if they are somewhat divorced from the open ocean, salinities may be less than 34 parts per thousand as in coastal lagoons and shallow isolated areas, such as the Waddenzee behind the Frisian Islands. The salinity also is less than 34 parts per thousand in the vicinity of Newfoundland, not only because freshwater is discharged to the sea by the St. Lawrence River and the rivers of Labrador but also because pack ice and icebergs float down from the north and release their meltwater. Meltwater generally contains less salt than ocean water, and in the case of icebergs it contains no salt at all. Another example is the Baltic, where the salinity is as low as 10 parts per thousand or less in many places; in the Gulf of Bothnia and the Gulf of Finland, it is as low as 5 parts per thousand. This low salinity is caused by the great freshwater discharge of many rivers in the area and by the fact that saline ocean water can enter only in very restricted measure through the narrows connecting the Baltic with the Atlantic Ocean. The surface salinity of the Arctic Ocean is also comparatively low, especially opposite the coast of northern Siberia, where many great rivers have their outlets. Because this freshwater is less dense, it remains on the surface; the salinity at greater depth in the Arctic Ocean is normal for seawater.

In the open oceans, the general salinity distribution is governed by the exchange of water with the atmosphere, that is, by the difference between precipitation and evaporation. Lower salinities occur in moderate and high latitudes and in the equatorial zone, where precipitation exceeds evaporation, and higher salinities prevail in the subtropical zones, where evaporation dominates. This general trend is depicted in Figure 4. In the Sargasso Sea, for instance, in the middle of the Atlantic Ocean at about latitude 25° N, the salinity in summer is more than 37 parts per thousand. Salinity is still higher in the Mediterranean (38 parts per thousand) and the Red Sea (41 parts per thousand). These high salinities are the result of

the small amounts of freshwater received by such seas and the high prevailing evaporation rates. Furthermore, because the water bodies are almost enclosed, they have poor communication with the open ocean. Communication of the Mediterranean Sea with the Atlantic Ocean, for example, is maintained through the narrow Strait of Gibraltar. There, water flows in from the Atlantic Ocean, notably in the upper layers, and at greater depths a countercurrent carries more saline water of the Mediterranean Sea over the sill of the strait into the Atlantic (see also DENSITY CURRENTS).

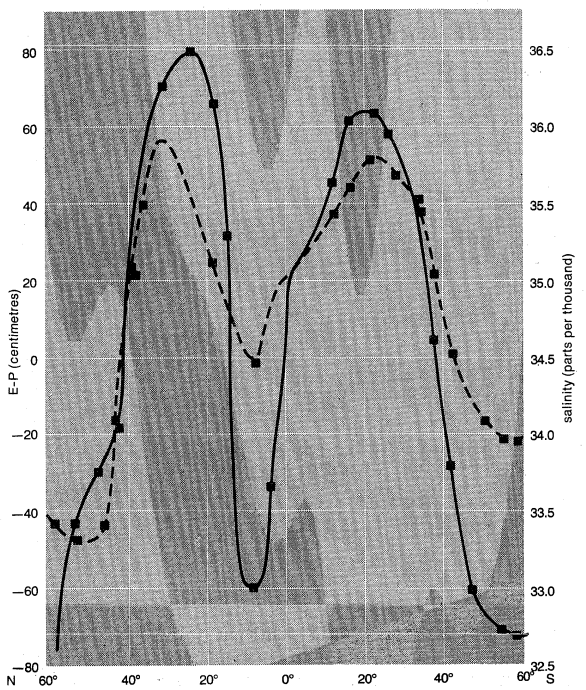


Figure 4: Mean meridional distribution of evaporation precipitation (E-P; solid line) and surface salinity (dashed line) for the entire ocean.

Differences in salinity in the deep sea are smaller than those at the surface, ranging from 34.5 to 35.0 parts per thousand. Very exceptional conditions have been found in a few places on the floor of the Red Sea, where salinities in small bottom depressions are as high as 256 parts per thousand. These values are accompanied by temperatures as high as 60°C (140° F). It is believed that mineral salts have been extracted directly from the underlying crust in these restricted pockets.

Surface temperatures

Temperature distribution. The temperature of the surface waters of the oceans and seas varies greatly in different parts of the world. It may be −1.9° C (29° F) in the polar seas and may rise to as much as 30° C (86° F) in such subtropical waters as the South China Sea and the Gulf of Mexico; temperatures may be higher still in such restricted marginal seas as the Persian Gulf, where 33° C (91° F) is not exceptional. Figure 5 shows mean annual temperatures of the ocean surface waters for the several latitudes, averaged along circles of latitude, together with the approximate annual temperature range. For the sea, the coldest and warmest months are usually February and August in the Northern Hemisphere, outside the equatorial belt, and the reverse in the Southern Hemisphere. In tropical seas and in the polar regions, the annual range is small, about 1° to 2° C (1.8 to 3.6° F). It is largest, as far as the open ocean is concerned, between latitudes 40° and 45° N in the North Atlantic Ocean, where it is roughly 8° to 9° C (about 14° to 16° F), and in the North Pacific, where it is about 9° to 10° C (16° to 18° F). In these regions, especially on the western sides of oceans, the annual range is increased by cold offshore winds in winter that lower the winter temperatures of the sea. There is scarcely anything of this kind in the Southern Hemisphere, where the largest annual temperature range, roughly 5° to 6° C (about 9° to 11° F), occurs between latitudes 30° and 40° S. In shallow marginal seas, the annual range will generally be greater than it is in the open ocean because of the influence of neighbouring land areas.

With increasing depth beneath the surface, the annual temperature variations become smaller. In general, they may be perceptible down to 300 metres (1,000 feet), but often these temperature variations extend downward no farther than about 100 metres (330 feet).

There are certain zones of significant temperature changes over a short distance in the general pattern of horizontal temperature distribution at the ocean surface, giving the appearance of a front. Such is the case on the left-hand side of the Gulf Stream, east of the northern United States and Newfoundland, where cold waters from the north meet the warm Gulf Stream waters and partly dive under it. The boundary here is called the Arctic Polar Front, or Arctic Convergence. In the Pacific, a similar phenomenon occurs to the northeast of Japan, where the cold waters of the Oyashio meet the warm Kuroshio, or Japan Current. In the Southern Hemisphere, there is the Antarctic Convergence, winding as a closed line around the globe between latitudes 50° and 60° S, along which cool waters from higher latitudes meet warmer waters from mid-latitudes; this meeting causes the surface temperature to jump two or three degrees within a short distance.

Subsurface temperatures

Temperatures generally decrease with increasing depth, except in the polar water masses, where temperatures are low from the surface to the bottom. The vertical temperature decrease often shows a jump, above which the water often is more or less isothermal; that is, it has the same temperature at different depths. This phenomenon may be caused by wind mixing or, in the cold season, by cooling from above, which induces vertical circulation (convection), or by both. These factors tend to make the mixed layer attain its greatest depth (to 100 metres or 330 feet) in winter and in spring. The jump also may be absent, particularly after a period of heating from above, or may be broken into sublayers. The drop in temperature found beneath the mixed layer is called the thermocline.

Beneath this thermocline, the decrease in temperature is more gradual, down to very low temperatures; even in tropical regions, temperatures of less than 1° C (34° F) have been found at depths of 5,000 metres (16,400 feet) or more, and the temperature already is 3.5° C (38.3° F) or less in most places at depths of 2,000 metres (6,600 feet). The cause of these low temperatures is the slow

From "Treatise on Marine Ecology and Paleoecology," vol. 1 (1967)

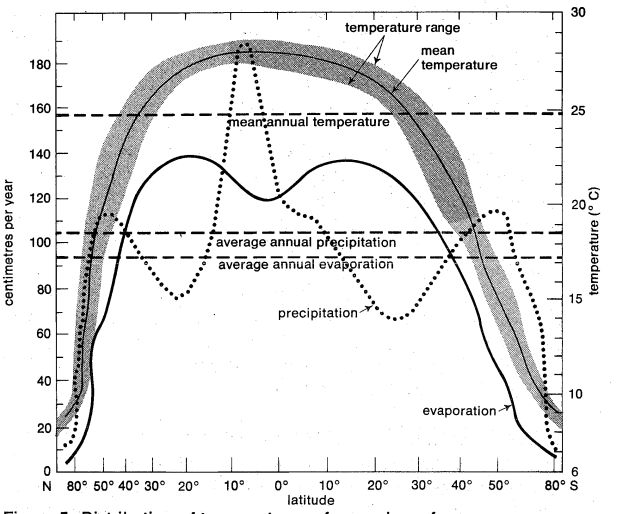


Figure 5: Distribution of temperatures of oceanic surface waters and variation of water gains (precipitation) and losses (evaporation) with respect to latitude. The latitude scale has been adjusted for ocean area to reflect the greater water expanse in the Southern Hemisphere.

deep-sea circulation, referred to earlier in connection with the oxygen content of these deep waters. These cold-water masses are of polar or subpolar origin, and the very coldest have come from the Antarctic. In order to have a true circulation, the water must, of course, eventually come up again from these great depths. This rise does occur, but it takes place in a diffuse manner and is extremely slow everywhere in the World Ocean. This deep-sea circulation, fed by cold polar and subpolar water masses, has its place mainly within the deep part of the body of the World Ocean. It is the colder part, the other part being the warm and moderately warm upper waters of tropical and subtropical regions. These warm waters form a layer several hundred metres thick, situated roughly between latitudes 45° S and 45° to 55° N; the boundary at the surface is not sharply defined, certainly not in the eastern part of the North Atlantic Ocean, which receives the Gulf Stream waters. The lower boundary of the warm layer also is not well defined. It is usually placed at the isothermal level of approximately 10° C (50° F). The two layers may be called the thermosphere (warm sphere) and the cryosphere (cold sphere), but the important point is that the thermosphere has a circulation of its own; the two circulations are associated, however, and water may pass sometimes from one layer into the other.

There are places in the deep sea where at some depth the temperature stops falling and then rises slightly at greater depth. Such temperature inversions occur in certain deep basins and troughs, particularly in the Moluccan seas, where the bottom water enters across a shallow sill and then flows downward for a considerable vertical distance. In flowing down, it is subjected to compression, which increases with depth, and this causes adiabatic heating—heating without any heat exchange with surrounding water.

Thermal properties. The adiabatic temperature change just mentioned is a consequence of the fact that water is compressible. An increase in pressure causes the volume to decrease slightly so that work is performed on the water, and, if this takes place adiabatically, then the energy gained causes a rise in temperature. For example, if seawater of 35 parts per thousand salinity and a temperature of 18° C (64° F) is lowered adiabatically from the surface to 1,000 metres depth (3,300 feet), the temperature increases by 0.18° C (0.32° F); for seawater of 2° C (36° F), the rise in temperature would be 0.06° C (0.1° F); taken down to 10,000 metres depth, the water would undergo a temperature increase of 1.3° C (2.3° F). These effects depend on the compressibility and the thermal-expansion coefficient of seawater, which have different values at different temperatures and pressures. They also depend, of course, on the specific heat (heat capacity) of seawater. If water is raised from a lower to higher level, the adiabatic temperature change has the same numerical value but is negative, which means a temperature decrease. In this connection, the concept of potential temperature must be mentioned. This is the calculated temperature that a quantity of water would assume if it were brought to a certain standard level of pressure adiabatically—for instance, to the sea surface. If water is actually lowered or raised, the actual temperature will change, but the potential temperature will not. The latter, called a conservative property, can therefore be used to trace water masses in the deep sea.

The specific heat of seawater is the amount of heat required to raise the temperature of one kilogram (2.2 pounds) of seawater by 1° C (1.8° F) under constant pressure or with constant volume. It is somewhat smaller in value than the specific heat of pure water at the same temperature. At a temperature of 17.5° C (63.5° F), it is 1,000 calories per kilogram for pure water (under constant pressure of one atmosphere), whereas it is 932 calories per kilogram for seawater of 35 parts per thousand salinity. Specific heat decreases slightly with increasing temperature and with increasing pressure.

The freezing point of saltwater is lower than that of freshwater. For various salinity values the freezing point is as follows:

salinity (parts per thousand)	0	10	20	30	35
freezing point	0.00	−0.53	−1.08	−1.63	−1.91° C

Water density. The density of seawater is the mass per unit volume, expressed in kilograms per cubic metre. It increases with increasing salinity and pressure and decreases with increasing temperature.

The dependency on salinity is exemplified by the following figures: at a temperature of 0° C (32° F), under atmospheric pressure, pure water has a density of 999.9 kilograms per cubic metre. Seawater of 20 parts per thousand and 35 parts per thousand salinity has a density of 1,016.1 and 1,028.1 kilograms per cubic metre, respectively. The effect of pressure becomes apparent from the following figures for seawater of 0° C and 35 parts per thousand salinity at various depths:

depth (metres)	0	1,000	2,000	10,000
density (kg/m^3)	1,028	1,033	1,037.5	1,071

The dependence of density on temperature deserves some special attention. Unlike freshwater, seawater attains its maximum density not at 4° C (39° F) but at a lower temperature. The higher its salinity, the lower the temperature at which its density is a maximum; and, if its salinity is 24.7 pro mille or more, seawater continues getting heavier with decreasing temperature until the freezing point is reached. This fact is of special importance in connection with the freezing over of seawater. If the sea is cooled from above, and the surface water thus becomes colder than the water underneath, it will sink downward, and water from below will take its place. For this reason, the surface water cannot drop to freezing temperature as long as the water beneath is not almost equally cold. This phenomenon explains why, apart from the lower freezing temperature, the sea does not freeze over as readily as freshwater, which remains at the surface when cooled below 4°C. The fact that the coldest seawater is the heaviest, provided its salinity is not less than 24.7 parts per thousand, also has an important bearing on global deep-sea circulation, which is driven by the sinking of ice-cold polar water masses.

The figures below give some values of the density of seawater of 35 parts per thousand salinity, for various temperatures, at atmospheric pressure (in oceanographic literature it is customary to use a quantity called sigma [σ], which is equal to density in kilograms per cubic metre minus 1,000; for the density as a function of temperature at atmospheric pressure, the symbol sigma-t [σ_t] is in use):

temperature	0	10	20	25	30° C
density (kg/m^3)	1,028.1	1,027.0	1,024.8	1,023.4	1,021.75
σ_t	28.1	27.0	24.8	23.4	21.75

Pressure. The pressure on water in the sea is expressed in a unit called the bar, which is equal to 10^5 newton per square metre, the newton being the physical unit of force according to the international system of units. The practical unit is the decibar, which equals 0.1 bar. One decibar corresponds approximately to the pressure of one metre of seawater of normal salinity. The exact pressure exerted by a column of seawater depends on its density. If it is in equilibrium with the force of gravity, the pressure difference between the top and the bottom of a column of seawater is equal to density times acceleration of gravity times the height of the column. Because density varies with salinity, temperature, and pressure, an exact computation of pressures in the sea requires the summation of partial pressures corresponding to portions of the whole water column involved, for which the values of salinity and temperature at various depths are known from observation. Calculation of pressures in the sea is of special importance with respect to the dynamics of ocean currents.

Optical properties. The interaction of light with seawater has two principal effects, namely, absorption and scattering. Absorption is defined as the conversion of

Absorption and scattering of light

radiant energy to other forms of energy (mostly heat) and scattering as the irregular deviation of light from straight-line propagation. A parallel beam of light that is propagated through seawater suffers attenuation by the combined action of these two processes; the relative loss of light intensity per metre of path length is called attenuance, which is absorptance plus scatterance. Sometimes the term absorption coefficient or extinction coefficient is used.

The absorption and scattering of light in seawater are caused by four constituents, namely, water, dissolved salts, dissolved organic substances, and suspended particles. Sea salts have negligible effect on attenuation. For pure seawater, Table 4 shows values of attenuance,

Table 4: Light Attenuance in Pure Seawater

wavelength (micron)	attenuance (percent per m)	scatterance (percent per m)	absorptance (percent per m)
0.375	4.4	0.7	3.7
0.400	4.2	0.5	3.7
0.450	1.9	0.3	1.6
0.500	3.5	0.2	3.3
0.550	6.7	0.1	6.6
0.600	16.7	0.1	16.6
0.650	25.0	0.1	24.9
0.700	39.3	0.0	39.3

scatterance, and absorptance for various wavelengths. Seawater is most transparent to blue light, and red light is strongly absorbed (infrared even more). The scattering produced by the water molecules is inversely proportional to the fourth power of the wavelength. Of the organic substances dissolved in seawater, a yellow substance, mainly carbohydrate-humic acids, especially adds significantly to the absorption of light, particularly in the shorter wavelengths. The substance is formed by decomposition of organic particulate matter and is especially abundant in northern coastal waters.

Suspended particles in the sea are responsible for absorption as well as scattering. The absorption is generally stronger for short wavelengths of light than for longer waves. Scattering, on the other hand, is virtually independent of wavelength. Table 5 shows some values of the attenuance for various wavelengths in different types of seawater.

Table 5: Loss of Light (Percent) in One Metre of Seawater*

	violet		blue		green	yellow	orange	red
Wave length (micron)	0.30	0.40	0.46	0.50	0.54	0.58	0.64	0.70
Oceanic water, most transparent	16%	4%	2%	3%	5%	9%	29%	42%
Oceanic water, least transparent	57%	16%	11%	10%	13%	19%	36%	55%
Coastal water, average		63%	37%	29%	28%	30%	45%	74%

*According to Jerlov.

Colour and transparency of seawater

The colour of seawater depends on the spectral distribution of attenuance. It is blue for clear ocean water and shifts toward greater wavelengths for less transparent (more turbid) waters, in accordance with the shift of the spectral-attenuation minimum.

The intensity of underwater light derived from the sunlight falling on the sea surface depends on the amount of reflection at the sea surface and on the depth and transparency of the intervening layer of water. The reflection from a smooth sea surface with the Sun at various elevations in a clear sky varies from 3 percent for a high Sun to very high values for a low Sun, as is shown by the following data:

Sun's elevation (degrees)	90	50	40	30	20	10	5
reflectance (percent)	3	3	4	6	12	27	42

If the sea surface is rough because of wave action, the reflectance is decreased at low elevations and increased at higher elevations of the Sun. Also, with an overcast sky, the differences of reflectance for different elevations of the Sun become smaller.

The transparency of the sea for daylight is usually smallest in the topmost few metres of water, because of the presence of small-sized drifting material and air bubbles (foam). The amount of light reaching various depths is, of course, different for different wavelengths. For average ocean water, the following values provide an example of the relative downward light intensities, expressed in percentage of the light intensity immediately below the sea surface:

depth (m)	0	10	20	50	130	200
relative intensity (percent)	100	9.5	3.7	0.31	5×10^{-4}	2×10^{-6}

Various instruments are used to measure optical properties of the sea, the most common being the transparency meter and the Secchi disk. The principal parts of the former, also called the turbidity meter, are a light source that emits a directed beam of light and a photoelectric cell placed at some fixed distance. The cell measures the intensity of the light transmitted through the intervening column of water. The Secchi disk, about 30 centimetres (12 inches) in diameter, painted white or yellow, is lowered into the sea, and the depth at which it vanishes from sight is taken as a measure of the transparency of the water.

Velocity of sound

Acoustical properties. The velocity of sound in seawater varies from about 1,450 metres per second to about 1,570 metres per second (about 4,760 to 5,150 feet per second); it increases with temperature at a rate of about 4.5 metres per second per degree C (8.2 feet per second per degree F), and it increases with salinity at a rate of 1.3 metres per second per part per thousand of salinity. The increase with depth (pressure) is at a rate of 1.70 metres per second per 100 metres (330 feet). At the sea surface, in water of 35 parts per thousand salinity and a temperature of 10° C (50° F), the velocity is 1,501 metres per second (4,923 feet per second). A typical sound-velocity profile along a vertical in the ocean generally shows a decrease of velocity with increasing depth down to about 1,500 metres, caused by the dominating effect of the temperature decrease, which is followed by a steady increase of velocity with greater depths. This velocity increase results from the effect of increasing pressure.

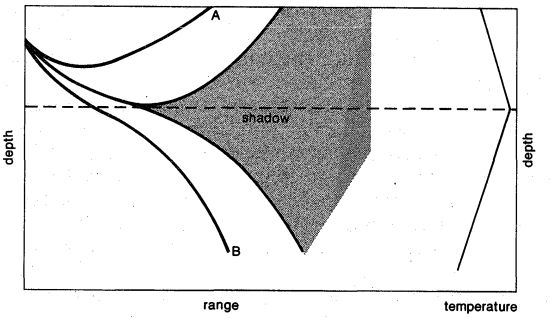


Figure 6: Acoustical properties of seawater: shadow zone formation that occurs when a positive temperature gradient lies above a negative temperature gradient. The sound projector is located within the positive gradient, and AB indicates boundaries of the sound channel (see text).

In the topmost ten to 100 metres of the sea, a slight increase of sound velocity with depth may occur because of the pressure effect if this top layer is a mixed layer with constant temperature and salinity.

The differences in sound velocity at different depths give rise to refraction, or bending, of the sound rays. When velocity decreases with increasing depth, a downward bending occurs; when it increases with depth, there is an upward bending of the sound rays. Refraction of sound gives rise to such phenomena as shadow zones and sound channels. A shadow zone is produced when a layer with downward increase of sound velocity occurs above a lay-

er with downward decrease of sound velocity. If a sound projector, as used in underwater acoustic detection, is in the upper layer, one gets a shadow zone as depicted in Figure 6.

A sound channel is caused by a sound-velocity minimum in the velocity profile, such as that mentioned above. Along the axis of the velocity minimum, sound rays are refracted successively upward and downward and essentially are trapped. The sound is focussed in the channel, and loss of intensity is minimized. Small charges exploded at the right depth (about 1,500 metres, or 5,000 feet) produce sounds that may be detected by hydrophones placed in the channel many thousands of miles from the source. This phenomenon is the principle of the SOFAR, sound fixing and ranging system used to locate positions at sea.

ICE IN THE SEA

There are two types of ice in the seas: sea ice, which is ice formed by the freezing of seawater, and ice that has come from land, such as icebergs and ice islands.

Sea ice. From an initial stage of so-called frazil crystals (floating needles and platelets) and sludge composed of them, sea ice grows to a compact aggregate of crystals of pure ice with pockets of seawater entrapped between them. Because of this composition, the salinity of sea ice is lower than that of the seawater from which it has grown. The initial sea-ice salinity may vary between 20 and two parts per thousand; the more rapid the freezing, the saltier the ice. After sea ice has formed, a process of salt removal by drainage of part of the enclosed brine sets in, because the cells in which it is contained are not completely isolated. Old ice has very low salinity, less than one part per thousand.

The growth rate of sea ice depends on surface temperature, the depth of snow cover, and the heat flux in the underlying water. In the central Arctic, the thickness of an ice cover formed in one growing season is about two metres (six and one-half feet). If the ice is not broken up, it finally reaches an equilibrium thickness of about three to four metres (10 to 13 feet) in five to eight years, when the annual ablation (loss by any means) at the top and the bottom equals the annual growth. In the Antarctic, perennial sea ice is found only in the Weddell Sea and a narrow strip around the continent. Most of the Antarctic sea ice is seasonal and reaches a thickness of about 1.5 metres (five feet) by the end of October.

The drastic change of reflectivity that occurs when the sea freezes over (from 5 to 10 percent to 80 percent) makes sea ice an important factor in the heat budget of the ocean. The boundaries of the sea ice are highly variable. In the Norwegian and Greenland seas, deviations of 300 kilometres (200 miles) north or south of the average position are not uncommon. The estimated mean areas of sea ice at the end of the summer and at the end of the winter in the Arctic are 9,000,000 square kilometres (3,500,000 square miles) and 12,000,000 square kilometres (4,600,000 square miles), respectively. In the Antarctic, the corresponding values are 4,000,000 square kilometres (1,500,000 square miles) and 20,000,000 square kilometres (8,000,000 square miles). The mean total volume of sea ice on Earth is 40,000 to 50,000 cubic kilometres (9,600 to 12,000 cubic miles), and the total amount of freezing and melting that occurs each year has been estimated at 30,000 cubic kilometres (7,200 cubic miles).

In the Arctic, it is possible to distinguish three regimes of sea ice: the great inner core, the permanent polar cap of sea ice (the Arctic pack), which covers about 6,000,000 square kilometres (2,300,000 square miles); around this the true drift ice or pack ice; and the landfast ice, which is present during nine months of the year, when it fringes the shores of the Arctic Sea out to the 22-metre (72-foot) depth line, approximately. Large amounts of pack ice drift southward each year. The ice discharge through the gap between Greenland and Spitsbergen is estimated to be 3,000 cubic kilometres (720 cubic miles) per year. On the west side of the North Atlantic, the pack ice reaches approximately latitude 45° N in winter and spring. On the east side, along the Norwegian coast, the sea remains open up to 73° N.

Ice islands and icebergs. Ice islands, of which a number have been found drifting in the Arctic Sea, are heavy sheets of ice that are far thicker than sea ice. Their thickness easily may amount to 50 metres (160 feet), five metres (16 feet) of which project above water. The surface area of the largest known ice island is approximately 1,000 square kilometres (400 square miles), but others are far smaller. Ice islands consist of a kind of glacier-like snow ice. The majority of them probably have been formed by the breaking of the shelf ice that borders the north coast of Ellesmere Island. The first ice island reported has undergone little change in configuration since its detection in 1946.

Icebergs and their distribution

Icebergs are formed by the calving (detaching of parts) of glaciers or of inland ice that reaches the sea. The main sources of icebergs in the northern seas are the valley glaciers of Greenland, which produce some 12,000 to 15,000 sizable icebergs annually. Almost as many are calved by the glaciers reaching the sea on the eastern seaboard as by those on the west coast, but the icebergs deriving from the east side do not travel much farther south than Kap Farvel, the southern tip of Greenland. The icebergs of the west coast, on the other hand, after travelling northward and across to the other side of Baffin Bay, are carried far south, along Baffin Island and Labrador, by the Labrador Current. It is estimated that about one in every 20 icebergs derived from west Greenland ends up south of Newfoundland (48° N), the greatest numbers arriving there in April, May, and June.

The icebergs of the Antarctic derive from an ice barrier, or shelf ice, a layer of ice that stretches out from the inland ice into the sea. It rests on the bottom near shore, but farther out to sea it floats on the water. Because of their origin, the Antarctic icebergs are very much longer than they are high, occasionally measuring some tens of kilometres in length. For this reason they are called table bergs.

The frequency with which icebergs occur in the Southern Ocean does not vary much with the season, in contrast to the North Atlantic occurrences. Generally speaking, October and November are the months in which they are most numerous in the south because of the release of the bergs from the pack ice in the southern spring. They reach farthest north from November to February. The average northern boundary for icebergs is about 40° S in the Atlantic Ocean, between 40° and 50° S in the Indian Ocean, and about 50° S in the Pacific. At least several thousands of them are adrift every year in the southern seas (see also ICEBERGS AND PACK ICE).

II. Dynamics and motions of the sea

OCEAN–ATMOSPHERE INTERACTION

Radiation and heat budget. Incoming sunlight is the primary source of energy for the atmosphere, the oceans, and the land. The heat flow from the interior of the Earth through the sea bottom is approximately 0.1 watt per square metre (see EARTH, HEAT FLOW IN). The supply of energy to the oceans by direct and scattered solar radiation, averaged over all seas and for the entire year, has been estimated at 150 watts per square metre, or 3.5 kilowatt hours per day per square metre, less than half of the incoming solar radiation at the top of the atmosphere, which averages 350 watts per square metre. The loss is due to backscattering and absorption by the air, reflection by clouds, and reflection by the surface of the sea. Besides this shortwave primary radiation, there is a considerable supply of infrared heat radiation from the atmosphere, derived from the clouds, the water vapour, and, to a small extent, from carbon dioxide. This radiation from the atmosphere is quite variable, depending on the water-vapour content, the temperatures in the lower layers of the atmosphere, and, particularly, on the cloud cover. Averaged over the whole World Ocean, it amounts to some 235 watts per square metre. On the other hand, the Earth's surface loses heat by radiation, by evaporation, and by direct conduction through contact with the atmosphere. The loss by radiation at any place is propor-

Energy gains and losses

tional to the fourth power of the absolute temperature of the Earth's surface at that place. For the whole World Ocean, the average outgoing radiation is estimated at 300 watts per square metre. The average heat loss by evaporation is estimated at 75 watts per square metre, and the average heat conduction to the overlying air is about ten watts per square metre. Thus, the total heat gains and heat losses both amount to 385 watts per square metre. The considerable effect of the atmospheric down radiation upon the temperatures of the Earth's surface is called the greenhouse effect of the atmosphere.

Heat losses by evaporation and by conduction vary greatly from place to place and from time to time. They may even be negative, because there may be a gain of heat by condensation on the sea surface instead of evaporation at certain places. Often there is conduction of heat from the air to the seawater if the seawater is colder, as is usually the case in summer and in regions of cold ocean currents. In general, evaporation is greatest in winter, when the seawater is usually warmer than the overlying air.

Although taken as a whole there is a balance between the thermal gains and losses of the oceans of the world, this need not be the case for any particular area, because transport of heat within the oceans occurs by current action. In the northeastern part of the Atlantic Ocean to the west of Norway, expenditure in the form of heat transfer to the atmosphere and upward radiation exceeds the incoming radiation. This deficit is balanced, however, by the internal transfer of heat carried through the ocean by the Gulf Stream system. The reverse occurs in the northwestern part of the Atlantic, where cold water carried down by the Labrador Current with its burden of ice fields balances out the surplus of heat received from above.

Water budget. The annual water budgets of the oceans and the lands are approximately as follows:

	oceans	lands
evaporation km³/year	360,000	70,000
precipitation km³/year	330,000	100,000
runoff km³/year	30,000	−30,000

Stated differently, the values are

	oceans	lands
evaporation m³/sec	11.4×10^6	2.2×10^6
precipitation m³/sec	10.4×10^6	3.2×10^6
runoff m³/sec	1.0×10^6	-1.0×10^6

The annual precipitation averaged over all the oceans and seas and expressed as the height of a water column amounts to 91 centimetres (36 inches). All these figures are estimates and incorporate a considerable degree of uncertainty, but they reflect one important aspect of atmosphere–ocean interaction (see also HYDROLOGIC CYCLE).

Air–sea transfer processes. The upward or downward transport of water vapour and heat from or toward the sea surface was referred to above in relation to the energy exchange of the sea with the air. This transport is performed by fluctuating vertical motions of the air immediately above the water surface. Except within about one millimetre (0.04 inch) of the interface, where molecular processes are of importance, turbulence is responsible for the vertical transport of matter and heat. Another property that is important in connection with transport is horizontal momentum. Momentum per unit mass is velocity, the rate of supply of momentum is force, and the rate of transfer of horizontal momentum per unit area of the surface is force per unit area, which is a stress exerted through the surface. A downward transfer of horizontal momentum across the sea surface means that a wind stress is exerted on the water.

Horizontal and vertical transport

The rate of turbulent transfer of any property depends on spatial differences of that property—that is, on its gradient (difference per unit distance)—and on the magnitude of the turbulent velocities. The definition and significance of turbulent velocity will be considered later in connection with turbulence in the sea. If in the lowest layers of air there is an upward gradient of horizontally averaged value of any mass property (Q)—*e.g.*, heat content, water-vapour content, or momentum per unit volume—a parcel of air moving downward at a certain level is liable to carry a higher value of Q with it than does an upward moving parcel of air at the same level; statistically, then, there will be a net downward flux of the corresponding quantity (heat, water vapour, momentum). Thus, the direction of flux is opposite to the direction of the gradient of Q and this is also the case with ordinary molecular diffusion. The relationship between flux and gradient is often written as

$$\text{flux} = -A \times \text{gradient of } Q,$$

in which A is the turbulent exchange (or Austausch) coefficient. Instead of A, it is customary to write $A = \text{K} \times$ density (of air), in which K is the turbulent, or eddy, diffusivity for the property concerned. In contrast to molecular diffusivity, K is not a constant characteristic of the air but is dependent upon the intensity of the turbulence and on the distance from the boundary surface. In the case of air over water, for example, it depends on the height above the water surface. In general, it will also depend on the property concerned; that is, it may be different for the exchange of water vapour, heat, and momentum. In the case of momentum transfer, K is called the eddy viscosity.

For practical purposes, instead of using the local gradient Q for determining the corresponding flux, the difference of Q between two levels is often used. For instance, the difference between the value at the sea surface and the value at some standard level, usually chosen to be ten metres (33 feet) above sea level, can substitute for the gradient of Q. The relationship used is then: upward flux $= -D\ (Q_1 - Q_0)$, in which D is the eddy diffusion factor for the property concerned. Empirically, this factor has been shown to be roughly proportional to the wind velocity (U_1) at standard height, at least if the air is not too unstable or too stably stratified. The formula for the flux then becomes: upward flux $= -CU_1\ (Q_1 - Q_0)$, in which the factor C represents a nondimensional quantity (a pure number), which depends on what meaning has been assigned to Q.

In the case of wind stress, which is a downward transfer of momentum, Q means density times wind velocity, so that, if the wind velocity at the very sea surface is supposed to be zero, the above formula becomes:

$$\text{wind stress} = C_1 \rho U_1^2,$$

in which rho (ρ) equals air density. C_1, called the drag coefficient, depends on the height of the standard level chosen and on the roughness of the sea surface or the small corrugations of that surface and consequently is not a true constant. Empirically, it has been shown to have values of about 0.001 to 0.002 for wind velocities of at least seven metres per second. It increases somewhat with increasing wind velocity. Also, it tends to be greater for pronounced instability of the air and to be smaller for pronounced stability.

Three movements of seawater

Three types of movements of seawater may be distinguished, namely, oscillating and rotating movements, as in water waves and tides; regular, or direct, currents; and irregular eddying movements, or turbulence that is superposed on waves or ocean currents as a form of "noise" (secondary or background disturbance).

Turbulence in the sea. The distinction between regular currents and turbulence is not absolute, because a varying current system of restricted dimensions within the framework of a larger current system may be looked upon as an element of turbulence with respect to that larger system. It may also be considered as a regular, though not constant, current system of its own with respect to still smaller eddies, with its own field of superposed turbulence. If, then, the turbulent velocity is defined as the momentary velocity at a certain point minus the average velocity of the current (the average being taken over a certain area), the outcome of this definition depends on the scale of averaging.

The effect of turbulent motions is to bring about an exchange of mass, heat, and momentum in much the same way as molecular motions bring about molecular diffu-

sion, heat conduction, and viscosity. The flux, or the net amount of substance, heat, or momentum that passes per unit time through a surface of unit area by turbulence, is proportional to the gradient of mean mass of the substance concerned per unit volume. The coefficient of proportionality, called eddy diffusivity, eddy conductivity, or eddy viscosity, respectively, is not a physical constant (as is molecular diffusivity, conductivity, or viscosity) but depends on the nature of the turbulent motion and on the scale of averaging. It depends on the scale of the phenomenon under consideration, and, consequently, it may have much larger values for horizontal than for vertical exchange. On the whole, it may vary from 0.1 to 10^9 square centimetres per second, increasing with increasing scale of the phenomenon. It has the same order of magnitude for the exchange of mass, heat, and momentum. Vertically, it also depends on the stability of the stratification; pronounced stability tends to counteract the development of turbulence. In the ideal case of statistical equilibrium between the eddies of different sizes, eddy diffusivity has been shown to be proportional to the 4/3rd power of the scale involved. As a consequence, the horizontal spreading of a patch of matter (suspended particulate matter or a dissolved substance) goes on at a rate that increases in time with the dimensions of the patch; indeed, as the patch becomes larger, ever larger eddies participate in spreading it; as long as the patch is of a smaller size, a larger eddy only displaces it, the spreading being left to the smaller eddies.

Eddy viscosity, concerned with momentum, is of paramount importance for the dynamics of currents in the sea, particularly for wind-driven currents. In the neighbourhood of boundaries, its value decreases with decreasing distance to the boundary (*e.g.*, bottom or sea surface) and depends on its roughness.

WAVES OF THE SEA

Waves of the sea are of many different kinds, and only gravity waves, of which gravity is the stabilizing factor, will be considered here. Gravity waves may be surface waves or internal waves; and, in either case, a further distinction can be made between running waves and standing waves.

The tides are essentially gravity waves that are running, standing, or in an intermediate state. The tides may be termed forced waves, because they have fixed, prescribed periods that are strictly determined by the relative movements of moon, Earth, and Sun. Sometimes the word tide is used in a wider, somewhat loose sense, including such phenomena as surges, which are called storm tides, or meteorological tides. Also, the term tidal wave is sometimes used in a general sense to denote very long gravity waves; that is, waves whose wavelength is very much greater than the depth of the sea. In the following discussion the use of the words tide and tidal will be restricted to the tides of astronomic origin and the forces and phenomena connected with them.

Surface waves. Of the nontidal kinds of running surface waves, three types may be distinguished, namely, wind waves and swell, wind surges, and sea waves of seismic origin (tsunamis).

Wind waves and swell. Wind waves are the wind-generated waves that are controlled and strengthened by the wind or wind field that made them. Afterward, when the wind has abated or shifted or when the waves have left the wind field, they run on independently as swell.

The dependence of the sizes of the waves on the wind field is a complicated one. A general impression of this dependence is given by the descriptions of the various states of the sea corresponding to the scale of wind strengths known as the Beaufort scale (Table 6), after the British admiral Sir Francis Beaufort, who drafted it in 1808, using as his yardstick the surface of sail that a fully rigged warship of those days could carry in the various wind forces. In the list given in Table 6, each Beaufort number is accompanied by a brief description of the appearance of the sea, on the lines adopted by the German sailing-ship captain Peterson. These descriptions were approved by the World Meteorological Organization for use at sea to determine the force of the wind. The Beaufort wind force is followed by the name given to such a wind at sea, and the next column gives the range of wind speeds.

Table 6: The Beaufort Scale

Beaufort number	name of wind	wind speed		description of sea surface
		knots	km/hr	
0	calm	<1	<1	like a mirror
1	light air	1–3	1–5	ripples with the appearance of scales are formed, but without foam crests
2	light breeze	4–6	6–11	small wavelets, still short but more pronounced; crests have a glassy appearance and do not break
3	gentle breeze	7–10	12–19	large wavelets; crests begin to break; foam of glassy appearance, perhaps scattered white horses
4	moderate breeze	11–16	20–28	small waves, becoming longer; fairly frequent white horses
5	fresh breeze	17–21	29–38	moderate waves, taking a more pronounced long form; many white horses are formed (chance of some spray)
6	strong breeze	22–27	39–49	large waves begin to form; the white foam crests are more extensive everywhere (probably some spray)
7	moderate gale	28–33	50–61	sea heaps up and white foam from breaking waves begins to be blown in streaks along the direction of the wind
8	fresh gale	34–40	62–74	moderately high waves of greater length; edges of crests begin to break into spindrift; the foam is blown in well-marked streaks along the direction of the wind
9	strong gale	41–47	75–88	high waves; dense streaks of foam along the direction of the wind; crests of waves begin to topple, tumble and roll over; spray may affect visibility
10	whole gale	48–55	89–102	very high waves with overhanging crests; the resulting foam, in great patches, is blown in dense white streaks along the direction of the wind; on the whole the surface of the sea takes a white appearance; the tumbling of the sea becomes heavy and shock-like; visibility affected
11	storm	56–65	103–117	exceptionally high waves (small and medium-sized ships might be for a time lost to view behind the waves); the sea is completely covered with long white patches of foam lying along the direction of the wind; everywhere the edges of the wave crests are blown into froth; visibility affected
12–17	hurricane	above 65	above 117	the air is filled with foam and spray; sea completely white with driving spray; visibility very seriously affected

When considering the descriptions, it must be remembered that the size of the waves depends not only on the strength of the wind but also on its duration and its fetch; that is, the length of its path over the sea. Moreover, the waves are liable to be modified considerably by tidal currents; the sea is affected by precipitation (rainfall, snow, hail); and at moderate and high latitudes, at the same wind speed at observation level, the motion of the sea is higher in an air mass colder than the water than it is in one that is warmer.

Wave characteristics and motion

The theory of waves starts with the concept of simple waves, those forming a strictly periodic pattern with one wavelength and one wave period and propagating in one direction; but real waves always have a more irregular appearance. They may theoretically be described as composite waves, in which a whole spectrum of wavelengths, or periods, is present and which have more or less diverging directions of propagation. In reporting observed wave heights and periods (or lengths) or in forecasting them, one height or one period is mentioned as the height or period, however, and some agreement is needed in order to guarantee uniformity of meaning. The height of simple waves means the elevation difference between the top of a crest and the bottom of a trough. The significant height, a

characteristic height of irregular waves, is by convention the average of the highest one-third of the observed wave heights. Period, or wavelength, can be determined from the average of a number of observed time intervals between the passing of successive well-developed wave crests over a certain point, or of observed distances between them.

Wave period and wavelength are coupled by a simple relationship: wavelength equals wave period times wave speed, or $L = TC$, when L is wavelength, T is wave period, and C is wave speed.

The wave speed of surface gravity waves depends on the depth of water and on the wavelength, or period; the speed increases with increasing depth and increasing wavelength, or period. If the water is sufficiently deep, then the wave speed equals the wave period times the acceleration caused by gravity, divided by two pi, or the square of the wave speed equals the wavelength times the acceleration caused by gravity, divided by two pi; these relations may be expressed as:

$$C = gT/2\pi \text{ or } C^2 = gL/2\pi,$$

in which C is wave speed, g is the acceleration caused by gravity (9.8 square metres per second), and T and L are again wave period and length. In this case, the depth does not appear explicitly in the formula. Here, deep and shallow have only a relative meaning, denoting the ratio of depth to wavelength. In practice, for the water to be called deep and the above formula applied, it is sufficient if the depth is more than half the wavelength. A few examples are listed below, giving the period in seconds, the wavelength in metres, and wave speed in metres per second:

period T (seconds)	1	2	4	8	16
wavelength L (m)	1.56	6.2	25.0	100	400
wave speed in deep water (m/sec)	1.56	3.1	6.2	12.5	25.0

For waves in very shallow water, another simple formula holds, one in which wavelength and period do not appear explicitly. It is: wave speed squared equals the depth times the acceleration caused by gravity, or:

$$C^2 = gD,$$

in which D means depth. In practice, this formula may be applied if the depth is less than $\frac{1}{25}$ of the wavelength or if L is more than 25 times the depth. Instead of denoting this expression as one for waves in very shallow water, it often is considered to be for long waves, both expressions meaning the same thing.

Wave trains and group velocity

Waves often appear in groups as the result of interference of wave trains of slightly differing wavelengths. A wave group as a whole has a group speed that generally is less than the speed of propagation of the individual waves; the two speeds are equal only when the waves are in very shallow water. For deep-water waves, the group velocity (V) is half the wave speed (C). In intermediate cases, V has a value from 0.5 C to C. In the physical sense, group velocity is the velocity of propagation of wave energy. From the dynamics of the waves, it follows that the wave energy per unit area of the sea surface is proportional to the square of the wave height, except for the very last stage of waves running into shallow water, shortly before they become breakers.

The height of wind waves increases with increasing wind speed and with increasing duration and fetch of the wind. Together with height, the dominant wavelength also increases. Finally, however, the waves reach a state of saturation, because they attain the maximum significant height to which the wind can raise them, even if duration and fetch are unlimited. For instance, winds of five metres (16 feet) per second, 15 metres (50 feet) per second, and 25 metres (80 feet) per second may raise waves with significant heights up to 0.5 metre (1.6 feet), 4.5 metres (15 feet), and 12.5 metres (41 feet), respectively, with corresponding wavelengths of 16 metres (53 feet), 140 metres (460 feet), and 400 metres (1,300 feet), respectively.

After becoming swell, the waves may travel thousands of miles over the ocean, particularly if the swell is from the great storms of moderate and high latitudes, whence it easily may travel into the subtropical and equatorial zones, and the swell of the trade winds, which runs into the equatorial calms. In travelling, the swell waves gradually become lower; energy is lost by internal friction and air resistance and by energy dissipation because of some divergence of the directions of propagation (fanning out). With respect to the energy loss, there is a selective damping of the composite waves, the shorter waves of the wave mixture suffering a stronger damping over a given distance than the longer ones. As a consequence, the dominant wavelength of the spectrum shifts toward the greater wavelengths. Therefore, an old swell must always be a long swell.

When waves run into shallow water, their speed of propagation and wavelength decrease, but the period remains the same. Eventually, the group velocity, the velocity of energy propagation, also decreases, and this decrease causes the height to increase. The latter effect may, however, be affected by refraction of the waves, a swerving of the wave crests toward the depth lines and a corresponding deviation of the direction of propagation. Refraction may cause a convergence or divergence of the energy stream and result in a raising or lowering of the waves, especially over nearshore elevations or depressions of the sea bottom.

In the final stage, the shape of the waves changes, and the crests become narrower and steeper until, finally, the waves become breakers (surf). Generally, this occurs where the depth is 1.3 times the wave height.

Wind surges. Running wind surges are long waves caused by a piling up of the water over a large area through the action of a travelling wind or pressure field. Examples include the surge in front of a travelling storm cyclone, particularly the devastating hurricane surge caused by a tropical cyclone, and the surge occasionally caused by a wind convergence line such as a travelling front with a sharp wind shift.

Generation of tsunamis

Waves of seismic origin. A tsunami (Japanese *tsu*, "harbour," and *nami*, "sea") is a very long wave of seismic origin that is caused by a submarine or coastal earthquake, landslide, or volcanic eruption. Such a wave may have a length of hundreds of miles and a period of the order of a quarter of an hour. It travels across the ocean at a tremendous speed; to a depth of 4,000 metres (13,000 feet), for instance, the corresponding wave speed is about 200 metres per second, or 720 kilometres per hour (450 miles per hour). Tsunamis have caused enormous destruction of life and property, because they ultimately pile up in coastal waters at places thousands of miles away from their point of origin, particularly in the Pacific Ocean.

A freestanding wave may arise in an enclosed or nearly enclosed basin as a free swinging of the whole water mass. Such a standing wave is also called a seiche, after the name given to the oscillating movements of the water of Lake Geneva, where this phenomenon first was studied seriously. The period of oscillation is independent of the force that first brought the water mass out of equilibrium (and that is supposed to have ceased after that) but depends only on the dimensions of the enclosing basin and on the direction in which the water mass is swinging. Assuming a simple rectangular basin of constant depth and supposing the most simple lengthwise oscillation, the period of oscillation (T) is equal to two times the length of the basin divided by the wave speed computed from the shallow-water formula above. This relationship may be written: $T = L/C$, in which L equals two times the length of the basin and C is the wavespeed found from the formula, using the known depth of the basin. Besides this fundamental tone (or response to stimuli), the water mass may also swing according to an overtone, showing one or more nodal lines across the basin.

The water in an open bay or marginal sea may also perform such a free oscillation as a standing wave, the difference being that in an open bay the greatest horizontal displacements are not in the middle of the bay but at the mouth. For the fundamental period of oscillation, the

formula given above is used with a wavelength equal to four times the length (from the mouth to the closed end) of the bay. In practice, of course, it is more difficult than that, because the form of a bay or marginal sea is irregular, and the depth differs from place to place. The North Sea has a period of lengthwise swinging of about 36 hours. The cause of such free oscillations may be a temporary wind or pressure field, which brought the sea surface out of its horizontal position and which afterward ceased to act more or less abruptly, leaving the water mass out of equilibrium.

Internal waves. Waves that have their maximum energy at some depth rather than at the surface are called internal waves. They are carried by an interface, or layer, separating lighter water above from heavier water below; the difference in density is caused either by a difference in salinity or by a difference in temperature. Internal waves manifest themselves by a regular rising and sinking of the water layers around which they centre, whereas the height of the sea surface is hardly affected at all. Because the restoring force, excited by the internal deformation of the water layers of equal density, is much smaller than in the case of surface waves, internal waves are much slower than the latter. Given the same wavelength, the period is much longer (the movements of the water particles being much more sluggish), and the speed of propagation is much smaller; the formulas for the speed of surface waves include the acceleration of gravity (g), but those for internal waves include the gravity factor times the difference between the densities of the upper and the lower water layer divided by their sum.

Role of density differences

The cause of internal waves may lie in the action of tidal forces (the period then equalling the tidal period) or in the action of a wind or pressure fluctuation. Sometimes, a ship may cause internal waves (dead water) if there is a shallow, brackish upper layer.

Tides of the sea. The tides may be considered as a kind of forced waves, partially running waves, partially standing waves. They are manifested by vertical movements of the sea surface (the height maximum and minimum are called high water [HW] and low water [LW]) and in alternating or rotating horizontal movements of the water, the tidal currents. The words ebb and flow or flood are used to designate the falling tide and the rising tide or the tidal currents accompanying the falling tide and the rising tide, respectively.

The forces that cause the tides are called the tide-generating forces. A tide-generating force is the resultant force of the attracting force of the moon or the Sun and the force of inertia (centrifugal force) that results from the orbital movement of the Earth around the common centre of gravity of the Earth–moon or Earth–Sun system. Details of this mechanism are treated in the article TIDES.

Tide-generating forces

Considering the Earth–moon system, at any time the tide-generating force is directed vertically upward at the two places on the Earth where the moon is in the vertical (on the same and on the opposite side of the Earth); it is directed vertically downward at all places (forming a circle) where the moon is in the horizon at that moment; at all other places, the tide-generating force also has a horizontal component. Because this pattern of forces is coupled to the position of the moon with respect to the Earth and because for any place on the Earth's surface the relative position of the moon with respect to that place has, on the average, a periodicity of 24 hours 50 minutes, the tide-generating force felt at any place has that same periodicity. When the moon is in the plane of the Equator, this force runs through two identical cycles within this time interval, because of the symmetry of the global pattern of forces described above. Consequently, the tidal period is 12 hours 25 minutes in this case; it is the period of the semidiurnal lunar tide. The fact that the moon is alternately to the north and to the south of the Equator causes an inequality of the two successive cycles within the time interval of 24 hours 50 minutes. The effect of this inequality is formally described as the superposition of a partial tide called the diurnal lunar tide, with the period of 24 hours 50 minutes, on the semidiurnal lunar tide.

In the same manner, the Sun causes a semidiurnal solar tide, with a 12-hour period, and a diurnal solar tide, with a 24-hour period. In a complete description of the local variations of the tidal forces, still other partial tides play a role, because of further inequalities in the orbital motions of the moon and the Earth.

The interference of the solar-tidal forces with the lunar-tidal forces (the lunar forces are about 2.2 times as strong) causes the regular variation of the tidal range between spring tide, when it has its maximum, and neap tide, when it has its minimum.

Although the tide-generating forces are very small in comparison with the force of gravity of the Earth (the lunar tidal force at its maximum being only 1.14×10^{-7} times the force of gravity), their effects upon the sea are considerable, especially because of their horizontal component. Because the Earth is not surrounded by an uninterrupted envelope of water but shows a very irregular alternation of sea and land, the mechanism of the response of the oceans and seas to the tidal forces is extremely complex, a further complication being brought about by the deflecting force of the Earth's rotation (Coriolis force).

In the Southern Ocean, the tide is propagated east–west around the world as a running wave, but it manifests itself between the continents—for instance, in the Atlantic Ocean—partly as a north-going running wave and partly as an east–west swinging standing wave; both wave movements are modified by the action of the Coriolis force. In gulfs and bays, the tide is generated by the tide of the open ocean as a forced standing wave to which the Coriolis force adds a swinging cross component, and both oscillations together result in a rotating tide. In such nearly enclosed seas as the Mediterranean, the Black, and the Baltic, a standing wave is generated by the direct action of the local tidal forces.

In these seas, the tidal range of sea level is only on the order of centimetres (one inch equals 2.5 centimetres). In the open ocean, it generally is on the order of tens of centimetres. In bays and adjacent seas, however, the tidal range may be much greater, because the shape of a bay or adjacent sea may favour the development of the tide inside; in particular, there may be a resonance (large co-oscillation) of the basin concerned with the tide. The largest known tides occur in the Bay of Fundy, where spring tidal ranges up to 15 metres (50 feet) have been measured.

OCEAN CURRENTS

Causes of ocean currents

Two primary causes of regular, nontidal currents in the sea may be distinguished, namely, wind and gravity. Winds act externally at the surface, and gravity acts internally, through pressure differences.

In the case of a wind-driven current system, the current may be the direct result of the wind locally, or it may be controlled by the wind elsewhere. The latter, for instance, may occur if the current is the returning part of a closed circulation of which one part only is directly driven by the wind. Examples are the North Equatorial Current and the South Equatorial Current, which flow east–west and are directly driven by the trade winds (on the north and south sides of the Equator), and the Equatorial Countercurrent, which runs from west to east and brings part of the water of those currents back though the belt of equatorial calms.

For the generation of gravity currents, pressure differences in a horizontal plane in the sea may be brought about by a slope of the sea surface, by juxtaposition of waters of different specific gravity, or, more often, by both. Combined forces operate in the Strait of Gibraltar, where the saltier (heavier) water of the Mediterranean Sea flows out in the lower layers, and the somewhat less saline (lighter) water of the Atlantic Ocean flows in in the upper layer. The same occurs at the mouth of the Red Sea, where it communicates with the Indian Ocean. Similar currents, called density currents (*q.v.*), flow through the entrances of the Baltic Sea and the Black Sea. There, however, the outgoing current is the upper one, because the waters of these seas have low salinities and are there-

fore lighter than the waters flowing in from the outside, which form the undercurrent.

Wind-driven-current patterns. In each of the three oceans, the general pattern of the great ocean currents is controlled largely by the systems of prevailing winds and is broadly as follows: to both sides of the Equator flow the North and South Equatorial currents, driven by the trade winds, an Equatorial Countercurrent flowing between them. On the west side of the ocean, the two equatorial currents bend poleward (except in the North Indian Ocean) and merge into currents that feed the broad west–east travelling currents in moderate and higher latitudes under the westerly winds prevailing there, and these eastbound currents, in their turn, feed currents that on the east side of the ocean flow toward the Equator and in low latitudes merge into the equatorial currents, thus closing the circuit.

In the North Atlantic Ocean, the northerly current on the west side is the Gulf Stream, coming from the Gulf of Mexico, into which part of the water of the North Equatorial Current has flowed. Much of the water carried by the Gulf Stream does not join the corresponding eastgoing current but flows in a northeasterly direction toward the coasts of northwestern Europe, finally entering into the Arctic Sea to the north of Norway and near Spitsbergen. As compensating currents, on the west side of the northern part of the Atlantic Ocean, the East Greenland Current and the Baffin Island-Labrador current carry polar waters to the south.

In the Pacific Ocean, a warm current, analogous to the Gulf Stream, is called Kuro-shio, or Japan Current. The cold countercurrent on the northwest side of it is called Oya-shio.

In the North Indian Ocean, the North Equatorial Current is present only during part of the year. In summer it is replaced by the Monsoon Current, driven by the southwest monsoons (*q.v.*).

A peculiar phenomenon in the equatorial belts of the oceans is the Equatorial Undercurrent, a subsurface current running at some depth, down to about 300 metres (1,000 feet) below the surface, from west to east.

In the Southern Ocean, the easterly currents of the three oceans merge into one broad current, encircling the whole of the Southern Hemisphere, the great West Wind Drift. Lastly, there is a narrow stream from east to west along the edges of Antarctica.

Current-generating forces. The dynamics of the currents of the sea are governed by the following forces: the pressure-gradient forces (vertical and horizontal), frictional forces, gravity, and the Coriolis force. The force of gravity is, on the average, balanced by the vertical pressure-gradient force. Thus, the three remaining forces, acting horizontally, are: the horizontal pressure-gradient force (G), the frictional force (F), and the horizontal Coriolis force (C). In general, dynamic systems in the ocean tend toward an approximate equilibrium of forces. In steady water movements, therefore, the three forces, G, F, and C, may be supposed to balance each other. Of these forces, the Coriolis force is always present wherever there is a current, because C is proportional to the velocity. Moreover, because it is proportional to the sine of the geographic latitude, it vanishes only at the Equator. In latitudes of about 45° N or S, for a velocity of one metre per second, the Coriolis force amounts to 0.0001 square metre per second per unit mass. It acts at right angles with the direction of motion, pointing to the right on the Northern Hemisphere and to the left on the Southern Hemisphere.

Effect of the Coriolis force

With respect to the other two forces, two special cases deserve attention: first, that in which C and G balance, friction being negligible; and, second, that in which C and F balance, pressure gradients being negligible.

In the first case, the current is called a gradient, or geostrophic, current, which results from a balance between the horizontal Coriolis and pressure-gradient forces; it is comparable to the geostrophic wind in meteorology. Because of the supposed equilibrium of forces, it runs at right angles with the horizontal pressure gradient —the higher pressures are on the right-hand side of the current if it is on the Northern Hemisphere and on the left-hand side if it is on the Southern Hemisphere. Near the surface, a horizontal pressure gradient in the water means a tilt of the sea surface; the surface slopes up to the right of the current in the Northern Hemisphere. At about 45° N or S, for a velocity of one metre per second, this slope amounts to 1:100,000; that is, one metre per hundred kilometres. In deeper layers, the horizontal pressure gradient corresponds in the same way with the slope of an isobaric surface (surface of equal pressure). A simple formula makes it possible to calculate the gradient current at any depth from the slope of the isobaric surface at that level. Assuming that the water at very great depths is nearly at rest, so that the isobaric surfaces there are about horizontal, it is possible to calculate the slopes of isobaric surfaces at higher levels from known values of the density and thus to calculate the gradient current. This method assumes that the current is, on the average, influenced to only a minor extent by the action of the wind or bottom friction, because wind and bottom friction are the main causes of internal friction between water layers.

In the second special case, drift current, wind stress is the main acting force: horizontal pressure gradients are supposed to be negligible, and the Coriolis force is balanced by the force of internal friction that accompanies the wind-made drift of the water. Because of the Coriolis force, the direction of the current deviates to the right of the wind direction in the Northern Hemisphere. At the surface, the angle of deviation is about 40°, but it increases with increasing depth, and the current velocity becomes smaller and smaller, virtually vanishing at some depth that is roughly proportional to the wind velocity. With a wind of ten metres per second, at moderate latitudes, this depth—called depth of friction—is about 90 metres (300 feet). The velocity at the surface amounts to about 1.5 percent of the wind velocity, at moderate latitudes, because depth of friction and surface velocity both decrease with increasing latitude.

This case of a pure drift current is rather theoretical. In reality, wind-driven currents usually are accompanied by a slope of the sea surface. This condition obtains for a wind blowing parallel to a coast. The Coriolis force causes the water to tend to move away from or toward the coast, thus creating a slope of the sea surface that may partly balance the Coriolis force. This condition is of particular importance in subtropical latitudes on the west side of the continents, where an upwelling of water from deeper layers is brought about in this way. The upwelling compensates for the surface water that flows away from the coast. Such upwelling of rather cold water has a profound influence on the climate of these coastal areas. Moreover, the upwelling water is rich in nutrients, so that there is abundant life in the zones in which upwelling occurs. Of course, upwelling is not confined to the regions just mentioned.

Deep-sea circulation. Water movements in the cold realms at great depths form part of a global deep-sea circulation, treated earlier in the sections on chemical and physical properties of seawater, particularly with respect to oxygen content and temperature. The water masses involved in this deep circulation come from source regions where waters sink from the surface because of their excessive density. The two prominent source regions of the deep-sea water masses are: (1) The Weddell Sea bordering the Antarctic continent, where the heaviest kind of seawater of all the oceans flows down to form the Antarctic Bottom Water and spreads northward over the beds of the three oceans, well into the Northern Hemisphere; and (2) the Irminger Sea between Iceland and Greenland and the area between south Greenland and Labrador, where cold and heavy water masses sink in winter to form the Atlantic Deep Water, which presses on southward, filling a large portion of the Atlantic deep sea (where it is joined by highly saline Mediterranean water), far into the Southern Hemisphere on top of the Antarctic Bottom Water.

Flow of water from the Arctic and Antarctic

These two water masses together form most of the deep-sea water of all the oceans, spreading eastward from

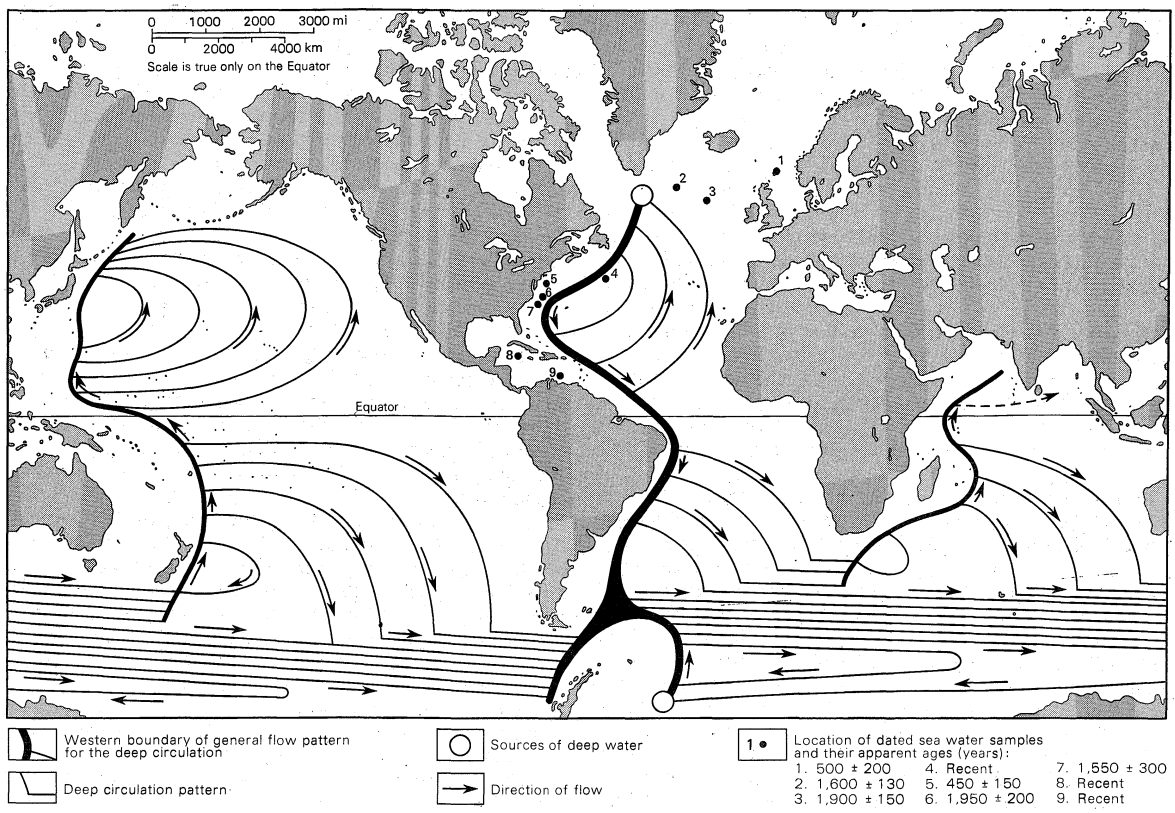


Figure 7: Deep-sea circulation pattern of the world's oceans.

From Henry Stommel, *The Gulf Stream: A Physical and Dynamical Description*, originally published by the University of California Press; reprinted by permission of The Regents of the University of California

the South Atlantic into the Indian Ocean and further on into the Pacific Ocean. Figure 7 illustrates this horizontal spreading. (P.Gr.)

III. Life in the open sea

THE SEA AS A BIOLOGICAL ENVIRONMENT

Despite its many moods and its ceaseless activity, the sea is the most constant environment of earth. This most equable habitat for living things is characterized by very narrow ranges of change in chemical and physical properties, in seasonal and horizontal fluctuations, in temperature (from about 0° to 30° C; 32° to 86° F), and in salinity. Although vertical pressure differences are significant, they are relatively constant at any particular depth.

Because of mixing processes, the relatively unstratified, upper lighted (euphotic) zone is one of vertical and horizontal constancy. The relative constancy of the euphotic zone is due to surface mixing processes that favour oxygenation of the water, maximum levels of photosynthesis, and thereby utilization of the existing nutrients. It is this zone of active plant–animal growth that determines oceanic productivity. From it, by a downward rain of organic material, arises the enormous stores of nutrients in the deep, dark layers of the open ocean.

In the deep strata, with high concentrations of carbon dioxide and relatively low concentrations of oxygen, vertical stratification occurs, but horizontal constancy remains unbroken. In the lighted strata, plants are responsible for seasonal variations in concentrations of inorganic salts, some of which, such as phosphorus, are physiologically important, and even slight departures from constancy are significant.

The open ocean possesses relative uniformity of kinds and numbers of planktonic organisms, especially of the smallest planktonic forms, nannoplankton, which do not vary appreciably, because they utilize dissolved organic substances available uniformly in the euphotic layer. The tendency toward biological constancy lessens for larger organisms that undergo periodic vertical or horizontal migrations in response to light and current.

It is reasonable to speculate that during early geologic time the lifeless waters encompassing the continents fluctuated in mineral content, augmented by periodic runoff from land areas, and eventually the salt content and surrounding physical factors reached levels favourable to the development of life. The first photosynthetic organisms were probably bacteria that could oxidize sulfides anaerobically (in the absence of free oxygen) without the production of free oxygen. (The ability to use hydrogen sulfide as a hydrogen donor permits photosynthesis with 1/10 the light energy required by oxygen-producing photosynthesis.) This circumstance seemingly had the advantage of increasing the depth zone for primitive life in the early ocean environment. The presence of blue-green algae, for example, in Precambrian sediments indicates that microscopic algae were well developed in relatively deep, upper strata.

First organisms

For higher marine animals, a water environment containing salts in such concentration that no effective movement of water (*i.e.*, no osmosis) takes place between it and the organism is the most favourable medium. This similarity has led to the supposition that the sea was already saline when life arose in the sea. Evidence that the sea is the original environment of animal life is based on the general composition of present-day faunas, similarity in chemical composition of body fluids and seawater, life histories, and paleontological relationships. Considering all animals, marine groups predominate. The role of the marine environment was to develop and maintain a wide diversity of lower forms of life. In contrast, the terrestrial environment, with more rigorous habitats, has produced less diversity of forms but a higher type of complexity.

During geological time, the runoff from land areas gradually produced increasing salt concentrations until the present levels of salinity were reached. It is not surprising then that as animal life developed in great structural complexity during geological time, providing the oceans with jellyfish and whales at opposite extremes of size, the fluid volume and fluid content of ocean life should show corresponding variability. The water content of animals varies generally between 70 percent and 85 percent —in jellyfish it is as high as 96 percent. Because their cell membranes are selectively permeable, marine animals are

able to tolerate varying changes in salinity encountered in changing environments. Most marine invertebrates are as "salty" as their sea environment. The plasma membrane of marine protozoa is permeable to water, but, in higher forms, the membrane possesses varying tolerances to salt concentration. The ability to control their fluid volume is possessed in varying degrees by marine animals such as worms. Some marine vertebrates, such as hagfish, alter the concentration of their body fluids toward that of their surroundings by adjusting their blood chloride. Teleost fish regulate the salt concentration of their blood by eliminating excess salt through the gills or by resisting water loss through the kidneys. The salt content (and osmotic pressure) of marine mammals is slightly higher than that of terrestrial mammals but still well below that of seawater. In general, the body fluid of marine forms such as seals and killer whales has a relatively low salt content (about 1 percent sodium chloride), whereas that of marine invertebrates is practically equivalent to that of seawater.

Habitat zonation

Life in the open sea occupies the vast extent of the sea bottom (benthic zone) and overlying water (pelagic zone). Assuming that 200 metres (660 feet) below the surface is the upper limit of the habitat of the deep-sea fauna, the surface of the deep-sea region lies under about 92 percent of the total surface of the sea, equivalent to about two-thirds of the total surface of the globe. This three-dimensional region of faint light or complete darkness, with its strange creatures, possesses monotonous uniformity and stands apart from other biological realms on earth.

The lighted, shallow, populated pelagic zone extends upward and oceanward from the upper boundary of the continental slope. The dark, deep, sparsely populated pelagic zone extends outward and upward from the continental slope and the deep-sea floor and abyssal region (see Figure 2). The corresponding zones of bottom-dwelling, or benthic, organisms are the archibenthic (800–1,100 metres; 2,600–3,600 feet) and abyssobenthic (below 1,100 metres) zones. The latter zone is sparsely populated and contains some of the most unusual animals known.

CHARACTER OF OCEANIC POPULATIONS

Organisms of upper layers of the open sea (Figure 8) are exposed to the full daylight spectrum. Relatively small numbers of organisms inhabit the offshore oceanic waters. The more conspicuous components of the surface population are the Portuguese man-of-war (*Physalia*), the by-the-wind sailor (*Velella*), the purple storm snail (*Janthina*), and a nudibranch mollusk, *Gaucus*.

The transient animal life of the surface layers includes many fishes. The relatively permanent surface fauna, apart from such large forms as cetaceans (whales), turtles, and sea snakes, are of two main types: animals adapted entirely to a surface existence partly in air and partly in water and animals that inhabit the immediately subsurface layers. The strictly surface forms usually have floats of various types; some are definite organs of buoyancy, while others are formed of bubbles or trapped air. The only known open-ocean insects are water striders that live on the surface film, buoyed by air trapped in the hairs on the body.

The subsurface layers of the lighted open sea contain a great variety of the free-floating, or planktonic, forms, including many kinds of animal larvae. Apart from the blue coloration that is a most striking characteristic of plankton of the open ocean, there are few signs of adaptive characteristics special to this environment. The minute copepods of the family Pontellidae are capable of jumping out of the water to distances of about six inches (15.25 centimetres). The paper nautilus (*Argonauta*) is able to ride on the upper surface of the bells of certain jellyfish.

Seaweed buoyed by bladders is widespread in the oceans of the world and carries with it many associated animals that show adaptations of colour and form suitable to the habitat (interestingly, most of these animals are from near the shore, not oceanic).

Plankters such as diatoms and copepods migrate up and down in response to light. Illumination in the upper few metres is too bright for survival of most phytoplankton and zooplankton; hence, this vertical zone of distribution is sparsely populated. The zooplankton of the deeper but still lighted zone in which photosynthesis is active undergo diurnal migrations determined by food and light intensity. Vertical movements of plankton diatoms are related to variations in structure. Increase in spination (the distribution and arrangement of spines) increases the relative surface area and so retards sinking or flotation of the organism when its specific gravity differs from that of the water. Special types of plants adapted to a floating existence have evolved.

In the open sea, the chief grazers are the copepod crustaceans, semimicroscopic in size, that feed mainly on diatoms, dinoflagellates, and other micro-organisms. Protozoans and the larval stages of larger invertebrates also graze upon the phytoplankton. Some fish, such as herring and menhaden, feed on phytoplankton but are mainly coastal forms.

Life in dark open water

Bathypelagic animals inhabit the deep, dark waters of the ocean above the bottom (Figure 8). The deep-sea plankton contains no algae except for a few so-called olive-green cells. At 50 metres (160 feet) or less, from 3,000 to 100,000 plankton forms occur per litre. The number of species in the plankton does not decline with depth as strikingly as its number of individuals. Unusual size characterizes some of the deep-sea plankton forms. Microscopic swimming forms such as crustaceans differ significantly from their relatives of shallow seas in usually being luminous and sometimes blind. Other forms have enlarged eyes and very long appendages that assist in floating.

There is a great diversity of features and forms among the deep-sea fishes. Many of them are quite blind and small, from about 10 to 30 centimetres (four to 12 inches) in length. One of the chief characteristics of this deep zone is food scarcity. Another is the size and extension of locomotory appendages of crustaceans as well as fish. The scarcity of food near the sea bottom results in considerable modification of feeding structures: the great teeth, the enormous jaws, and the uncannily extensible stomachs characteristic of these deep-sea forms.

Organic material in various states of decomposition rains down from above and is consumed by middepth and abyssal detritus feeders. In fact, plant material as such probably never reaches the bottom but goes into solution to be reclaimed by bacteria and other consuming forms. The previously mentioned olive-green cells have maximum distribution below the euphotic zone and are believed to reclaim significant amounts of dissolved organic matter.

Deep-sea benthic animal life shows special features. The abyssal region extends from 100 to 1,500 metres to the lowest depths. It is a zone of muddy bottom characterized by low temperatures (usually between 1° and 2.5° C; 34° and 37° F) and by food scarcity. The population is sparse. Sessile animals (those that tend to remain in a given locale but that are not necessarily fixed) are characteristic and include sponges, stalked polyps, sea anemones, sea lilies, and rhizopod protozoans. Characteristic forms are the echinoderms, such as the sea urchins, some of which have abnormal body forms. There are giant isopods, 15 to 20 centimetres (six to eight inches) long, with large eyes. Some blind isopods have very long legs and tactile feelers that enable them to move over the muddy ooze and feel for their prey. Deep-sea crabs often have long, slender extremities, sometimes armed with great spikes that keep them from sinking into the ooze.

ADAPTATIONS TO MARINE CONDITIONS

Structural adaptations. Structural adaptations pertain to the rigid conditions of existence at great depths. They relate to acquisition of food, absence of light, stillness of waters, low temperature, and great pressure. The uniformity of these factors is more marked in the deeps than elsewhere in the ocean. The enormous pressure experi-

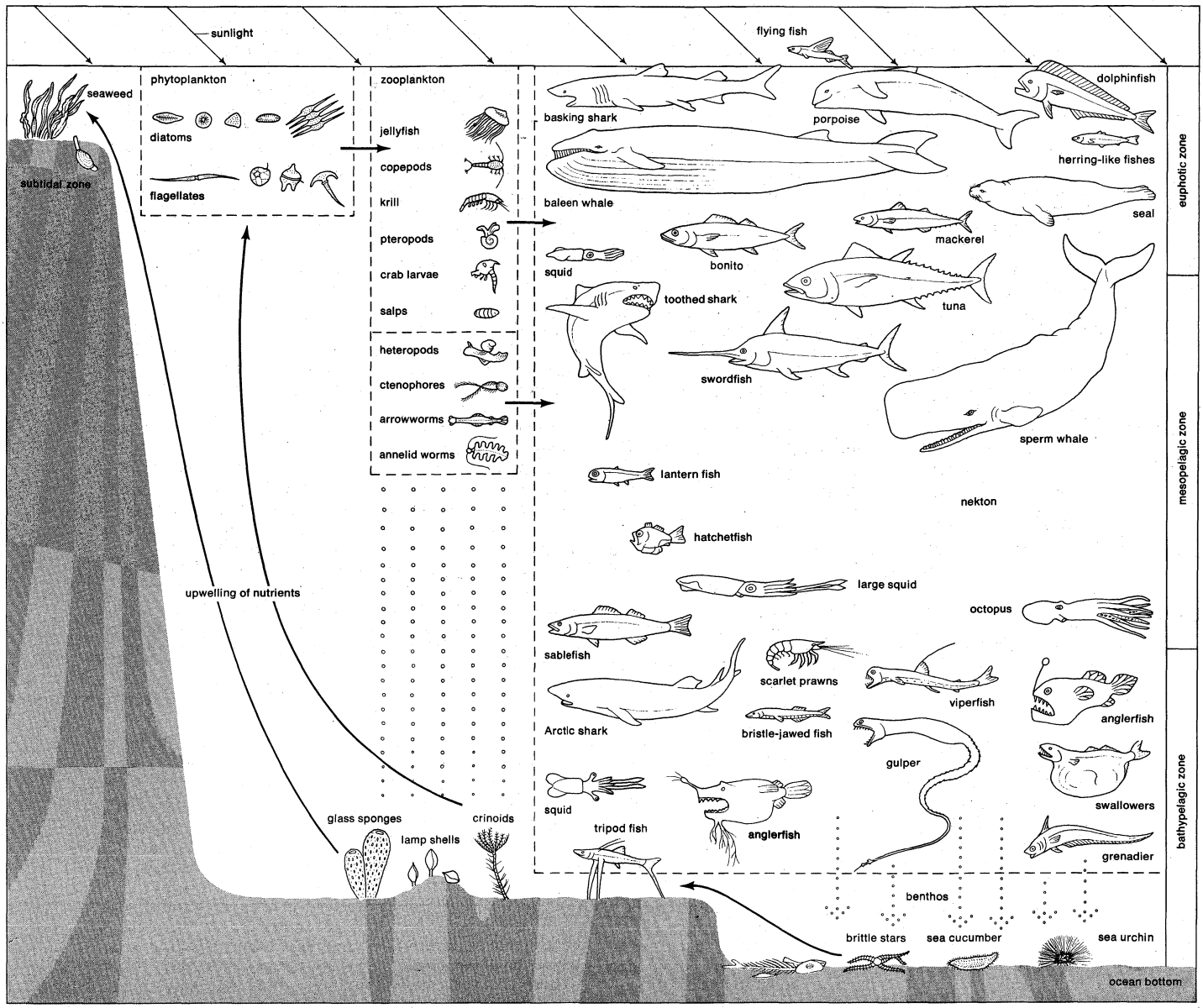


Figure 8: *The food chain in the marine environment.*
All life in the sea is dependent upon sunlight, which is directly converted into food by microscopic plantlike floating organisms (phytoplankton). Phytoplankton in turn is eaten by animal-like plankton (zooplankton). The plankton supports a succession of actively swimming predators (nekton). Organic debris rains down (dots) and provides food for animals at lower depths. Near-shore organisms are supported by land drainage. Coastal upwelling (upward arrows) provides the phytoplankton with nutrients released by decomposition of organic matter on the bottom. (Organisms not drawn to scale.)

enced by deep-sea organisms is not countered by any special development. Because of the balancing of external and internal pressures, water pressure is apparently not felt. Pressure changes can be endured if they are gradual enough to permit adjustment. A swim bladder, the hydrostatic apparatus so widespread in surface-living fishes, enables them to float or, by varying the gas content, to rise or sink in the water. A number of deep-sea fishes possess swim bladders. At a depth of 1,000 metres (3,300 feet), with a pressure of 100 atmospheres, the gas content of the swim bladder is compressed to 1/15 of the volume it would have at a ten-metre (33-foot) depth.

Aids to buoyancy

The most widespread means of reducing specific gravity among pelagic forms is through the absorption of large amounts of water in, for example, connective tissue, thus producing transparent gellike tissue found in cnidarians (coelenterates) snails, annelid worms, and pelagic cephalopods. Certain decapods are so transparent because of such tissue that print may be read through their bodies. Plankton fishes and eel larvae are also watery and transparent.

More effective than the absorption of water is the storage of lighter materials, such as low-salinity water, fat, or even air. The fluid of vacuoles of radiolarians and certain ctenophores has a lower specific gravity than seawater. The accumulation of fat is widely distributed among pelagic animals, thus lowering the general specific gravity. It is present in protozoans, such as radiolarians that contain oil drops, in crustaceans such as cladocerans and copepods, in some mollusks, and in many fishes that store food in the form of fat in their livers and that have oil drops in their eggs.

The most effective buoyancy means is the inclusion of air or other gases in the body. Jellyfish have air sacs filled with gas from a gas-producing gland. Certain cephalopods, such as *Nautilus*, have a shell containing air in chambers. The air sacs of fishes have already been discussed.

The relative immobility of deep water favours special animal developments. Fragility of the skeleton and other structural elements of the body is an example. Occasional development of giant forms—sponges and coelenterates

more than two metres high and sessile tunicates one metre high—are a product of perpetually still water. The deep-sea environment also produces the largest of the sea urchins, crabs, ostracods, isopods, and sea spiders. This unique environment also permits the degeneration of all supporting and strengthening elements of the body. Thus, the shells of unicellular radiolarians, skeletal spicules of sponges and echinoderms, shells of sea urchins, mussels, and snails, and bones of fish are often so thin and scarce that they could not provide support or protection in disturbed waters. Below 2,500 metres, the water is so deficient in calcium that that element dissolves readily in it; hence, it is difficult for animals to build up and maintain calcareous supporting structures.

The common factor of all pelagic animals is their independence of the bottom. They have obvious means, already discussed, for maintaining themselves in open water without sinking. In addition, shelled pelagic protozoans have thinner shells than bottom species. The calcium carbonate content of the shells is higher and the size of the pores is larger in the pelagic forms, thus favouring suspension. The shells of pelagic crustaceans are less calcified than those of their benthic relatives; they also have a higher fat content. Pelagic snails have delicate shells or none at all. The pelagic bivalve mollusk *Planktomya* has uncalcified shells. The skeleton of many pelagic fishes is weak, little calcified, or significantly reduced. Reduction of weight in pelagic copepods is achieved by depositing their eggs singly instead of carrying the egg sacs with them.

Effects of light and oxygen content. Lack of light in oceanic depths has important implications for physiological vitamin balance. For example, vitamin D, needed for building bones of vertebrates, can form and function only in the presence of light. Perhaps the lack of this vitamin accounts for the slightness and delicacies of bones of deep-sea fishes and for the markedly distorted forms of many bottom-living fishes.

Pelagic forms of the deep usually have an adequate supply of oxygen. Because of the low temperatures, these animals have low oxygen requirements. But the benthic occupants of the deep-water sediments require adaptations to meet the problem of low oxygen content. Some bottom dwellers can survive for relatively long periods without oxygen in the liquid environment. Respiratory pigments may provide an oxygen reserve during short periods in oxygen-deficient depths. The shipworm (*Teredo*), which may be found at great depths, makes use of glycogen as an oxygen supply, the glycogen comprising half the dry weight of its tissues. The biology and physiology of inhabitants of such anaerobic sediments are specially adapted to their environment. Besides a biochemical adaptation through the breakdown of glycogen, some animals cease to move. Others have morphological features that enable them to supply aerated water to vital respiratory organs.

Luminescence of fishes

Bioluminescence, or the production of light by organisms, is especially characteristic of the deep-sea population. The bottom of the abyssal swarms with light producers, including coelenterates, echinoderms, and annelid worms. As many as 44 percent of the fishes of depths below 900 metres (3,000 feet) are light producers. The light organs of deep-sea fishes vary in size and form. Their source is functionally metamorphosed skin glands or trapped luminous bacteria. Bioluminescence serves to lure prey, to frighten enemies, to enable the sexes to find each other, and to enable species that move in schools to keep together (see BIOLUMINESCENCE).

Maximum diatom abundance in the open sea is at around 30 metres (100 feet) depth. The daytime concentration of upper zooplankton is at about 125 metres (410 feet), the level varying with location and light intensity. Most of the zooplankton ascend to the surface at night. The echo phenomenon known as the scattering layer is caused by plankton or nekton or both throughout the oceans of the world at from about 400 to 800 metres (1,300 to 2,600 feet). Some of the animal components of this layer undergo diurnal migration of 400 metres or more.

The distribution of euphausid crustaceans has been explained in terms of temperature and salinity. In the Antarctic, there appears to be a regular cycle in which the immatures and adults are carried into lower latitudes by shallow water movements, and the eggs are returned to high latitudes by deeper, southward flowing water movements.

Associations. Mutualism is illustrated by the widespread cleaning of fish surfaces of undesirable parasites by associated organisms in search of food. Many species of fish and shrimps, as well as other organisms, clean parasites from other animals. The host is relieved of irritation while the cleaning species gains food. An example is the Spanish hogfish (*Bodianus rufus*), which swims into the mouth of the barracuda and forages among its teeth for food.

The advantage derived by a commensal (species associated with the host) may be the provision of a resting place, shelter, or transport and often of food. Commensals in somewhat permanent contact with their hosts include a great variety of animals and plants that simply require physical support. Certain barnacles are found only on the backs of whales, where they benefit by being transported great distances. Pilot fish eat leftovers from the sharks, to which they attach by means of a suction disk. Whiting and man-of-war fish shelter among the tentacles of jellyfish. Tiny gobies hide under the gill covers of larger fish.

Fish serve as hosts to a large number of parasites, including protozoans and many kinds of worms. They are also plagued by fish lice or copepods. In many deep-sea angler fish, the male lives as a tiny permanent parasite upon the body of the female and obtains his entire nourishment from her blood supply. These species have solved the problem of finding mates in the inky blackness of the deep sea.

PRODUCTIVITY OF MARINE COMMUNITIES

The productivity of the oceans may be judged by the biological oxygen consumption of the water or by nutrient concentration, including soluble inorganic phosphates, nitrates, nitrites, ammonium salts, and silicates. These minerals are consumed in the upper lighted layers, the inorganic phosphorus and nitrogen are regenerated by bacterial decomposition or by dissolved organic debris, and silicates are reformed by the dissolution of planktonic tests. The return of nutrients to the upper photosynthetic layers is accomplished by upwelling and the action of currents.

Nutrient layers

There are vertical-distributional strata that may differ markedly in nutrient concentrations in any one ocean. Four layers are recognized: (1) the surface, well-mixed, euphotic layer of relatively low concentration that varies little with depth; (2) the layer in which the concentrations increase significantly with depth; (3) the layer of maximum concentration, which has a range of between about 500 and 1,600 metres; and (4) the thick bottom layers of relatively uniform phosphate and nitrate concentrations and in which silicate content increases significantly with depth.

Differences in the nutrient concentrations of the oceans depend largely on the composition of the deep water masses at their origin and upon subsequent changes induced by circulation and biological processes. Meaningful comparisons of oceans are thus difficult to make. Available data for the Indian Ocean suggest annual biological oxygen-consumption levels of around 0.45 millilitre of oxygen per litre of seawater in the upper 400 metres (1,300 feet) in the Antarctic shelf area and corresponding levels in the equatorial region of about 1.5 millilitres per litre. The North Indian bottom water levels below 2,000 metres (6,600 feet) average about 0.04 millilitre per litre, and at depths between 600 and 1,200 metres (2,000 and 4,000 feet) biological oxygen consumption levels have been found to range from 1.5 to 2.0 millilitres per litre.

The standing crop, or abundance, of zooplankton in the Antarctic waters in summer has been found to be about ten times greater than that in the tropical Atlantic waters.

At low temperatures, a larger total number of organisms can be supported on the same amount of food; and, in colder waters, nutrients may be supplied more rapidly in relation to their use. In tropical waters with great upwelling of nutrients from the bottom layers, a large standing crop of plankton exists and can support a productive food chain.

Probably the total marine biomass is far greater than the combined biomass of land and freshwater. Standing-crop data are commonly reported as milligrams of carbon (C) per cubic millimetre (or other unit of volume). The range is commonly ten to 1,000 milligrams of carbon per cubic metre. Values are usually derived from measurement of phytoplankton pigments, especially chlorophyll *a*. Values for open-ocean areas may range around five to 150 milligrams of carbon per cubic metre.

Distribution of the yield

The entire oceanic productivity falls in the range 1.6 to 15.5×10^{10} tons of carbon per year, compared to 1.9×10^{10} tons for the land. About 58×10^6 tons of fish and shellfish are produced annually by the oceans, an amount believed to represent only about 0.03 percent of the total amount of organic material estimated to be produced annually in the sea. Most of the fishery yield comes from waters over the continental shelves and within the 200-metre depth range where nutrient concentrations are high.

This productive area represents only about 3 percent of the total ocean surface. The deep, open-ocean areas are comparable to continental deserts. The reduction in biomass from shallow coastal waters to oceanic depths is about 10^6 for a benthos (organism living on or near the bottom); and for the plankton it approaches 10^4. Undoubtedly, this very low rate is associated with and indeed reflects the very small amount of detritus and other food material reaching the bottom at great depths. The low temperatures of the ocean deeps reduces respiratory rates and growth rates. Maintenance demands in terms of food supply are correspondingly low. In summary, the pelagic and benthic zones of the deep open ocean are sparsely populated and the least productive of all oceanic areas.

Instances of productivity variations among the Atlantic, Pacific, Indian, and Arctic oceans are associated with latitudinal temperature differences and upwelling of deep water rich in nutrients. The maximum–minimum variations in standing crop of zooplankton during the year at a particular station in the open ocean may amount to a factor of 2 or 3. With approach to continental regions, the productivity of the water is greater, and the seasonal variations in volume of dominant zooplankton forms is much greater, perhaps seven or eight times richer. Temperate areas of the Atlantic are much richer in macroplankton numbers than the sparsely populated waters of the Sargasso Sea, which are about four times less populated than continental-slope areas. The standing crop of zooplankton in cold northern waters is sometimes over eight times that of tropical waters, and there is believed to be a possible tenfold to twentyfold increase in zooplankton from the winter minimum to the late summer maximum. (C.N.)

IV. Economic aspects of oceans and seas

The sea is generally accepted by scientists as the place where life began on earth. Without the sea, life as it is known today could not exist. Among other functions, it acts as a great heat reservoir, levelling the temperature extremes that would otherwise prevail over the earth and expand the desert areas. The oceans provide the least expensive form of transportation known to man, and the margins of the sea serve as one of his major sites of recreation. The sea is a major source of food and a dumping ground for many wastes. And the sea is a major potential source of protein, minerals, and power, all of which are required in ever-increasing quantities by all industrialized societies.

In this section, the principal economic aspects of the sea —both realized and potential—in the general areas of transport and communications, food and water, energy resources, and waste disposal are dealt with.

TRANSPORT AND COMMUNICATIONS

From the beginning of recorded history, man has used the sea as a means of transport, first for himself, and then as a means of distributing products throughout the world. The bulk of the tonnage of products transported throughout the world today is moved in ocean vessels. The size of these vessels ranges from small boats capable of carrying a few tons to bulk carriers capable of transporting almost 500,000 tons of oil. The cost of transporting goods on the ocean depends on the product, the form of shipment, and the type of vessel. Probably the cheapest form of transportation known to man is that of the great oil carriers in which a ton of oil may be shipped for an average cost of hundredths of a cent per ton mile. This cost is about 100 times cheaper than shipping such material on land, if the use of pipelines is excluded. As the per capita consumption of materials increases, the outlook for marine transportation is one of ever-increasing tonnages and size of carrying vessels, not only in conventional vessels but also in ground-effect vessels (already in commercial operation in Europe) and in bulk-carrying submarines to pass beneath the polar ice cap.

Since the laying of the trans-Atlantic cable in the 19th century, the oceans have served as a major means of communication between continents and islands. Hundreds of sea-floor cables connect all major centres of world population. With the development of satellite communications, ocean-floor cables as a means of communication may tend to decrease in importance, but they will continue to carry information for many decades to come.

In addition to communications, cable and pipes laid on the ocean floor carry electrical energy, oil, and other commodities in many parts of the world.

FOOD AND WATER

Fishing. Man extracts about 60,000,000 tons of food from the ocean annually by fishing. The food-producing potential of the sea, however, is several hundreds of times the present rate of production. The methods by which man takes food from the sea are inefficient, and the fact that he takes only certain choice species of fish makes fishing in the ocean doubly inefficient. The development of a process to extract protein concentrates from all types of fish might rectify this shortcoming. This protein concentrate could be stored, transported, and utilized very efficiently. It has been estimated that the daily protein requirements for a human being can be produced from fish for less than a penny. It is estimated that, by efficient harvesting of all the fish of the sea, the ocean could produce a sustained yield of about 2,000,000,000 tons of food annually, but it must be noted that such utilization requires, in many instances, a change in human attitudes. In many parts of the world, wheat has proved to be an unacceptable substitute for rice where rice is a customary, integral part of the diet of the inhabitants.

Fish-farming

The Japanese have instituted a substantial program of continental-shelf-development studies to develop the eventual farming of adequate supplies of fish and edible plants there. While Japan has farmed oysters in its oceanic bays for many years, its fishermen have been active around the world's oceans in the manner of those from other countries. With that activity seemingly at a point of diminishing returns, Japan has thus turned to home waters.

Fish need not be fenced in to be farmed. They will stay where the food supply is. By creating an artificial food supply in a given location, fish can be kept where they are wanted. Creating a food supply for the fish need not mean the adding of fish nutrients to the sea but the development of some means of mixing the nutrient-rich bottom layers of water in the ocean with the life-rich upper layers of water. Wherever a natural upwelling of the bottom layers of water occurs, such as off Peru, a tremendous fish population also is found. By encouraging such an upwelling artificially, the fish population could be greatly increased at some more convenient point rather than at the location nature has provided. Developing this technique could probably increase the potential yield of the ocean by ten times or so over the present potential

productive capacity of the ocean. The energy for sustaining this upwelling in the ocean can be produced by several sources; for example, by nuclear reactors on the bottom of the ocean. The development of this or any other means of producing upwelling, however, will probably await economic necessity, if not actual crisis; reactors on the sea floor do not appear to be likely in the immediate future.

Desalination. Although springs of freshwater issuing on the ocean floor are known to occur in water depths as great as 1,000 metres (3,300 feet), these springs never will prove to be a major source of freshwater for the world. Of greater interest is the prospect of desalting seawater itself. Throughout the world, hundreds of desalination units, producing from a few thousand to 10,000,000 or more gallons per day, already are in operation. In general, the desalination plants in production are in areas where the population has outstripped the onshore water supply and where high-cost desalinated water can be afforded. This situation tends to arise in coastal-desert areas, or on densely populated islands because the costs of pumping water through pipelines to interior areas would add prohibitively to the basic cost—at the sites of desalination.

A population usually can afford to pay about ten times as much for water for domestic purposes as it does for agricultural water. Proposals for large-scale nuclear desalination facilities, when constructed, promise to lower the cost of desalinated water to 10 cents to 30 cents per 1,000 gallons—at the desalination sites—a price that all domestic users, most industries, and a few agricultural enterprises can afford.

The bulk of the water produced from seawater is produced by some form of evaporation and condensation. Although the principle of this technique is quite simple, the mechanics of achieving high efficiencies can become quite complicated. Superheated water and multiple evaporation and condensation units, operating at varying temperatures and pressures, are employed in a number of these facilities. The choice of construction materials is quite important, because the brines produced in extracting pure water can be corrosive.

Other processes under consideration as potential economic methods of desalting seawater are freezing, reverse osmosis, ionic processes, electrodialysis, and techniques that change the physical or chemical properties of water itself so that it can be separated from the salts in seawater. In the future, it can be expected that the ocean will become an increasingly important source of freshwater. If production and transportation costs can be lowered sufficiently, it may be possible to produce freshwater to irrigate large areas that border the oceans in many parts of the world.

ENERGY RESOURCES

There are a number of recognized techniques by which energy can be extracted from the sea. The major problem in taking energy resources from the sea is that they tend to be diffused over a large lateral area. A point-concentration energy source is necessary if it is to be exploited economically.

Power generation. Although energy presently is extracted from the tides (*q.v.*) of the ocean, it is unlikely that man will develop a technique for extracting efficiently large amounts of energy from such sources as the waves of the sea. Another potential power source in the ocean, however, is the temperature differential between the surface layers and the lower layers of water; it can be as large as 50 degrees over vertical distances of as little as 300 feet (90 metres) in some areas of the ocean. For many years, the French have experimented with techniques of using this temperature differential to generate electricity. The processes they used were a technical success; but they were not economic, mainly because the plant was located on land, and most of the energy obtained from this system was used in pumping the seawater into and out of the plant.

Utilization of temperature differentials

A group in the United States has developed a system whereby the energy can be extracted by a floating power plant that eliminates the pumping problem. Propane or some similar fluid that will boil at the temperature of the surface layer of water is used, the gaseous phase then being released to a turbogenerator to manufacture the power. After the gas is condensed to a liquid in the cooler, lower layers of water, it is pumped to a surface boiler to be recycled. To be economically practical, the temperature differential should be several tens of degrees, and major markets for the energy created should not be more than a few hundred miles from the power plant. The areas of the ocean in which it would be most efficient to operate such a plant, normally in the tropics, are not near major markets.

Some techniques have been developed in the 1970s for handling liquid hydrogen and liquid oxygen on a large scale. If power generated in remote areas of the ocean were used to reduce seawater to hydrogen and oxygen, these gases might be utilized economically in shoreside thermal power plants. This procedure is, admittedly hypothetical. Followed to its logical though futuristic conclusion, it suggests that oceanic thermal power plants may someday be used to power the artificial upwelling previously mentioned.

A conservative estimate of the energy resources presently available in the ocean is several thousands of years, at present rates of world energy consumption. Even more significant, however, is the fact, that because this thermal differential is generated by the sun, it is renewing itself about 100 times as fast as it could be used to provide the whole world with its present total energy consumption. Man could take from the sea about 200,000,000 megawatt hours of energy each day. The United States presently consumes about 1,000,000 megawatt hours of power each day; thus, the sea can produce power at a rate of about 200 times present consumption rate of the United States. There are, of course, no fuel costs involved in generating power in the sea, and the estimated plant operational and capital costs indicate an overall power-production cost substantially lower than that of power generated from conventional sources.

Petroleum. In the mid-1950s, the production of oil and gas from oceanic areas was negligible. By the 1970s, about 6,000,000 barrels per day, or about 16 percent of the world's production, came from ocean wells; it is predicted that by 1980 the offshore areas will yield about 23,000,000 barrels per day, or about 35 percent of the 1980 world production. About 300 offshore drilling and production rigs were at work in the early 1970s, at more than 60 offshore locations throughout the world drilling, completing, and maintaining offshore oil wells (see PETROLEUM; NATURAL GAS). Estimates have placed the potential offshore oil resources at about 2,000,000,000,000 barrels, or about half of the presently known onshore potential oil sources.

It was once thought that only the continental-shelf areas contained potential petroleum resources, but discoveries of oil deposits in deeper waters of the Gulf of Mexico (about 3,000 to 4,000 metres [10,000 to 13,000 feet]) have led to a revision of this idea. It is now believed that the continental slopes and neighbouring ocean-floor areas will contain large oil deposits, thus enhancing potential petroleum reserves of the ocean bottom.

Minerals. The rivers of the world dump billions of tons of material into the sea each year. Sea-floor springs and volcanic eruptions also add many millions of tons of elements. Even the winds contribute solid materials to the sea in appreciable quantities. Most of these sediments rapidly settle to the sea floor in nearshore areas, in some cases forming potentially valuable placer mineral deposits. The dissolved load of the rivers, however, mixes with seawater and is gradually dispersed over the total oceanic envelope of the earth. Because of the nature of the minerals and their mode of formation, it is convenient to consider the occurrence of sea deposits in several environments, namely marine beaches, seawater, continental shelves, sub-sea-floor consolidated rocks, and the marine sediments (*q.v.*) of the deep-sea floor. Minerals are mined from all of these environments except for the deep-sea-floor area, which was only in the 1970s recog-

nized as a repository for mineral deposits of unbelievable extent and significant economic value.

Placer and nearshore deposits

Minerals that resist the chemical and mechanical processes of erosion in nature and that possess a density greater than that of the earth's common minerals have a tendency to concentrate in gravity deposits known as placers. During the ice ages (about 10,000 to 2,500,000 years ago), sea level was appreciably lowered as the ocean water was transferred to the continental glaciers. Because of the cyclical nature of the ice ages and the intervening warm periods, series of beaches were formed in nearshore areas both above and below present sea level. Also, when sea level was lowered in past ages, the streams that now flow into the sea coursed much further seaward, carrying placer minerals to be deposited in channels that are now submerged. With geophysical-exploration techniques, these channels and beaches can be easily delineated, even though these features are totally covered by Recent or Holocene sediments; that is, those deposited during the last 10,000 years.

Sand and gravel, mined from a number of offshore locations around the world, generally with hydraulic dredges, are used mainly for construction purposes or for beach replenishment or nearshore fills.

Sulfur, which is taken from salt domes in the Gulf of Mexico, is mined by a process in which pressurized hot water is pumped into the sulfur-containing cap of the dome, melting the sulfur and forcing it to the surface. Compressed air is also used to pump sulfur to the surface; the still-molten sulfur is then pumped ashore through insulated pipelines.

Of considerable interest are the sea-floor phosphorite deposits on the coastal shelves of many nations. The phosphorite off California occurs as nodules that vary in shape from flat slabs, several feet across, to small spherical forms termed oölites. The nodules commonly are found as a single layer at the surface of coarse-grained sediments. Phosphorite composition from the California offshore area is surprisingly uniform and contains potentially economically attractive amounts of phosphorus.

Another type of phosphate deposit has been found off the west coast of Mexico, namely, a fine-grained, unconsolidated deposit in about 50 metres (160 feet) of water. It contains as much as 40 percent apatite (common phosphate mineral), and it is speculated that there is as much as 20,000,000,000 tons of recoverable phosphate rock in this deposit.

Deep-sea deposits

Mineral deposits of monumental size and potential economic significance are found in the deep-sea areas of the ocean. Minerals formed in the deep sea are frequently found in high concentrations, because there is relatively little clastic material generated in these areas to dilute the chemical precipitates.

An estimated 10^{16} tons of calcareous oozes, formed by the deposition of calcareous shells and skeletons of planktonic (floating) organisms, cover some 130,000,000 square kilometres (50,000,000 square miles) of the ocean floor. In a few instances, these oozes, which occur within a few hundred miles of most nations bordering the sea, are almost pure calcium carbonate; but they often show a composition similar to that of the limestones used in the manufacture of portland cement.

Covering about 39,000,000 square kilometres (15,000,000 square miles) of the ocean floor in great bands across the northern and southern ends of the Pacific Ocean and across the southern ends of the Indian and Atlantic oceans are other oozes, consisting of the siliceous shells and skeletons of plankton animals and plants. Normally these oozes could serve in most of the applications for which diatomaceous earth is used, for fire and sound insulation, for lightweight concrete formulations, as filters, and as soil conditioners.

An estimated 10^{16} tons of red clay covers about 104,000,000 square kilometres (40,000,000 square miles) of the ocean floor. Although compositional analyses are not particularly exciting, red clay may possess some value as a raw material in the clay-products industries, or it may serve as a source of metals in the future. The average assay for alumina is about 15 percent, but red clays from specific locations have assayed as high as 25 percent alumina; copper contents as high as 0.20 percent also have been found. A few hundredths of a percent of such metals as nickel and cobalt and a percent or so of manganese also are generally present in a micronodular fraction of the clays and probably can be separated and concentrated from the other materials by screening or by some other physical method.

Underlying the hot brines in the Red Sea are basins containing metal-rich sediments that potentially may prove to be one of the great metal-deposit discoveries of all time. It has been estimated that the largest of several such pools, the Atlantis II Deep, contains over $2,000,000,000 worth of copper, zinc, silver, and gold values in relatively high grades. These pools lie in about 2,000 metres (6,600 feet) of water, midway between The Sudan and the Arabian Peninsula. Because of their gel-like nature, pumping these sediments to the surface should prove relatively uncomplicated. These deposits are forming today under present geochemical conditions, and they are similar in character to certain major ore deposits on land.

Manganese nodules

From an economic standpoint, the most interesting oceanic sediments are manganese nodules—small, black to brown, friable lumps found to be widely distributed throughout the three major oceans in the late 19th century by the famous "Challenger" and "Albatross" expeditions.

Many theories have been proposed to account for the formation of manganese nodules, the best probably being that the ocean is saturated at its present state of acidity–alkalinity in iron and manganese. For this reason, these elements precipitate as colloidal particles that gradually increase in size and filter down to the sea floor. Colloids of manganese and iron oxides collect many metals and bear an electrical charge; they tend to agglomerate as nodules at the sea floor rather than settle as particles in the general sediments.

An estimated 1,500,000,000,000 tons of manganese nodules are on the Pacific Ocean floor alone, and they are estimated to be forming at an annual rate of about 10,000,000 tons. Averaging about four centimetres (slightly less than two inches) in diameter (Figure 4) and found in concentrations as high as 100,000 tons per square mile, these manganese nodules contain as much as 2.5 percent copper, 2.0 percent nickel, 0.2 percent cobalt, and 35 percent manganese. In some deposits, the content of cobalt and manganese is as high as 2.5 percent and 50 percent, respectively. Such concentrations would be considered high-grade ores if found on land, and, because of the large horizontal extent of the deposit, they are a potential source of many important industrial metals. Table 7 lists some statistics concerning the amount of reserves and rate of accumulation of various metals in the nodules.

Relatively simple mechanical cable bucket or hydraulic dredges with submerged motors and pumps can effect the mining of the nodules at rates as high as 10,000 to 15,000 tons per day, from depths as great as 6,000 metres (20,000 feet). The estimated costs to mine and process the nodules indicate that copper, nickel, cobalt, and other metals can be economically produced from this source.

WASTE DISPOSAL

One of the least known but most important uses of the sea is as a garbage dump. In the past, the ocean was capable of assimilating the wastes of societies without noticeable pollution effects. Because society is becoming increasingly affluent, however, the dumping of wastes into the ocean is overwhelming the ocean's capacity to digest them. Great marginal areas of the world's oceans have been heavily polluted by many forms of man's wastes, the most obvious of which is sewage. Less apparent forms of pollution are atomic wastes, chemical wastes, trash and oil spills from vessels, oil spills from offshore wells, and heat. Great power plants are generally located along coastlines to reduce the costs involved in cooling their condensers by water-circulation systems. Although the

Table 7: Reserves of Metals in Manganese Nodules of the Pacific Ocean

	amount of element in nodules (billions of tons)	reserves in nodules at consumption rate of 1964 (years)	approximate world land reserves of element (years)	ratio of: (reserves in nodules) divided by (reserves on land)	rate of U.S. consumption of element in 1964 (millions of tons per year)	rate of accumulation of element in nodules (millions of tons per year)	ratio of: (rate of accumulation) divided by (rate of U.S. consumption)
Magnesium	25	600,000	*	—	0.04	0.18	4.5
Aluminum	43	20,000	100	200	2.0	0.3	0.15
Titanium	9.9	2,000,000	*	—	0.3	0.069	0.23
Vanadium	0.8	400,000	*	—	0.002	0.0056	2.8
Manganese	358	400,000	100	4,000	0.8	2.5	3.0
Iron	207	2,000	500	4	100	1.4	0.01
Cobalt	5.2	200,000	40	5,000	0.008	0.036	4.5
Nickel	14.7	150,000	100	1,500	0.11	0.102	1.0
Copper	7.9	6,000	40	150	1.2	0.055	0.05
Zinc	0.7	1,000	100	10	0.9	0.0048	0.005
Gallium	0.015	150,000	—	—	0.0001	0.0001	1.0
Zirconium	0.93	100,000	100	1,000	0.0013	0.0065	5.0
Molybdenum	0.77	30,000	500	60	0.025	0.0054	0.2
Silver	0.001	100	100	1	0.006	0.00003	0.005
Lead	1.3	1,000	40	50	1.0	0.009	0.009

*Present reserves so large as to be essentially unlimited at present rates of consumption.
Source: After Mero, 1965.

whole of the ocean never will be affected by the waste heat dissipated by these plants, great environmental effects can be caused in the immediate area of the power-plant outfall. Crude oil from leaking offshore oil wells, from sinking tankers, and from bilges pumped by practically all vessels using oil for power have polluted countless beaches throughout the world. Lead, which is highly toxic to life, is being added to the ocean in great quantities by fallout from internal-combustion exhausts, which enters the earth's atmosphere and ultimately precipitates and reaches the seas. Pollution by mercury, another highly toxic substance, became of great concern around 1970 after analyses revealed its presence in marine and freshwater fish. The effects of dumping these wastes into the ocean are becoming great enough to pollute the entire ocean in some manner. Man can continue to use the ocean as a dumping ground for his wastes, but he must control the form and point of dumping into the ocean so that the pollution effects are reduced to an acceptable level. (J.L.M.)

BIBLIOGRAPHY

Physical and chemical aspects of oceans and seas: There are many works that provide general coverage of oceanography but the interested reader should not fail to consult the classics by M.F. MAURY, *The Physical Geography of the Sea* (1855, reprinted 1963); J. MURRAY and J. HJORT, *The Depths of the Ocean* (1912); and O. KRUMMEL, *Handbuch der Ozeanographie*, 2 vol. (1907–11). Recent texts on oceanography that are both readable and comprehensive include W.S. VON ARX, *An Introduction to Physical Oceanography* (1962); G.L. PICKARD, *Descriptive Physical Oceanography*, 2nd ed. (1966); G. NEUMANN and W.J. PIERSON, JR., *Principles of Physical Oceanography* (1966); P. GROEN, *De wateren der wereldzee*, 2nd ed. (1961; Eng. trans., *The Waters of the Sea*, 1967); and P.K. WEYL, *Oceanography: An Introduction to the Marine Environment* (1970). In this same vein, no list of general texts would be complete without citing H.U. SVERDRUP, M.W. JOHNSON, and R.H. FLEMING, *The Oceans* (1942), which was in many ways the forerunner of all modern general texts; it covers biological aspects as well as the physics and chemistry of the oceans. More mathematical treatment of oceanography is provided by J. PROUDMAN, *Dynamical Oceanography* (1953); A. DEFANT, *Physical Oceanography*, 2 vol. (1961); and the collection of papers on selected topics edited by M.N. HILL, *The Sea: Ideas and Observations on Progress in the Study of the Seas*, 3 vol. (1962–63). Also worthy of mention here is the Soviet oceanographic atlas, Mopckon atrac, 3 vol. (1950–58); and *the Encyclopedia of Oceanography*, ed. by R.W. FAIRBRIDGE (1966), which contains many readable articles on specific water bodies as well as on general oceanographic topics.

Life in the open sea: W.V. CROMBIE, *The Living World of the Sea* (1967), a popular book on marine subjects of wide general interest; K. GUNTHER and K. DECKERT, *Wunderwelt der Tiefsee* (1950; Eng. trans., *Creatures of the Deep Sea*, 1956), a classic work on deep, underwater animals; B.E. and N. MACGINITIE, *Natural History of Marine Animals*, 2nd ed. (1968), an excellent, generalized treatment of the ecology of widely representative marine animals stressing those from the Pacific Coast of North America; R.C. MILLER, *The Sea* (1966), a fascinating book on marine life, habitats, and phenomena written in an exceptional style with a unique format; H.B. MOORE, *Marine Ecology* (1958), authoritative book on marine ecology suitable for use as a college text; J.A. COLIN NICOL, *The Biology of Marine Animals*, 2nd ed. (1967), a scholarly exhaustive treatment of physiological topics and mechanisms pertaining to marine organisms; F.S. RUSSELL and C.M. YONGE, *The Seas*, 3rd ed. (1963), a well-written popular account of life and habitats in the seas of the world, their importance and use to man.

Economic aspects of oceans and seas: J.H. and J.H. ANDERSON, "Power from the Sun by Way of the Sea?" *Power* 109:64–65 (1965), the original paper describing a method of extracting power from the thermal differential of the ocean using a floating power plant; J.E. BARDACH, "Aquaculture," *Science*, 161:1098–1106 (1968), a comprehensive review on growing food in bodies of water; J.L. BISCHOFF and F.T. MANHEIM, "Economic Potential of the Red Sea Heavy Metal Deposits," in E.T. DEGANS and D.A. ROSS (eds.), *Hot Brines and Recent Heavy Metal Deposits in the Red Sea*, pp. 535–541 (1969), a paper dealing with the economic potential of the heavy metal deposits underlying the Red Sea brines; R.H. CHARLIER, "Harvesting the Energies of the Ocean," *MTS Journal*, 3:13–32 (1969), a comprehensive paper dealing with the techniques of extracting power from the tides and waves of the ocean; J.L. MERO, *The Mineral Resources of the Sea* (1965); J. MURRAY and A.F. RENARD, *Report on Deep-Sea Deposits Based on the Specimens Collected During the Voyage of H.M.S. Challenger* (1891), two classic references on the mineralogy of deep-sea deposits; D.F. OTHMER, "Desalination of Seawater," in F.C. FIRTH (ed.), *The Encyclopedia of Marine Resources*, pp. 162–169 (1969), a review article on the techniques of desalting seawater; L.A. WALFORD, *Living Resources of the Sea* (1958), a classic; L.G. WEEKS, "The Ocean's Resources," *Offshore*, 28:39–48, 87–88 (1968), a comprehensive review of the petroleum resources of the sea.

(P.Gr./C.N./J.L.M.)

Ockham, William of

William of Ockham, the most influential philosopher of the 14th century and a controversial figure because of a bitter dispute with the Pope, was a late Scholastic thinker who is regarded as the founder of Nominalism—the school of thought that denies that universal concepts such as "father" have any reality apart from the individual things signified by the universal or general term. He is considered by some, though not all, historians as a forerunner of Martin Luther, the great 16th-century Reformer.

Early life

Ockham was an Englishman, born probably in Surrey about 1285. It seems that he was still a youngster when he entered the Franciscan order. At that time a central issue of concern in the order and a main topic of debate in the church was the interpretation of the rule of life composed by St. Francis of Assisi concerning the strictness of the poverty that should be practiced within the order. Ockham's early schooling in a Franciscan convent concentrated on the study of logic; throughout his career, his

interest in logic never waned, because he regarded the science of terms as fundamental and indispensable for practicing all the sciences of things, including God, the world, and ecclesiastical or civil institutions; in all his disputes logic was destined to serve as his chief weapon against adversaries. After his early training, Ockham took the traditional course of theological studies at Oxford University and apparently between 1317 and 1319 lectured on the *Sentences* of Peter Lombard—a 12th-century theologian whose work was the official textbook of theology in the universities until the 16th century. His lectures were also set down in written commentaries, of which the commentary on Book I of the *Sentences* (a commentary known as *Ordinatio*) was actually written by Ockham himself. His opinions aroused strong opposition from members of the theological faculty, and he left the university without obtaining his master's degree in theology; he thus remained, academically speaking, an undergraduate—known as an *inceptor* ("beginner") in Oxonian language or, to use a Parisian equivalent, a *baccalaureus formatus.* Ockham continued his academic career, apparently in English convents, simultaneously studying points of logic in natural philosophy and participating in theological debates. When he left his country for Avignon, France, in the autumn of 1324 at the Pope's request, he was acquainted with a university environment shaken not only by disputes but also by the challenging of authority: that of the bishops in doctrinal matters and that of the chancellor of the university, John Lutterell, who was dismissed from his post in 1322 at the demand of the teaching staff.

However abstract and impersonal the style of Ockham's writings may be, they reveal at least two aspects of Ockham's intellectual and spiritual attitude: he was a theologian-logician (*theologicus logicus* is Luther's term). On the one hand, with his passion for logic he insisted on evaluations that are severely rational, on distinctions between the necessary and the incidental and differentiation between evidence and degrees of probability—an insistence that places great trust in man's natural reason and his human nature. On the other hand, as a theologian he referred to the primary importance of the God of the creed whose omnipotence determines the gratuitous salvation of men; God's saving action consists of giving without any obligation and is already profusely demonstrated in the creation of nature. The medieval rule of economy, that "plurality should not be assumed without necessity," has come to be known as "Ockham's razor"; the principle was used by Ockham to eliminate many entities that had been devised, especially by the Scholastic philosophers, to explain reality.

Treatise to John XXII

Ockham met John Lutterell again at Avignon; in a treatise addressed to Pope John XXII, the former chancellor of Oxford denounced Ockham's teaching on the *Sentences,* extracting from it 56 propositions that he showed to be in serious error. Lutterell then became a member of a committee of six theologians that produced two successive reports based on extracts from Ockham's commentary, of which the second was more severely critical. Ockham, however, presented to the Pope another copy of the *Ordinatio* in which he had made some corrections. It appeared that he would be condemned for his teaching, but the condemnation never came.

At the convent where he resided in Avignon, Ockham met Bonagratia of Bergamo, a doctor of civil and canon law who was being persecuted for his opposition to John XXII on the problem of Franciscan poverty. On December 1, 1327, the Franciscan general Michael of Cesena arrived in Avignon and stayed at the same convent; he had been summoned by the Pope in connection with the dispute over property. They were at odds over the theoretical problem of whether Christ and his Apostles had owned the goods they used; that is, whether they had renounced all ownership (both private and corporate), the right of property and the right to the use of property. Michael maintained that because Christ and his Apostles had renounced all ownership and all rights to property, the Franciscans were justified in attempting to do the same thing. The relations between John and Michael grew steadily worse, to such an extent that, on May 26, 1328, Michael fled from Avignon accompanied by Bonagratia and William. Ockham, who was already a witness in an appeal secretly drafted by Michael on April 13, publicly endorsed the appeal in September at Pisa, where the three Franciscans were staying under the protection of Emperor Louis IV the Bavarian, who had been excommunicated in 1324 and proclaimed by John XXII to have forfeited all rights to the empire. They followed him to Munich in 1330, and thereafter Ockham wrote fervently against the papacy in defense of both the strict Franciscan notion of poverty and the empire.

Instructed by his superior general in 1328 to study three papal bulls on poverty, Ockham found that they contained many errors that showed John XXII to be a heretic who had forfeited his mandate by reason of his heresy. His status of pseudo-pope was confirmed in Ockham's view in 1330–31 by his sermons proposing that the souls of the saved did not enjoy the vision of God immediately after death but only after they were rejoined with the body at the Last Judgment, an opinion that contradicted tradition and was ultimately rejected. Nevertheless, his principal dispute remained the question of poverty, which he believed was so important for religious perfection that it required the discipline of a theory: whoever chooses to live under the evangelical rule of St. Francis follows in the footsteps of Christ who is God and therefore king of the universe but who appeared as a poor man, renouncing the right of ownership, submitting to the temporal power, and desiring to reign on this earth only through the faith vested in him. This reign expresses itself in the form of a church that is organized but has no infallible authority—either on the part of a pope or a council—and is essentially a community of the faithful that has lasted over the centuries and is sure to last for more, even though temporarily reduced to a few, or even to one; everyone, regardless of status or sex, has to defend in the church the faith that is common to all. For Ockham the power of the pope is limited by the freedom of Christians that is established by the gospel and the natural law. It is therefore legitimate and in keeping with the gospel to side with the empire against the papacy or to defend, as Ockham did in 1339, the right of the king of England to tax church property. From 1330 to 1338, in the heat of this dispute, Ockham wrote 15 or 16 more or less political works; some of them were written in collaboration, but *Opus nonaginta dierum* ("Work of 90 Days"), the most voluminous, was written alone.

Excommunication

Excommunicated after his flight from Avignon, Ockham maintained the same basic position after the death of John XXII in 1334, during the reign of Benedict XII (1334–42), after the election of Clement VI, and even after the death of Louis the Bavarian in 1347. In these final years he found time to write two treatises on logic, which bear witness to the leading role that he consistently assigned to that discipline, and he discussed the submission procedures proposed to him by Pope Clement. Ockham died at a convent in Munich, probably in 1349, apparently of the Black Death.

BIBLIOGRAPHY. A critical edition of Ockham's political works (*Guillelmi de Ockham Opera Politica*) is in preparation by the University of Manchester (already published are: vol. 1, 1940; vol. 2, 1963; vol. 3, 1956); and of the nonpolitical works in 25 vol. by the Franciscan Institute, St. Bonaventure, N.Y., vol. 1 (1956). LEON BAUDRY, *Guillaume d'Occam: sa vie, ses oeuvres, ses idées sociales et politiques,* vol. 1 (1949), is an authoritative study of Ockham's life and works, with emphasis on a discussion of the problems of authenticity and chronology. A fairly complete bibliography of editions, manuscripts, and studies to date of publication is included. JURGEN MIETHKE, *Ockhams Weg zur Sozialphilosophie* (1969), also an authoritative work, studies Ockham's career through his writings, showing the development of his political theory, including his doctrine of the church.

(P.D.V.)

Oda Nobunaga

Oda Nobunaga was a 16th-century Japanese general who, after a century of feudal wars had engulfed Japan in seemingly never-ending anarchy, made himself virtual

dictator of the country. Aided by Toyotomi Hideyoshi and by Tokugawa Ieyasu, the first of the Tokugawa shoguns, he succeeded in restoring a stable government and in unifying the country.

Nobunaga was born in 1534, in the province of Owari, the son of a government official who had amassed wealth and a respectable force of military retainers. In

By courtesy of the International Society for Educational Information, Tokyo, Inc.

Oda Nobunaga, portrait by Kano Munehide, 1583. In the collection of Choko-ji, Aichi Prefecture.

1549 Nobunaga succeeded to his father's estate and soon overpowered his relatives and the principal family of the province. By 1560 he had proved his brilliant strategic gifts by bringing all of Owari under his sway; and in the same year he astonished all of Japan by defeating the huge forces of Imagawa Yoshimoto, one of the overlords of his neighbourhood provinces, his first step toward unification of the country.

Stout-hearted, audacious, and autocratic, Nobunaga was quick to seize on any promising new invention. Ahead of other feudal barons, or daimyo, he organized units equipped with muskets. He also brought under his control the agricultural production of the fertile Owari plain, as well as the rising merchant class of the city of Nagoya in the centre of the plain. With an economic base thus assured, he planned to advance on the Kinki district, the prosperous area surrounding Kyōto, long the centre of Japanese power.

Rise to power

In 1562 he entered into an alliance with Tokugawa Ieyasu, a capable feudal lord of the neighbouring province of Mikawa; and in 1567 Nobunaga, feeling that he had secured his rear flank, moved his base of operations north to the city of Gifu. In the following year, he supported Ashikaga Yoshiaki, who escaped from the rebel band, hoping to become shogun after the assassination of his elder brother, the former shogun Ashikaga Yoshiteru. Nobunaga marched on Kyōto, the capital, and made Yoshiaki shogun. Soon, however, he fell out with Yoshiaki, and at last in 1573 he deposed him, thus ending the Ashikaga shogunate, even though it nominally lasted until Ashikaga Yoshiaki's death. In 1576, in order to consolidate his hold on the area, Nobunaga built for his headquarters a magnificent castle at Azuchi on the shore of Lake Biwa in the neighbourhood of the capital.

Meanwhile, he promoted a new economic policy by abolishing the checkpoints for collecting tolls on the roads as well as from the guilds, both of which had been privileged sources of income for the local daimyo (lords of manors). He also strengthened his military forces; and in 1571 he destroyed the Enryaku-ji monasteries, which had maintained their traditional power in politics and religion ever since the beginning of the Heian period in the 8th century. In the meantime, the fanatically religious Buddhist Ikkō sect held out against Nobunaga's attempts to unify the country by retaining the loyalty of minor local lords, extending its secular power, aiding Yoshiaki, and allying its members with the powerful daimyo of many provinces. In fact, Nobunaga had fought the Buddhist Ikkō sect directly and indirectly for more than ten years. It was only through the mediation of the Imperial court that Nobunaga in 1580 finally achieved the surrender of the fortress-monastery of Hongan-ji at Ōsaka, the most important political and military centre of the Ikkō. After capturing a great number of manors and temple estates, Nobunaga established his hold on the samurai and the wealthier farmers by investing them with the newly won estates. He thus gained a firm political and economic basis, which he strengthened by reducing even further the traditional influence of the Buddhist temples.

Elimination of Buddhist influence

Once established in Kyōto, he extended his protection to the Jesuit missionaries and assisted them in building a church in the capital and a seminary in Azuchi. He did so not only because of his interest in European culture but because he regarded the encouragement of Christianity as a further means of restraining the influence of the Buddhist temples. Nobunaga was a non-believer; his attitude toward Christianity was frankly political.

In 1582 he conquered central Japan and then made an attempt to conquer western Japan. In June of that year, however, he was killed during the rebellion of a discontented vassal. By the time of his death, Nobunaga had succeeded in bringing nearly half of the provinces of Japan under his control. He had overthrown the *ancien régime* and old order of fractionized power held by the daimyo and had paved the way for the political and economic unification of the country.

BIBLIOGRAPHY. TADACHIKA KUWATA, *Oda Nobunaga* (1964); and RYOICHI SUZUKI, *Oda Nobunaga* (1967), are the standard biographies (both in Japanese). GEORGE B. SANSOM, *The Western World and Japan* (1950) and *A History of Japan, 1334–1615*, 3 vol. (1961); and MICHAEL COOPER (ed.), *The Southern Barbarians: The First Europeans in Japan* (1971), provide useful information on Nobunaga's period.

(A.Eb.)

Oder River

The Oder River, a vital economic artery in eastern Europe, runs through the western portions of Poland and has considerable contemporary regional significance. It is one of the largest rivers in the catchment basin of the Baltic Sea, second only to the Vistula in size. For the first 83 miles from its source, it passes through Czechoslovakia. For a distance of 100 miles in its middle reach, it constitutes the boundary between the Polish People's Republic and the German Democratic Republic, before reaching the Baltic Sea via a lagoon north of the Polish city of Szczecin. Called the Odra in Polish and Czech and the Oder in German, the river is an important waterway, kept in good operating condition. It forms a link, by way of the Gliwice Canal, between the great industrialized areas of Śląsk (Silesia), in southwestern Poland, and the trade routes of the Baltic Sea and beyond. The Oder is connected with the Vistula, Poland's largest river, by means of a water route utilizing the Warta and Noteć rivers, together with the Bydgoszcz Canal, and is tied in with the waterway system of western Europe by way of the Oder-Spree and Oder-Havel canals in the German Democratic Republic.

The total length of the Oder River is 550 miles (886 kilometres), 464 miles of which lie in Poland. The total watershed area has been calculated at 46,000 square miles (119,150 square kilometres), of which close on 90 percent is in Polish territory. The mean elevation of the Oder Basin above sea level is 535 feet. From the river's source and over the greater part of its course, the Oder flows in a generally southeast to northwest direction; only from the junction with the Neisse (Nysa łużycka) River does the northward trend toward the Baltic commence. The principal left-bank tributaries are the Opava of Czechoslovakia, and the Osobłoga, Nysa Kłodzka, Oława, Ślęza, Bystrzyca, Kaczawa, Bóbr, and Neisse; while from the east the Oder receives the following tributaries: Olše, Kłodnica, Mała Panew, Strobrawa, Widawa, Barycz, Obrzyca, Warta, Mysła, and Ina. From the junction with Opava, the Oder is navigable for a distance of 472 miles for 220 to 230 days of the year. Towns of

particular importance along the Oder are Ostrava in Czechoslovakia; Frankfurt in East Germany; and Racibórz, Opole, Brzeg, Wrocław, Nowa Sól, and Szczecin in Poland.

The course. The Oder starts its course in Czechoslovakia, at an altitude of 2,080 feet in the Oder Mountains of the Hrubý Jeseník Range. Initially it runs as a mountain stream with a steep gradient that progressively lessens until the river reaches the floor of the structural depression called the Moravian Gate; from here on, the Oder continues its course in a wide valley. After receiving the Olše River, a tributary, the Oder enters Poland, and makes its way as a lowland river that in a characteristic manner alternates between following ancient east–west stream valleys of glacial origin and crossing gaps cut in the intervening uplands. Where the Oder takes advantage of these pre-existing valleys, it reaches widths as large as six miles or more, while in gaps it narrows down to about a mile. Near Koźle, the Gliwice Canal enters the Oder; and from here as far as Brzeg Dolny, the river has a navigable channel 116 miles long, controlled by 23 locks. From Brzeg Dolny downstream until the final outflow into the Szczecin Lagoon, the river channel is fully improved. Beginning with the mouth of the Neisse River, the Oder becomes the borderline between Poland and East Germany. In this part of the valley, there branch off the Oder-Spree and the Oder-Haval canals. Farther downstream, the Oder Valley shows numerous cross branches and parallel channels. Fifty-two miles from its outflow into the Baltic, the Oder splits into two main branches; the left canalized branch, called the Western Oder, passes through the city of Szczecin and enters the Szczecin Lagoon directly, while the right branch, the Eastern Oder (in its final section known as Regalica) passes east of Szczecin via the large Lake Dąbie and then also enters the Szczecin Lagoon.

Origin

Hydrology. The Oder has a limited flow volume. During low-water periods, in summer and autumn, the river is fed from storage reservoirs built in the upper tributaries. The mean water depth in the Oder channel is three feet, and the mean rate of flow is three feet per second. In summer the upper reaches of the Oder system are flooded by heavy precipitation, while in spring the middle and lower reaches suffer from meltwater floods. Flow volume varies from as low as 530 cubic feet per second in the upper reaches to up to 50,000 cubic feet per second in the lower. Annually the Oder delivers into the Baltic an average volume of no less than 3.8 cubic miles of water. The ice cover on the river lasts up to 40 days per year. As is the case with many of the world's great rivers flowing through heavily industrialized regions, the Oder's waters are, unfortunately, heavily polluted; of the fish that are still found in the river, the most common are bream and eel.

Improvements

The first hydraulic works—embankments and other methods of flood prevention—were started in the Oder Valley as early as the 12th century; spillway dams built in the 13th century were in operation until the 18th century, when work was initiated on channel straightening by means of excavated cuts. Improvement of the straightened part of the Oder channel was for the most part completed around 1900 (although final improvements were not made until after World War II), while control works in the middle and lower reaches were carried out in the interwar period.

Historical background. Due to its geographic situation, the Oder was, in ancient times, of major importance as the zone where people inhabiting southern and northern Europe came into contact with each other and exchanged cultural values. The first agricultural population arrived from the south after passing the Moravian Gate, which separates the Sudety from the Carpathian mountains. Along the middle reach of the Oder there developed the pre-Lusatian and the Lusatian cultures (of the Bronze Age), which greatly affected the later evolution of the Slav population. In the area surrounding the Oder estuary, there was a mutual interpenetration by Scandinavian, Germanic, and Slav cultures. Finally, in the 9th and 10th centuries, the Polish state developed between the Oder and the Vistula. In the 13th century, the German expansion dislodged Poland eastward, away from the Oder Basin. But on the basis of the 1945 Potsdam Conference between the Soviet Union, the United States, and Great Britain, the Polish nation returned to its former lands bordering the Oder River.

Regional significance. The Oder River is an important element in the Polish economy, serving as a supplement to the heavily overburdened railway and highway systems linking the highly industrialized regions of the south with the largest Polish seaport, Szczecin, at the Oder's Baltic mouth. The river carries about 10 percent of the total tonnage handled by the port. The Oder is also used by the barges of the German Democratic Republic, which travel over a system of navigable canals that connect the Oder with the Central European waterway network.

A system of navigable canals connects the Oder with the Vistula, Poland's largest river, and also with the rivers of the eastern portion of the country and the waterway system of the Soviet Union. This creates the possibility that the entire system may evolve into an all-water commercial route, transporting commodities from west to east and from east to west. Before full advantage of this situation can be taken, however, further investments, notably the construction of river port and embankment facilities, appear indispensable. (W.Pa./Je.P.)

Odonata

Large, active by day, and often strikingly coloured, dragonflies (order Odonata) are among the few insects that have secured a major place in folklore and art. Though they cannot inflict injury to man, longstanding vernacular names such as "horse stinger" and "devil's darning needle" testify to their formidable appearance. In Japan, where a journal (*Tombo*) is devoted solely to reports of their biology, dragonflies traditionally have been held in high regard.

Dragonfly is the common name for large, winged, sun-seeking insects comprising the order Odonata. Adult dragonflies have two pairs of narrow transparent wings, a long slender abdomen, and prominent eyes; they can be recognized by the skewed thorax, forwardly directed legs, and small inconspicuous antennae. Typically they fly near ponds and rivers in which their larvae develop. Dragonflies are predatory insects; the adults catch small flying insects on the wing. Although they consume mosquitoes and other insects troublesome to man, they are catholic feeders and not considered of appreciable value in controlling specific pests.

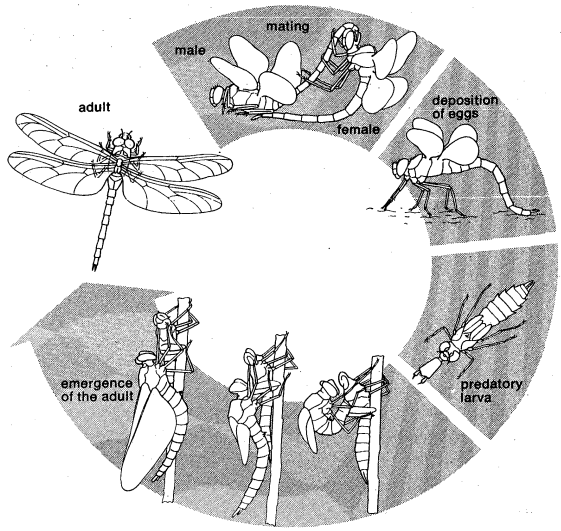
Distribution and size range

The order Odonata is small, well known, and widely distributed. The world fauna probably does not greatly exceed the 5,000 species now known. Dragonflies occur from the tropics, where they are most numerous and varied, to the latitudinal limit of trees. At a medium-sized pond, populations seldom exceed a few hundred adults of each species. Though the structure of adults is uniform, there is a wide range of sizes, and the wingspans of the smallest and largest species are about 1.8 and 19.3 centimetres, respectively.

Natural history. Since it is fully aquatic, the larva (or nymph) differs markedly in structure and behaviour from the flight-oriented adult. During development the larva molts approximately 8 to 15 times and passes through a corresponding number of stages or instars; the number of stages varies within and between species. There is no pupal stage.

A few minutes after hatching from an egg, the sheathlike cuticle of the first instar larva (prolarva) splits and releases the spider-like second instar. The early instar larvae feed actively on various small water animals including minute Crustacea and Protozoa; during later instars, the larva feeds on midge larvae, aquatic beetles and snails, and even small fishes. Wing pads appear at an early stage and grow larger with subsequent molts. Toward the end of the last instar, the organs become adult inside the larval skin. A few days later the larva climbs out of the water and molts to disclose the adult, a process known as emer-

gence. The largest dragonflies usually leave the water after sunset and take to flight just before sunrise; therefore, their emergence is seldom witnessed. Smaller species may emerge during the daytime.



Life cycle of the dragonfly.

The newly emerged adult dragonfly is soft, reproductively immature, and lacks full coloration. One of its first actions is to fly away from water. Adult life consists of two stages, a prereproductive (or maturation) period that lasts two days to two weeks (depending on the species and weather) and a reproductive period that lasts one to six weeks.

Mating behaviour

During the first stage the adult feeds actively, away from water. The second stage begins when, as a sexually mature adult, the dragonfly flies to the mating rendezvous, usually a pond or river where the eggs will be laid. Males assemble there slightly earlier than females and space themselves along the shore or over the water. Each male, once established, defends an area of characteristic extent, much as birds defend territories. When a male adult approaches or enters a "territory" occupied by another individual of the same species, the occupant acts aggressively and a contest often ensues; thus, territories are held by the most vigorous males. If a female adult approaches or enters a territory the resident male tries to mate with her. In some species mating is preceded by a courtship display during which, by her responses, the female accepts or rejects the male. Usually, however, there is no evident prelude, and copulation occurs immediately.

The mating posture is unique among insects. With the claspers at the end of his abdomen, the male grasps the top of the female's head or prothorax; then, by his movements he induces her to bring the tip of her abdomen forward so that it meets the intromittent organ at the base of his abdomen; in this way the female receives a sperm capsule. This "copulation wheel" can be formed in flight, though mated adults usually alight promptly. Copulation lasts between a few seconds and several hours, depending on the species. After mating, the female may lay eggs immediately or after a delay of hours or days. If she lays immediately, she may do so alone, with her partner still attached in tandem, or with him hovering nearby darting at other males that approach. In this way females lay eggs with little interference.

Eggs are laid in one of two ways. Species with a well-formed ovipositor place them within or on plant tissue, above or in the water. Some climb beneath the water surface to lay and remain submerged for an hour or more. Species without an ovipositor wash the eggs off the end of the abdomen or drop them onto the water surface. Eggs of some species are laid in running water and usually possess adhesive or tangling devices that prevent their being swept downstream.

In order to reproduce, the adult dragonfly requires warm weather, food, and water in which to lay its eggs. In equatorial waters reproduction occurs throughout the year; elsewhere, however, the life cycle is modified so that the reproductive stage coincides with the appropriate season. Where rainfall is a seasonal constraint, as in the tropics, the adult survives dry periods in one of two ways, either by retiring to humid woodlands (*Lestes, Gynacantha*) or by becoming nomadic and following the moving rainbelt, laying eggs in temporary pools that appear in its path (*Pantala, Tramea*). In both cases the larval stage is brief (one to three months), and the adult stage relatively long. Outside the tropics, where the seasonal constraint is imposed by low temperatures rather than drought, winter is passed in a resistant stage. In a few species the resistant stage is the immature adult and sexual activity is postponed until spring (*Aciagrion, Indolestes, Sympecma*); in other species it is the egg (*Lestes, Aeshna*); but most often the resistant stage is the larva, which, depending on the species and location, may require one to five years to complete development. Thus the larval stage lasts considerably longer than the adult one, which may last only a few weeks. In temperate regions, each species flies at a characteristic time of year, as determined by larval responses to climatic variables such as temperature and day length.

Requirements for reproduction

Dragonfly larvae are eaten by fishes, birds, and each other; predators at the time of emergence include birds and small crocodiles. Once they have left the emergence site, the adult insects have few regular enemies. In flight they are able to evade almost all predators by their agility. Larvae and adults are parasitized by flukes (Trematoda) and water mites (Hydracarina), and eggs by minute wasps.

Form and function. *Adult structure.* The principal structural features of adult dragonflies reflect adaptations to flight. The large wings are strengthened with a network of veins, and each has a thickened patch (the pterostigma) on the anterior edge near the tip. In certain migratory Anisoptera (*Libellula, Pantala, Tramea*), the expanded hindwing permits gliding flight. The sloping thorax accommodates large wing muscles and sets the legs forward where they can grasp prey. Where the head joins the thorax, a delicate orientation organ maintains equilibrium during flight. The large compound eyes, acutely responsive to movement and form, play an important role when adults capture food or interact with other dragonflies. In most Anisoptera the compound eyes meet dorsally, and in certain Aeshnidae that fly only in subdued light (*e.g., Gynacantha*) the eyes occupy almost the entire surface of the head.

Adaptations to flight

The unique method of copulation exhibited by Odonata can be regarded as a special way of transferring a sperm capsule. Instead of placing the capsule on the ground or a leaf for the female to pick up (as certain primitive insects do), the male puts it on his own body, where accessory genitalia at the base of the abdomen are especially constructed to receive it and transmit it to the female during flight. Interlocking structures on the male claspers and female head or prothorax make the copulation wheel secure.

Larval structure. Larval structure, though broadly similar to that of the adult, reflects demands of the aquatic environment, particularly with respect to respiration, locomotion, feeding, and concealment. Oxygen is replenished by diffusion through gills; in most Zygoptera these are three leaflike plates at the tip of the abdomen, while in most Anisoptera they are outgrowths of the hindgut wall, ventilated as water is pumped in and out of the anus. Both respiratory systems are used in emergencies as methods for locomotion; the jet propulsion employed by Anisoptera is particularly effective. Unlike the adult, the larva exhibits wide variation in form. Species living within fine sediment are cylindrical, with a terminal siphon that maintains respiratory contact with the water above (*Aphylla*); those residing on the surface of mud or sand often are flattened, with lateral spines (*Ictinogomphus*); those living among plants near the surface are streamlined and very active (*Anax, Lestes*).

Adaptations to aquatic environment

The larva captures prey in an unusual way. Typically it remains motionless until it detects a victim by sight or

touch. When the prey comes within range, the larva shoots out a prehensile organ (the labium) that grasps the prey and draws it back to the mandibles. In the resting position the labium lies beneath the head and thorax, sometimes obscuring the front of the head like a mask. It can be extended almost instantaneously by a localized increase in blood pressure that is controlled by a muscular diaphragm in the abdomen.

Evolution, paleontology, and classification. The extinct Protodonata possessed a complete series of alternate convex and concave wing veins; the two suborders, Palaeodonata and Archodonata, are known from the Upper Permian and Upper Carboniferous, respectively. The Odonata, in which the series of wing veins lacks the posterior media and anterior cubital, contain four extinct and three living suborders. The extinct ones are Meganisoptera (Upper Carboniferous and Upper Permian), which included the gigantic *Meganeura monyi* with a wing-span of 70 centimetres; Protanisoptera (Lower and Upper Permian); Protozygoptera (Lower Permian); and Archizygoptera (Triassic and Jurassic). Living suborders are listed below.

Distinguishing taxonomic features. Features used by taxonomists in classifying members of the order Odonata are, for adults, the shape and venation of the wings, distance between the compound eyes, form and development of caudal appendages, and presence of an ovipositor. Larvae are classified according to the type and form of respiratory organ, shape and armature of labium, number and arrangement of body spines, and shape of abdomen.

Annotated classification. The classification given here is that of F.C. Fraser (1957); it takes account of the fossil record of this group, which is unusually long and rich. In this scheme the superorder Odonatoidea (Odonatoptera of Lameere) comprises two orders: the Protodonata and Odonata.

ORDER ODONATA

Carnivorous insects with aquatic larvae (except *Megalagrion*), terrestrial and winged adults; adults with two pairs of similar richly veined wings and a pterostigma, small movable prothorax, large obliquely set thorax with forwardly placed legs, intricate accessory genitalia on second and third sternites in the male; larvae with labium modified to seize prey, gills in hindgut or as caudal or lateral abdominal structures; about 5,000 known species, of which all but two are Zygoptera or Anisoptera.

Suborder Zygoptera (damselflies)

Upper Permian to present; both pairs of wings similar, petiolate at the base; eyes more than one diameter apart; male adult with four caudal appendages, two superior and two inferior, female with ovipositor well developed; larva with caudal or lateral gills, gizzard with 8–16 radially symmetrical fields; one extinct, 17 living families; small, delicate, weakly flying insects; long slender abdomen attains greatest development in *Mecistogaster; Megalagrion oahuense*, small forest dweller in Hawaii, only known member of the Odonata with a terrestrial larva; size: body length about 1.9–15.0 cm, wingspan about 1.9–19.3 cm.

Suborder Anisoptera

Jurassic to present; well represented in Tertiary; hindwings broader at base than forewings; wings not petiolate at base; eyes less than one diameter apart or touching; male adult with three caudal appendages, two superior and one inferior, female with ovipositor well developed (Aeshnidae, Petaluridae) or reduced; larva characterized by rectal gills, anus closed by three robust spines, gizzard with four to eight dental folds; eight living families; robust, strongly flying, agile insects; some accomplished migrants; size: body length about 1.8–11.9 cm, wingspan about 2.4–15.7 cm.

Suborder Anisozygoptera

Mainly Mesozoic; adults intermediate in structure and appearance between the other two living suborders; larvae resembling Anisoptera; 11 families, all extinct except one (Epiophlebiidae), known from two similar species inhabiting highland streams in Japan and Nepal; size: body length about 5.1 cm, wingspan about 6.5 cm.

Critical appraisal. The Odonata, as the only living descendants of the Odonatoidea, occupy a uniquely isolated position in the phylogeny of insects, having separated from all other winged groups at a very early stage of evolution. There is general agreement among specialists regarding the status and affinities of living families and genera and, with few exceptions, classifications based on the adult and larva correspond. The evolution of the two major suborders has, however, been a matter for continuing debate. Some authorities consider that the Zygoptera and Anisoptera arose independently from the Protodonata. According to Fraser (1957), the Anisoptera, in their descent from the Protodonata, passed through a zygopterous stage. This view of the genealogy of the order Odonata has gained wide acceptance.

BIBLIOGRAPHY

General: R.J. TILLYARD, *The Biology of Dragonflies—Odonata or Paraneuroptera* (1917), a classic, college-level textbook emphasizing systematics and functional morphology; P.S. CORBET, *A Biology of Dragonflies* (1962), a college-level textbook emphasizing ecology and behaviour.

Regional: C. LONGFIELD, *The Dragonflies of the British Isles* (1949); P.S. CORBET, C. LONGFIELD, and N.W. MOORE, *Dragonflies* (1960); F.C. FRASER, *Odonata*, vol. 1–3, *The Dragonflies of India*, in "The Fauna of British India Series" (1933–36); E.M. WALKER, *The Odonata of Canada and Alaska*, vol. 1–2 (1953–58); J.G. NEEDHAM and M.J. WESTFALL, *A Manual of the Dragonflies of North America (Anisoptera)* (1954); E.C.G. PINHEY, *The Dragonflies of Southern Africa* (1951); P.A. ROBERT, *Les Libellules (Odonates)* (1958), a German work on western European Odonata. Works listed in the above section include identification manuals incorporating biological observations.

Technical: F.C. FRASER, *A Reclassification of the Order Odonata* (1957), the definitive work on classification; S. ASHINA, *A Morphological Study of a Relic Dragonfly Epiophlebia Superstes Selys (Odonata, Anisozygoptera)* (1954), a comprehensive study of the only living genus of Anisozygoptera; P.S. CORBET, "The Life History of the Emperor Dragonfly, *Anax Imperator* Leach (Odonata: Aeshnidae)," *J. Anim. Ecol.*, 26:1–69 (1957), with emphasis on larval growth, metamorphosis, and emergence; M.E. JACOBS, "Studies on Territorialism and Sexual Selection in Dragonflies," *Ecology*, 36:566–586 (1955), a detailed treatment of sexual behaviour and egg-laying; C.H. KENNEDY, "The Relation of American Dragonfly-Eating Birds to Their Prey," *Ecol. Monogr.*, 20: 103–142 (1950), an exhaustive study of bird predation; F.X. WILLIAMS, "Biological Studies in Hawaiian Water-Loving Insects, Pt. II, Odonata or Dragonflies," *Proc. Hawaii. Ent. Soc.*, 9:273–349 (1936), a description of the ecology of the only known species with a terrestrial larva.

(P.S.C.)

Office Machines

For a variety of reasons the work load of business offices has increased rapidly during the 20th century, and, because much of the work is repetitive and routine, many kinds of office machines and appliances have been developed to handle it efficiently and accurately. The manufacture of such machines and appliances has become a major world industry in which new products are constantly under development.

Most office machines fall into either of two broad categories, writing and reproducing machines and computing and accounting machines. Among the former are typewriters, dictating machines, and the various types of duplicating and copying machines, In the second group are calculating machines, cash registers, accounting machines, and classifying and tabulating machines. In addition, there are several miscellaneous types of office machinery such as coin sorters, mail handlers, electronic paging devices, and others.

History. While the origins of modern office machinery may be said to date back to ancient times and the invention of the abacus, the major developments occurred in the late 19th and 20th centuries. This is attributable in part to the multiplication of office tasks with the growth of large-scale enterprise, and in part to certain basic advances in technology. The machine-tool industry, which developed in the United Kingdom in the mid-19th century, made possible the mass production of precision components, a prerequisite to such devices as the typewriter, which appeared in the U.S. in the 1870s. Important problems in a totally different field, sound recording, had to be solved before the typewriter's companion, the dictating machine, could come into being. Although Thomas Edison patented his original phonograph in 1877, it was a

dozen years before he produced a modification suitable for office usage.

Similarly, the original invention of photography in the mid-19th century required considerable refinement before an economically practical office photocopying process could be developed.

The most important advance in basic technology came with the birth of electronics in the early years of the 20th century. Most modern office equipment, of which the computer is the outstanding example, has been based first on the utilization of the electron tube and, more recently, on the transistor.

With the increasing sophistication of office machinery, and the consequent increasing capital investment represented by it, considerable study has been given to the rationale of using such equipment. This may be briefly summarized as depending on the economy of labour effected by the greater speed and volume of output, the value of possible improvements in accuracy, and the effects on office personnel, such as retraining problems.

WRITING AND REPRODUCING MACHINES

Typewriters. The typewriter, the most common office machine, came into use in the 1870s and has virtually eliminated handwritten business letters and reports. The many improvements made in the typewriter since its invention have culminated in electric and automatic electric typewriters. The automatic typewriter produces typewritten copies of form letters or other documents in any number desired. The copy for such applications is first typed in the form of coded perforations on a roll of paper tape or in the form of impulses on magnetic tape. The tape activates the type bars on the typewriter. Automatic typing may be stopped to enable an operator to insert particular amounts, names, dates, and other information. In the late 1960s the automatic typewriter was connected to a computer, which can be programmed to enter names and other individualized information on the forms being produced (for a detailed discussion of this equipment see TYPEWRITERS).

Automatic typewriters

Dictating and transcribing machines. Dictating and transcribing machines provide for the storage and later reproduction of spoken messages. Dictating machines normally employ a magnetic process, recording the voice on plastic disks or belts, wire, or coated tape, which can be removed from the machine after dictation and forwarded to the point of transcription. The transcribing machine reproduces the dictated message so that it can be typed. By the use of these machines, stenographic transcription can be carried out while additional dictation progresses.

The dictation and transcription units are often portable and can be housed in a single unit. Recent developments include centralized systems in which individual desk controls with microphone or telephone are connected to a central bank of dictating–transcribing machines. Fewer machines are thus necessary and there is less manual handling of the message media. A highly sophisticated arrangement requires no handling of discs, tapes, or related supplies. Instead, an automatic handset smaller than a telephone permits the person dictating to record the message, correct errors or modify his statements, and review one word or the entire dictation, merely by operating a switch on the handset. The message is recorded in a transcription unit that lights up for the secretary when a message is to be typed. This signal goes off when there are no more messages in line. The system also allows for dictation from regular and out-of-plant telephones at any time.

Duplicating machines. Around 1000 BC a cylinder mounted on a frame and rolled by hand over a slab of clay became an early duplicating process. Later, damp tissue applied against a gum-base ink served the same function. Many new and sophisticated kinds of duplicating machines have since been developed. Yet some modern machines still use variations of these early techniques. The major classifications of duplicating machines are stencil, hectograph, multilith (or offset), and imprinting, all of which will be discussed later. Regardless of variations in these processes, all duplicating machines require the preparation of a master from which copies are made by a machine. Duplicating machines in this sense are thus differentiated from copying machines (discussed later), in which copies are made from an existing original in an exposure–image-forming process. It should also be noted that on the basis of this distinction the typewriter is not a duplicating machine even when carbon paper inserts are used, because the copies are prepared simultaneously with the original rather than printed from a master.

Early duplicating process

Stencil. The stencil method of duplicating involves the use of a coated fibre sheet for a stencil. Using a typewriter with the ribbon shifted out of the way so that the keys do not strike it, the operator types the material to be duplicated on the stencil, cutting the coating on the stencil and exposing the fibre base so that ink can pass through to it. Corrections are made by a special sealing fluid that permits retyping over the patched-error position. Signatures or drawings may be placed on the stencil with a hand stylus.

Two machines are widely used to produce copies from stencils. In one the stencil is fastened to the ink-saturated surface of a hollow, rotating cylinder. As the cylinder rotates, the ink flows through the cuts in the stencil to the sheets of paper fed under the cylinder. The other machine has two cylinders around which a silk-screen belt rotates. The stencil is attached to the screen, and the ink is distributed to the surface of the rotating cylinders by two rollers. The ink is thus pressed through the small openings in the screen and through the stencil openings, or image areas. Paper is fed between the stencil, and an impression roller to create the copy. Up to 5,000 copies can be made in either process from a single stencil, and stencils can be stored for considerable periods of time for re-use.

Hectograph. There are two types of hectograph or direct-process duplicators: gelatin and spirit.

Gelatin and spirit processes

The gelatin process, now rarely used, requires the preparation of a special master paper upon which the copy to be duplicated is typed, written, or drawn with a special ink or ribbon. This sheet is then pressed face down against a moist gelatin surface, to which the image is transferred in reverse form. Sheets of paper pressed against this impregnated gelatin receive an image impression. Either a flatbed or rotary machine, similar to that used in the stencil method, can make the duplicate copies. The master copy can be prepared in a variety of colours by using ink and carbon sheets of different shades. Multicoloured copies may thus be produced in one operation. The practical limit on copies produced by the gelatin process is about 200.

The spirit method is also referred to as the direct, or fluid, process. The master copy is prepared by typewriter, handwriting, punched card, or computer-printing devices. Master copies can also be prepared by copying machines and microfilm reader-printers. The master sheet is then fastened to a rotating drum. As copy sheets, slightly moistened by a special liquid, are brought into direct contact with the master sheet, a minute amount of the carbon is transferred to them, resulting in finished copies. Multicolour duplication in one operation is possible, as it is with the gelatin process. A further advantage of the spirit process is that information can be added to or deleted from the master. Up to 300 copies can be made from one master sheet. Special hectographic masters can be produced directly from an original or from certain reproductions by using the thermocopier, which employs a facsimile-like process.

Offset. Offset duplicating processes, also known as multilith or lithograph, require either chemically fixing copy on a metal sheet or preparing a paperlike master copy by typing, printing, or drawing, or by an electrostatic or xerographic copying process. This master impression is inked and brought against an intermediate composition agent, usually rubber, to which the image is transferred. The image is immediately transferred (offset) again to a copy sheet. The process operates on the principle that grease will not mix with water. Since a special grease-base ink forms the image on a moisture-retaining

surface, the moisture remains in blank areas, repelling the ink to the image area. Thousands of copies can be produced, and masters can be stored for indefinite periods and re-used. In photo offset the material is prepared on a layout sheet and photographed on an offset plate or master. After the completed master is attached to the drum of a duplicator, copies are prepared as explained in the offset process.

Imprinting. There are three basic methods of imprinting: (1) spirit hectograph master cards, (2) stencil cards, and (3) metal or plastic plates. Hectograph master cards are made with the aid of hectograph carbon. The imprint on the card is transferred by means of a chemical solution to envelopes or other forms. Up to 250 imprints may be made from a single master card. Stencil cards consist of small pieces of stencil tissue mounted in a cardboard frame. The copy to be imprinted is typed or written on the stencil tissue. When a drawer of stencil cards is attached to the addressing machine, the machine automatically brings together the cards and the envelopes or other forms to be imprinted, performs the imprinting, places the stencil cards in a tray, and stacks the imprinted forms. Metal or plastic plates of various sizes may be embossed, with the text standing out in relief, on a special machine. Forms or envelopes are imprinted through an inked ribbon as the raised-image plates are fed through another machine, which operates automatically at high speed. The plates are practically indestructible.

The most common office machine for this kind of imprinting is the addressing machine, which addresses postcards, envelopes, shipping labels, or tags. It is also used widely for imprinting checks, production orders, schedules, requisitions, time tickets, routing slips, tax bills, and similar forms.

The mailing selector

Attachments for these machines are designed for special services. A selector, for example, permits certain plates to pass through the machine without imprinting, thus directing mailings only to selected groups. A cutoff device permits the printing of selected data only. A repeating device prints the data several times before the plate is ejected, thus permitting names and addresses to be entered on envelopes and invoices in one operation. A dating device that enters the date on forms along with the other data embossed on the plates is used frequently in entering names and addresses on letters and statements. It is also possible to use these machines for distribution and addition of figure information. Holes punched in plates at predetermined points activate a sensing mechanism that translates desired information and lists it on forms. The preparation of billing records and payrolls is an example of ways in which this device may be used.

Copying machines. There are several major types of copying processes, including photocopy, thermography, electrostatic copy, and facsimile transmission, all of which are now discussed. In none of these is a master copy needed as an intermediary step, as in the case of duplicating machines. Some copying machines are fully automatic, compact, and capable of making copies at very high speeds. Such machines, especially electrostatic copiers, are beginning to compete favourably with stencil and hectographic processes in speed and on a cost-per-copy basis, so that the distinctions between duplicating and copying are becoming less clear.

Photocopy. One of the oldest processes involves photographing a negative in the making of a positive image. Although other light-utilizing processes and some sophisticated variations have evolved, the most commonly used photographic processes are diffusion transfer, dye transfer, and diazo, which are contact processes, and microfilm, which utilizes a camera.

In the diffusion-transfer process the original copy is made by typing, drawing, or printing on a translucent paper or cloth sheet, which is placed, image side up, on light-sensitized paper. Both sheets are exposed to light, which passes through the translucent sheet and deactivates the nonimage areas. Since the printed areas absorb light and thus do not permit it to pass through to the sensitized sheet, these image areas remain activated. Another variation employs a two-step process in which a light-sensitized negative sheet is placed on top of the original, sensitized side down. The light source passes through the negative sheet to the original, which absorbs the light in the image areas. The nonimage areas reflect the light to the sensitized side of the negative paper, turning these areas black. The negative is developed in contact with positive transfer paper in a chemical solution. The two sheets are then squeezed and peeled apart.

In the dye-transfer process, the original is placed against the sensitized paper as in two-step diffusion transfer and then exposed to light. The image is formed in a solution in which areas not exposed to light are transformed into a dyelike substance that, when placed against nonsensitized paper, transfers to it and forms the copy. Dye-transfer machines can also produce masters for offset duplicating machines.

In the diazo process, an improved form of blueprinting, a translucent original and a sheet of copy paper treated with azo dye are used. With the original and copy paper placed together as in the one-step diffusion-transfer process, an ultraviolet light source is used. The copy paper is developed by passing it either between two rollers moistened with developer or, in the case of dry diazo, through ammonia vapours.

In contrast with the contact processes are the photographic processes in which cameras are used to take pictures of the material to be reproduced. Most processes produce a negative from which positive copies in various sizes may be made.

Small strips of film carrying information and attached to carrier pigeons were used for communications as far back as 1870 during the Franco-Prussian War. Today microfilm is widely used in office files. The capacity of the microfilm camera for information miniaturization is such that it can transform 100 file drawers of original documents into as few as two drawers of filmed records. Records-retention management can often be facilitated in this way, with the added advantage that the film can be noncombustible.

Origins of microfilm

Microfilm information can be stored in any of a number of forms. Rolls and cartridges of film necessitate sequential access to given records, while jackets or card-held microfilm frames are unit records with sequential storing within each unit. Variations of this latter form have led to microfiche, which consists of rows of frames placed directly on films, and aperture cards, which are regular punch cards with film inserts for computerized storage and retrieval of visual and punched information. The film may be studied through a viewer, which shows the image in its original size, or, if duplicate sets are needed, it may be used to make reproductions of the original copy, either on paper or on microfilm.

Innovations provide for high-speed filming and reduction–expansion of computer output data as well as fast copy generating units. Much greater reduction capacities are also available for specialized storage needs, as is the case with libraries; a whole book can easily be printed on a single file card.

Thermography. The commercial application of heat or infrared copying methods—sometimes called thermography—is relatively new. Special copy paper and the original copy are placed in contact with each other and run through a machine in which they are exposed to infrared, or heat, rays. The original copy absorbs the rays in areas darkened by print, line drawings, or other illustrations, making the same impressions on the heat-sensitive surface of the copy paper. The original with the new copy paper attached may be put through the machine repeatedly to produce the number of copies desired.

Electrostatic copy. Exceptionally flexible, electrostatic copying processes can copy anything that is written, printed, typed, or drawn. They will reproduce halftones (photographs) and will enlarge or reduce copy.

Xerography: "dry writing"

In the transfer electrostatic process, or xerography (from the Greek "dry writing"), the original document is scanned by a mirror that reflects the image onto a drum coated with amorphous selenium, a nonmetallic element that is able to hold an electrostatic charge in the dark while allowing it to dissipate when exposed to light. From

here the process operates on the principle of opposite charges attracting and like charges repelling one another. The selenium on the drum is given a positive electrostatic charge that is held in the dark areas of the reflected image but dissipates in the light, nonimage areas. The coating is then treated with negatively charged powder, or toner, which adheres to the positive image area. The darker the image, the more toner adheres. This reverse image is transferred to positively charged ordinary paper after it has passed over the drum, drawing the powder onto itself. The toner is fused by infrared heat to form a permanent copy.

A direct electrostatic process differs from xerography in that the image is shown directly on specially coated paper with no intermediate transfer needed.

Facsimile. In facsimile reproducing processes, a transmitter scans the original material, converting the image into electrical signals that are sent over telephone lines or by other means of communication, such as radio or microwave. A receiver reconverts the signals into the original image on paper, the clarity of the copy depending on the number of scanning lines per inch utilized in the process. The combination transmitter receiver is normally housed in one unit small enough for use on an office desk. Printed, written, graphic, and photographic material can thus be copied at great distances from the original documents. Some units transmit the image over telephone lines and then allow for normal voice transmission over the same lines.

COMPUTING AND ACCOUNTING MACHINES

Calculating machines. The earliest and most long-lasting calculating device ever devised is the abacus. Various versions of this instrument were extensively used in ancient and medieval Europe and Asia. The abacus probably grew out of a more primitive instrument, a board sprinkled with sand, on which tallies were marked. When lines were drawn in the sand dividing the board into columns, to represent tens, hundreds, and thousands, the calculating board, or abacus, was born, with pebbles or bones used as counters. The abacus continues to be widely used in much of Asia today. Because a skilled operator can perform a variety of calculations with great speed and accuracy, the simple and economical abacus retains its popularity even in industrially advanced Japan.

The rise of interest in natural philosophy in western Europe in the 17th century spurred the development of mechanical calculating instruments. John Napier, a Scottish mathematician, in 1614 discovered the logarithm, which made the task of multiplying very much easier. The mechanization of the logarithm led to the development of a whole class of calculating machines called analog, or measurement, devices.

Pascal's digital calculator

In 1642 a French scientist, Blaise Pascal, invented an adding machine that in a sense represents the first digital calculator. It is called digital since it performs its operations by the counting of digits, the numbers 1 through 9, and 0. In 1671 a German mathematician, Gottfried Wilhelm Leibniz, invented a machine that performed the multiplication process by repetitive adding. His stepped gear wheel, a gear wheel with two or more complete circular sets of teeth arranged adjacently on the same rim so that corresponding teeth form a series of steps, still appears in a few 20th-century devices.

During the last quarter of the 19th century, inventors developed machines that were more compact and made possible modern desk calculators, which can perform the processes of addition, subtraction, multiplication and division, and, by shifting numbers left or right, the operations of multiplying or dividing by powers of 10. They can store the result of an operation on counter wheels visible to the operator, and data can be introduced by means of keys on a board. In all these machines, however, it is necessary for the operator to intervene at each step of a calculation. The time required for the calculation thus depends on the speed of the user.

In carrying out a mathematical calculation, man performs three distinct functions that must be incorporated into any calculating machine: first, the basic arithmetic operation of adding, subtracting, multiplying, or dividing; second, the decision on what to do next (*e.g.*, take the number in column one, multiply it by the quantity just produced); and third, storage of the initial, intermediate, and printing of final results on sheets of paper or other media.

Many attempts were made to make this entire process automatic. Technical difficulties, however, stood in the way of the goal until an American scientist, Vannevar Bush, and associates at the Massachusetts Institute of Technology developed and placed in operation the first differential analyzer in 1930. It was a fully automatic calculating instrument that could be programmed to do different types of problems. In addition to the three basic steps of computation, the analyzer had an input–output device that permitted communication between the mechanism and the user. Since 1930 the developments in this field have been extraordinarily rapid. (For a detailed discussion see COMPUTERS.)

Basic types of calculators. Modern calculating machines fall into one of three basic types: adding–listing, key-driven, and rotary. The first and third type are either ten-key or full keyboard machines. Ten-key machines are designed for one-hand touch operation. Numbers are entered by depressing one key at a time, beginning with the left-most digit and proceeding to the right until all digits, including zeros, have been recorded. Full keyboard machines provide a vertical column of keys for each digit position from one to nine. The machines vary in capacity of digits, depending upon the number of columns of keys included.

Adding–listing machines

Adding–listing machines, either ten-key or full-keyboard, are capable of adding, subtracting, dividing, and multiplying. Because multiplying and dividing operations are slow, they are generally used for adding and subtracting. The individual amounts entered as well as the totals are listed on a paper tape.

In key-driven machines, which are full-keyboard, calculations are performed directly by depressing the keys. The results of the calculations appear on a dial. Some machines of this type have two dials, one indicating individual calculations and the other registering grand totals and net results. Although key-driven machines can perform all four arithmetic functions, they are especially suitable for problems that require rapid addition and multiplication.

Rotary calculators need the additional operation of a lever or motor bar to record a number in the register. Although they are capable of performing the four arithmetic functions, they are used primarily for division and multiplication problems. Both listing and nonlisting and either ten-key or full keyboard models are available. Calculators of this type vary in the number of answer registers available. Some have only one; others may have as many as five. A register holds a figure or amount until it can be used with amounts in other registers to perform additional calculations.

Some rotary calculators have split registers, enabling the operator to add individual items in one half while accumulating them in the other half of the register. A variety of special features are available, including automatic squaring, decimal placement, and high-speed, short-cut multiplication. In fully automatic machines of this type, complex problems can be solved by recording the factors and depressing the multiply or divide key. The machine performs the calculation, repositions the carriage, and then stops.

Modern electronic calculators. The electronic desk calculator, though fully developed only in the 1960s, has a long engineering design history. The first electronic calculators were slow and cumbersome. When solid-state devices were introduced in the 1950s, calculator size was reduced by the miniaturized transistor and diode.

Use of integrated circuits

When a special photo-reduction process made it possible to miniaturize these solid-state circuits the integrated-circuit generation of electronic calculators came of age. Solid-state circuits were reduced to an eighth of an inch square on a silicon wafer known as a metal oxide silicon chip. Each chip, the size of a half dollar, houses an entire

circuit capable of performing the same functions of their solid-state ancestors. The entire chip, produced under dust-free conditions to prevent imperfections in manufacture, is embedded in special plastic with gold leads connecting the outside to the integrated circuit (see also INTEGRATED CIRCUITRY).

The latest generation of circuitry involves the further integration of a number of metal oxide silicon chips in forming a large-scale integrated circuit. Such circuits, which require fewer connections in the machine, become less costly as their size increases. As manufacturing techniques improve, costs may be expected to decline and increasing numbers of electronic calculators will doubtless appear for nonbusiness and home use.

Developments in electronic calculating machines include such sophisticated capabilities as display screens to show the results of operations. The electronic display characters are created in various ways, including cathode-ray–tube projection, neon lumination, or luminated vacuum-tube processes. Usually from two to five display registers are available, affording visual access to operational results as well as to variables of the operation. Some units feature instead a listing or printing device housed within the regular package. These are electromechanical in design and are the source of noise and heat. While fast, they are much slower than the electronic operations of the major unit, which performs in milliseconds. The calculator unit itself has no internal moving parts and is therefore noiseless.

The modern electronic calculator is simple to operate. Besides the basic processes of addition, subtraction, multiplication, and division, most units handle square roots, squaring, percentages, and logarithms with the pressing of an appropriate key. The machine can also often handle several operations automatically. Decimal placement is easily accomplished, with as many as 15 or more significant places possible in many machines. In addition to the several display registers for results and variables, registers are also available for the storage of constants and intermediate results.

Exponential and trigonometric operations

Very sophisticated machines handle exponential and trigonometric operations through the use of special keys that utilize core memories much like those in modern computers. Accessory card programmers that accept punch cards, magnetic tape cassettes, and other input devices provide some units with computer-like capabilities, all in a desk-size machine. The keying in of the necessary data for the calculation is still necessary. These machines go through programmed routines processing numeric data on command. Graphic output has also been achieved with the use of a special accessory printer, and some machines signal the operator if an impossible task is included in the problem.

It should be noted that electronic calculators perform operations once thought to be possible only in computers. Meanwhile, computer sophistication has resulted in smaller units, some of which are beginning to resemble electronic desk calculators. The two may be distinguished by the abilities of the computer to handle alphabetic and symbolic information and to "branch"—*i.e.*, decide which sequences or subsequences of instructions to follow when certain conditions arise in the data. Perhaps even these differences will be less distinct with future developments in the two lines of machines.

Specialized calculating and accounting machines. *Cash registers.* Cash registers are recording and adding machines that help a merchant control his business and assist salespeople in serving customers. Cash registers make records of each transaction for the customer and the merchant. There are many different types, some extremely simple and others highly complex. Most cash registers have certain fundamental features and functions. The typical machine indicates the amount of a transaction at the top of the register so that it can be seen by both customer and salesman. It keeps separate totals of sales by various classifications. It prints and issues a receipt on cash sales or overprints a record of the transaction on a sales docket on charge sales. It keeps within the register an audit strip, which is a complete printed record of every transaction that has been made. It has special counters to show the number of customers handled and the number of each kind of transaction. In a recent development, a clerk records on a preprinted card the number of each of several items sold. The card is inserted into an electronic calculator, which calculates and enters on the card the cost of each item and the total. The price of each item is stored in advance in the calculator. These prices may be changed by merely running a revised price card through the calculator. Information on the items sold and cash received is simultaneously transmitted to a central computer, which provides inventory and cash-on-hand data to management on a minute-by-minute basis. This application is especially useful in branch office and franchising operations.

Accounting machines. Accounting machines are complex devices that combine features of the computing machines already described with some provision for writing. They have taken over much of the labour of bookkeeping. Billing machines, for example, are designed to type names, addresses, and descriptions, to multiply, to figure discounts, and to add net total. A simple accounting machine generally has only one or two registers and no typewriter facility. Abbreviated description keys are used to denote the type of transaction, and the carriage is designed to hold a number of forms at once, a capability that enables the operator to post associated documents, such as ledger card and daybook, simultaneously. More elaborate accounting machines have many registers, often have a typewriter built in, and can carry out a number of operations automatically. Accounting machines post to one account at a time and frequently accumulate in the manner of cash registers to permit summarizing and distributing accounts. The window-posting accounting machine, in common use in banks, hotels, and retail stores, prints the transaction on the customer's statement or passbook as well as on an audit sheet and automatically calculates new balances.

Billing machines

Classifying and tabulating machines. These machines use cards on which coded data are recorded. Their function is to facilitate the sorting and tabulation of diverse data. The simplest card systems function by notching the perforated perimeter of the card according to codes. These systems are limited to the performance of sorting tasks. They are operated by a simple machine or device, frequently merely a metal rod, which is inserted in a specific perimeter location of a stack of cards. It will thus penetrate both notched and unnotched holes. When the rod is lifted, those cards that have notched holes will drop out.

The first modern device of this type was designed for the automatic tallying or tabulating of statistics. Following extensive development, however, modern tabulating machines can now automatically figure and print a report based on the information given to it in the form of perforations in cards or tape. Between 1834 and 1854, Charles Babbage in England designed and partly built an "analytic engine" in which both the amount to be operated upon and the nature of the operation were entered by perforated cards. His machine was never completed, largely because modern precision techniques of fabricating metal parts to close tolerances were not yet developed. The modern tabulating machine was finally produced by a U.S. statistician, Herman Hollerith.

Census tabulation

Hollerith tallied census returns by an electrical reading of the data punched into cards (which provided coded information on each individual in the census) and accumulated the totals in separate registers. The system was successfully used in the U.S. census of 1890 and the British census of 1911. During the next decade notable improvements were made, including provisions for automatically feeding cards under brushes for reading and a machine for sorting the cards rapidly into desired groupings. The U.S. census in 1900 furnished particularly good evidence of the possibilities of the system. The population volume was ready one year and seven months from the start of enumeration. It was estimated that hand tabulation of only three factors—sex, nativity, and occupation—would have required the services of 100 clerks for 7

years and 11 months. Since 1901 the commercial utility of a punch-card system has been greatly increased. The tabulating or unit record method is now to be found in general use throughout the world.

Whenever a transaction involves a number of amounts that are later to be used as units two or more times in compiling various totals, it is generally economical to translate it onto a perforated card. If, for instance, the data connected with the cost of production of a certain part are recorded on a card, this card can be used in computing totals of individual wages, departmental costs, machine costs, and productive hours.

The card upon which the data are recorded in the form of perforations has a column for each number or symbol to be entered. The digits 0 to 9 are denoted by holes punched proportionate distances from the bottom edge of the card, starting with the 9 position. Above the 0 position is space for two additional perforations used principally for actuating certain control operations of the machine. These additional positions may also be used for special classification numbers or in combination with perforations in the regular digit position to denote alphabetic characters. The card is divided vertically into fields, each of which denotes a particular fact of the total information recorded. Some of the information is descriptive and some of it is quantitative. The former controls sorting and indicating; the latter represents amounts to be totalled. Sorting and totalling are accomplished by different machines. Three separate mechanical devices are necessary in the system: a perforating machine or punch, a sorter, and a tabulator. The ordinary form of perforating machine has a set of 12 punches under which the card is advanced a column at a time. By means of a drum card or skip bar and stops, certain fields can be skipped where no punching is to be done. More complex types of perforating machines are arranged to permit the automatic duplication of information already on one card, or other cards, to punch any predetermined number of cards with identical data, and to number cards consecutively as they are punched. The speed with which trained operators can perforate the cards depends on the nature of the information being recorded as well as its extent.

Sorting devices

The sorter arranges cards in numeric or alphabetic sequence or sorts them into related groups for tabulation. A sort brush electrically reads the punches in a card, one column at a time, at the rate of from 400 to 2,000 cards per column per minute. One well-known type of sorting machine has 13 pockets, one for each possible columnar perforation and one for rejects. As the perforations are sensed by the sorting brush, the cards are dropped into appropriate pockets, depending upon the information punched into the column being processed. Newer sorters are capable of sensing perforations in cards by photoelectric cells instead of by brushes and are thus faster. Some are equipped with special devices which are capable of providing certain types of computational and statistical information.

The tabulator, or accounting machine, reads data from perforated cards by means of reading brushes and processes the data in accordance with instructions it receives from a control panel or wired unit. The control panel, which is similar to a telephone switchboard, determines which card columns are to be read, what is to be done with the data in these columns, and where to print the processed information. The tabulator can either calculate amounts punched in cards—about 150 per minute—and print only the total, or list each amount individually, followed by a total for the group. Generally, tabulators contain several counters, making it possible to take totals of several card fields simultaneously and store the amounts for grand totals. Subtraction is also possible, and devices can be added to the printing mechanism for handling report and bill forms. As a result, tabulating machines have become automatic accounting and billing machines.

MISCELLANEOUS OFFICE MACHINES

In offices where quantities of coins must be handled, machines are used to sort, count, wrap, or dispense coins. Sorting is accomplished simply by sifting the various sizes. Trip counters count the coins, and wrapping machines wrap a standard number in paper. The coin-dispensing machine releases a specific sum when an appropriate key on a keyboard similar to that on an adding machine is depressed.

Mail-handling machines

Mail-handling machines are designed to expedite handling of large quantities of incoming and outgoing mail. Since outgoing mail requires the most work, machines are designed to aid in this task. The main types are folding machines, inserting machines, and envelope sealers. Stamp-affixing machines mechanically separate stamps from a roll, moisten them, and apply them to envelopes. Postage meters print stamp charges directly on envelopes or on adhesive strips to be affixed to envelopes. The machines are set by postal employees to print postage equal to the amount of deposit paid. An automatic postage meter and mail-scale machine feeds, stamps, seals, and stacks letters at high speed. It also automatically weighs and computes the postage on parcel-post packages.

The modern office is frequently of such size that efficient communication becomes a factor in maintaining the flow of work. Two-way voice systems operate on the telephone principle to permit communication between specific areas. In a commonly used version, the individual may call any one or any combination of a limited number of stations by depressing appropriate keys on the receiver–transmitter box. When he speaks into the box, his voice, amplified electrically, is broadcast from similar boxes at the various stations. Others may then reply in similar fashion from their stations.

Paging systems utilizing either the telephone switchboard operator or private internal telephone systems are very popular. Flashing lights, bells, and similar signals locate persons by a code or speakers call out their names. Those paged then call the operator or designated persons to receive the message. This process frees individuals from a fixed position in the office or plant because the signalling devices can be spread throughout the building and grounds.

Teletype systems provide typed communications between two or more points. Messages are received and recorded even though no person is in attendance at the receiving machine. These systems are especially valuable for rapid intercommunication among a number of widely separated offices and frequently span an entire country or link offices in different countries. Telegraph services are also available to subscribers on a basis much like that of telephone service.

Recording units mounted on regular telephones also receive spoken messages after first announcing to the caller why his call is being answered by machine. One of these has become popular because of its ability to use regular telephone lines with special transmitter and receiver handling computer data from cards by converting them into appropriate tones. Machines can thereby "talk together" for the price of a regular phone call.

Dataphone

Closed-circuit television installations permit viewers to see documents and hear explanations transmitted from distant locations. Rapid transmission of recorded information is thus possible for verification, identification, or computation. Signatures or account balances may be transmitted for verification from a central office to many points instantaneously, eliminating the need for large numbers of duplicate records. Monitoring of operations is also a key application. Closed-circuit television holds much promise in its capacity to handle visual and animated information.

Paper cutters, punches, binding equipment, and folding equipment are available in a wide range of sizes for office use. Document-destroying equipment disposes of outdated paper documents by shredding or chopping. Collating equipment assembles pages of printed matter in a specific order.

Time-recording machines combine a clock mechanism with a recording device to provide a quick, correct recording of the time of occurrence of certain routine events. Time-stamping machines are manually actuated to stamp the time upon letters and documents. Time-re-

cording clocks record the times of arrival and departure of employees. The employee inserts a time card into the machine when he arrives and again when he departs. Job-recording clocks function in identical fashion to record time of beginning and finishing particular tasks.

Check-writing machines that print amounts on checks so that they cannot be altered; check-signing machines; and numbering, dating, and receipting machines are examples of other office devices.

BIBLIOGRAPHY. P.L. AGNEW *et al.*, *Secretarial Office Practice*, 7th ed. (1966), a classroom text on basic secretarial tasks in the modern office, including a discussion of copying, duplicating, and calculating machines; BUSINESS EQUIPMENT MANUFACTURERS ASSOCIATION, *Office Machines: Systems Devices* (1967), an excellent discussion of the uses of modern office equipment in systems and of logical work-flow designs; M.J. HANNA *et al.*, *Secretarial Procedures and Administration*, 5th ed. (1968), a modern classroom text treating secretarial office duties and discussing copying and duplicating equipment and basic and special telephone services; G.R. TERRY, *Office Management and Control*, 6th ed. (1970), an up-to-date college level text on the many facets of modern scientific office management, with a useful section on office machines and electronic developments in that area; S.J. WANOUS *et al.*, *Fundamentals of Data Processing* (1971), a comprehensive treatment for classroom or general use, including the history of data processing development and modern hardware and software.

(S.J.W.)

Ogata Kōrin

One of the most typically Japanese artists of the Tokugawa or Edo era (1603–1867), Ogata Kōrin was a versatile painter and designer whose work and life were a perfect expression of the luxury-loving Genroku period (1688–1703), during which the well-to-do merchant class of Kyōto became the patrons of an elegant art that reflected their wealth and social position. Much admired during his lifetime, Kōrin's fame and influence have continued unabated since his death, and today he is regarded as one of the two great masters of the school of decorative painting known as the Sōtatsu-Kōrin school. In addition to the splendid, colourful screen paintings for which he is primarily famous, he was an expert lacquer artist whose work has had a tremendous influence on craftsmen working in this field. He is also famous for his textile designs and the pictorial decorations that he supplied for the ceramics of his brother, Ogata Kenzan, who is regarded by many critics as Japan's greatest potter.

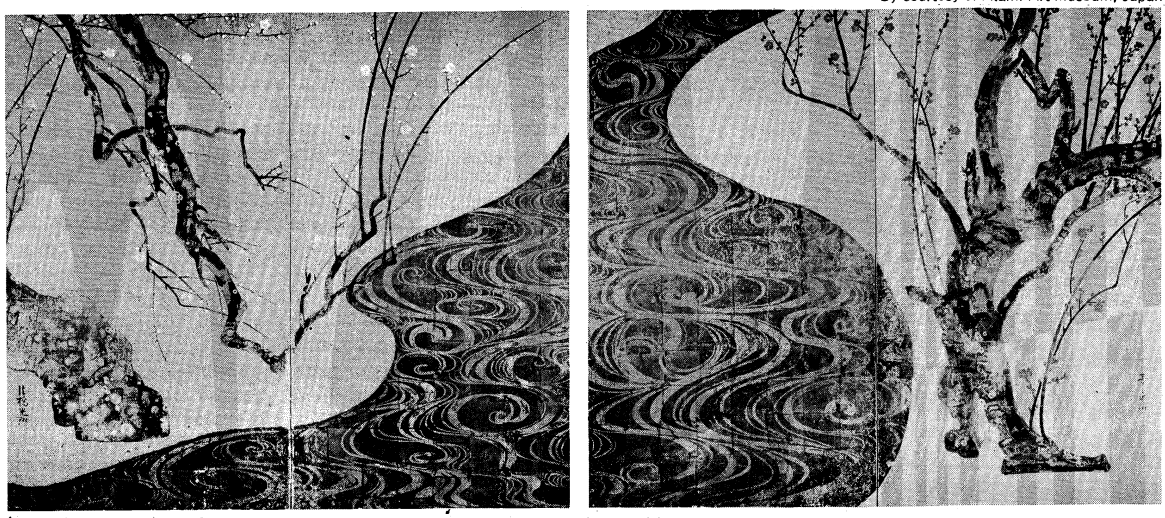
Family and youth

Coming from an old and distinguished Kyōto family, Kōrin was descended from a samurai (warrior aristocrat) who had served under the Ashikaga rulers and was related to a famous calligrapher and swordsmith, Kōetsu. Kōrin's grandfather and father were owners of Karigane-ya, a prosperous store dealing in fabrics for kimonos that was patronized by some of the most powerful ladies of the capital. Members of his family were also keenly interested in the arts. Kōrin's grandfather, Sōhaku, spent the later years of his life in Takagamine, an art colony established at the outskirts of Kyōto by Kōetsu; and Kōrin's father, Sōken, was an accomplished calligrapher of the Kōetsu school, as well as a lover of Nō drama.

Born in 1658 in Kyōto, the ancient capital of Japan, Kōrin grew up in an environment of luxury and aesthetic refinement. His original name was Ichinojō, but when he was 34 or 35 years old, he changed it to Kōrin. Having received a considerable inheritance from his father, the artist spent his youth leading a carefree life filled with the pursuit of pleasures and amorous adventures. He was not married until 1697, when he was almost 40. The extravagance of Kōrin's life-style is best illustrated by a story of a lavish picnic party that Kōrin and his friends gave at Arashiyama outside of Kyōto. As each of the participants showed off his splendid dishes, Kōrin provided the climax by producing his food wrapped in bamboo leaves decorated with gold. When the meal was over, the artist tossed the leaves into the river, an action for which he was banished from Kyōto because it violated the law forbidding the use of gold and silver among the common people. Because of such extravagances, Kōrin lost the fortune he had inherited and had to turn to art for a living. Earlier in his life, he had studied painting for many years, at first probably under the tutelage of his father, who was an accomplished painter, and later under Yamamoto Sōken, a member of the officially recognized Kanō school. Sōken, who was skillful in both Chinese-style ink painting and the traditional Tosa school painting, which employed Japanese subject matter and a colourful decorative style, instructed his pupil in both these manners. Other influences on his early work were paintings of Kanō Yasunobu and especially the work of Sōtatsu, who were the two most outstanding decorative painters of the early 17th century. Very few of Kōrin's early paintings have been identified, and those works that can be attributed to this phase of his career appear to have been painted in ink in the traditional Kanō style.

Kōrin's artistic maturity began in 1697, when he established himself as a professional painter. In 1701, at the age of 44, he was given the rank of Hokkyō, indicating that he was an accomplished artist, and after that virtually all his work bears the signature Hokkyō Kōrin. Because almost none of his paintings bear dates, it is difficult to determine the chronology of his artistic output, but it appears that all his important work was produced in the 20-year period after 1697. These years may be divided into three parts: the formative Kyōto period, from 1697 to 1703, when he gained recognition as an art-

By courtesy of Atami Art Museum, Japan



"Red and White Plum Trees" by Ogata Kōrin, pair of *byobu* screens painted with colour over gilded paper. In the Atami Art Museum, Japan.

Artistic maturity

ist; the period from 1704 to 1710, when he lived in Edo (present-day Tokyo); and the years 1711 to 1716, when he reached his artistic climax. His first masterpiece was a screen representing autumn grasses and flowers (now in a private Japanese collection), which combined his two most outstanding characteristics: a fine sense of abstract decorative design and a close study of nature. The culmination of this phase was in the famous iris screen, a work that is believed to date from 1701. After moving to Edo in 1704, he enjoyed the patronage of rich merchants and some of the great lords of feudal Japan. Among the works attributed to this period are a hand scroll depicting the flowers of the four seasons that has been handed down in the Tsugaru family of Tokyo; a charming hanging scroll of red and white azaleas (Hatakayama Museum, Tokyo); and a twofold screen representing waves. After his return to Kyōto in 1711, the work of Sōtatsu became the overwhelming influence on his art. In fact he based the compositions of two of his most famous screens on the painting of this master. These are the two sixfold screens depicting the waves at Matsushima and the twofold screen "God of Thunder and God of Wind" (Tokyo National Museum). The work that is usually regarded as the supreme achievement of this period is the screen "Red and White Plum Trees," now in the Atami Museum (see illustration). In this work, Kōrin's sense of flat, decorative design and his feeling for nature combined with an emphasis on an abstract colour pattern are seen at their very best. Although he died at the age of 59 in 1716, he left many pupils and followers. The most outstanding of these was Sakai Hōitsu, who was active during the early years of the 19th century.

MAJOR WORKS

Kōrin was the originator of the style named after him; below are a few works attributed to Kōrin (active late 17th–early 18th century): "Irises" (Nezu Art Museum, Tokyo); "Red and White Plum Trees" (Atami Art Museum, Shizuoka, Japan); "Narihira at Yatsuhashi" (Tokyo National Museum); "Narihira Going Eastward" (S. Hisamatsu Collection, Matsuyama, Japan); "Waves at Matsushima" (Museum of Fine Arts, Boston); "Waves" (Metropolitan Museum of Art, New York).

BIBLIOGRAPHY. I. TANAKA (ed.), *The Art of Kōrin* (1965), the standard Japanese work on the artist, with English summary; S. TAJIMA, *Masterpieces Selected from the Kôrin School*, 5 vol. (Eng. trans. from the Japanese, 1903–06), still the most complete study of Kōrin and his school, although outdated and not always reliable; D. RANDALL, *Kōrin* (1960), a brief popular account of the artist and his work in English.

(H.Mu.)

O'Higgins, Bernardo

Bernardo O'Higgins, the revolutionary leader and first Chilean head of state, was a key figure, as both a political and a military leader, in the winning of Chile's independence from Spain. After Chilean independence was established, he led the nation for six years, maintaining its security through the first critical years, and built the administrative foundation of the republic.

By courtesy of the Organization of American States, Washington, D.C.

O'Higgins, oil painting by Jose Gil. In a private collection.

Bernardo O'Higgins was probably born on August 20, 1778, in Chillán, a town in southern Chile, then a colony of Spain. As noted in his certificate of Baptism, he was the illegitimate son of Ambrosio O'Higgins, a Spanish officer of Irish origin, who became governor of Chile and later viceroy of Peru; his mother was Isabel Riquelme, a prominent lady of Chillán.

Education and early influences

Bernardo's father had only indirect contact with his son, who used his maternal surname until his father's death. At 12, Bernardo was sent to Lima, Peru, for his secondary education. Four years later he went to Spain. At 17 he was sent to England for further education. In London he became imbued with a sense of nationalist pride in Chile, a pride largely fostered by his contact with several political activists, of whom Francisco Miranda, the Venezuelan champion of Latin-American independence, exerted the greatest influence on him. Along with several other future revolutionary leaders, he belonged to a secret Masonic lodge, established in London by Miranda, the members of which were dedicated to the independence of Latin America. In 1799 he left England for Spain, where he came into contact with Latin-American clerics who also favoured independence and doubtless further strengthened his views. His political position was remarkable in view of the fact that his father was viceroy of Peru.

Bernardo's father died in 1801, leaving him a large hacienda near Chillán; by 1803 he was working the estate. This interlude may have been the most satisfying period of his life. The hacienda began to prosper almost immediately and Bernardo was soon maintaining a house in Chillán. In 1806 he became a member of the local town council.

Before O'Higgins had time to settle into his agrarian way of life, however, the foundations of Chilean society were threatened. In 1808 Napoleon I invaded Spain, which, occupied with its own defense, left its colonies, including Chile, largely uncontrolled; the first steps toward national independence began to be taken throughout Spanish America. On September 18, 1810, a national junta, composed of local leaders who replaced the governor general, was established in Santiago, and by 1811 Chile had its own congress. O'Higgins was a member, and during the next two years he played a key role in the country's turbulent political affairs.

By early 1813 Chile had a constitution and a junta that seemed able to control the country and to avert the threat of civil war. In 1814, however, the Viceroy of Peru sponsored an invasion to re-establish royal authority. Within a few months, O'Higgins rose from the rank of colonel of militia to general in chief of the defensive forces. Soon, he was also appointed governor of the province of Concepción, in which the fighting took place. But the war went badly and O'Higgins was superseded in command. In October 1814, at Rancagua, the Chilean patriots lost decisively to the Peruvian-led forces. For the next three years the country was occupied by the royalists.

Several thousand Chileans, including O'Higgins, crossed the Andes into Argentina in flight from the royalists. O'Higgins spent the next three years preparing for the reconquest of Chile. In January 1817, O'Higgins returned to Chile with the Argentine general José de San Martín and a combined army of the Andes consisting of Argentine troops and Chilean exiles. At Chacabuco, on February 12, 1817, they decisively defeated the Spanish, and with Chile largely reconquered, O'Higgins was elected interim supreme director.

Supreme director

For the next six years, as supreme director, O'Higgins maintained, on balance, a successful administration. He created a working governmental organization and provided the essentials of the new nation—peace and order. Under adverse circumstances he succeeded in building a national navy and in mounting a major military expedition against Peru to fight the royalists.

O'Higgins, however, was not politically astute: by 1820 he had antagonized the conservative church and the aristocracy with his reforms. Later he alienated the business community. He did not perceive the importance of a solid

political base, and, because his support was based on his prestige as a war leader in a threatened country, his fall was assured once the danger of war had disappeared. O'Higgins was associated with a grand scheme of continental independence that was essentially Argentine in its conception; by the time of his resignation—under pressure—in January 1823, a growing Chilean nationalism had rendered him and his Argentine colleagues much less attractive than they had been in 1817.

In 1809, at the age of 31, O'Higgins had observed: "The career to which I seem inclined by instinct and character, is that of laborer"; in rural life, he would have come to be "a good *campesino* and a useful citizen." As supreme director, O'Higgins had the positive attributes of solid moral principles, an eagerness to work hard, and singular honesty. In the countryside, as he himself understood, these virtues would have been ample, but in public administration they were not enough.

Last years

From 1823 until his death in October 1842, O'Higgins lived in exile in Peru, dividing his time between his hacienda and Lima. His last years were poignantly similar to his first: in his youth circumstances required that he live away from home; now in maturity circumstances again conspired to keep him abroad. In both periods he longed to return home.

Little is known of O'Higgins' personal life. Though he never married, he managed to acquire a family, in the same manner as his father had. His natural son Pedro Demetrio O'Higgins was his companion in exile.

O'Higgins was a liberal in the 19th-century sense of the term and an admirer of the British constitutional system. Although not as conservative as some contemporary Chilean leaders, he was not a democrat either. While his reputation since his death has fluctuated with the political predilections of governments and historians, his leading role in establishing Chile as a republic remains unquestioned.

BIBLIOGRAPHY. SIMON COLLIER, *Ideas and Politics of Chilean Independence, 1808–1833* (1967), the best study of the topic in English; JAIME EYZAGUIRRE, *O'Higgins*, 6th ed. rev. (1965), the best recent biography of O'Higgins (in Spanish); JAY KINSBRUNER, *Bernardo O'Higgins* (1968), a recent biographical study of O'Higgins and his period in English.

(J.Ki.)

Ohio

The first state to be carved from the Northwest Territory when it became the 17th member of the Union in 1803, Ohio, in the 20th century, reflects the urbanized, industrialized, and ethnically mixed United States that developed from an earlier agrarian period. The pattern of its life is so representative of the nation as a whole that it is often used to test attitudes, ideas, and programs in education, politics, and industry. Significantly, Ohio has supplied by birth or residence more than one-fifth of the presidents of the United States—William H. Harrison, Ulysses S. Grant, Rutherford B. Hayes, James A. Garfield, Benjamin Harrison, William McKinley, William H. Taft, and Warren G. Harding.

Overview of the state

The state's accessibility has been perhaps the key factor in its growth. Its location between the populous states of the Eastern seaboard and the growing Middle West and its lack of natural barriers to movement made it a corridor for east–west travel, whether by riverboat down the Ohio River, packet across Lake Erie, or wagon, railroad, or automobile and truck across its land. With Pennsylvania on the east, West Virginia and Kentucky on the southeast and south, Indiana on the west, Michigan on the northwest, and Lake Erie on the north, Ohio lies in the heart of the United States' most industrialized area, close to major resources of raw material and manpower and to the markets of the East, Middle West, and South.

Ohio's area of 41,222 square miles (106,764 square kilometres), excluding 3,457 in Lake Erie, ranks only 35th in size among the states; it is the smallest, after Hawaii and Indiana, west of the Appalachian Mountains. The more than 10,500,000 Ohioans counted in the 1970 census, however, made it sixth in population, with a density of about 260 persons per square mile, compared with the national average of just over 50. The state capital, after being located in Chillicothe and Zanesville during the early years of statehood, was finally settled in newly founded and centrally located Columbus in 1816. (For information on related topics, see the articles UNITED STATES; UNITED STATES, HISTORY OF THE; NORTH AMERICA; and GREAT LAKES.)

THE HISTORY OF OHIO

Prehistory and settlement. Remains of archaic civilizations dating from 5,000 to 7,000 years ago have been found in Ohio. Among later cultures, the Hopewell peoples, who disappeared about 1500, represented the highest prehistoric development; their burial mounds contained exquisitely crafted artifacts. The earliest white explorers in the 17th and 18th centuries found the region occupied largely by Miami, Shawnee, Wyandot, and Delaware Indians, as well as a host of smaller tribes.

Indian cultures

The region between the Appalachians and the Mississippi River was in nearly constant contention between France and Britain until 1763. Ordinances of 1785 and 1787 by the new United States government set the pattern for settlement and for government, though the Virginia Military District in southern Ohio and the Western Reserve (*i.e.*, of the state of Connecticut in 1786) in northern Ohio were set aside for veterans of the Revolution.

Statehood. In the decades after Ohio attained statehood in 1803, the foundations for future social and economic diversification were laid. Threats of Indian warfare were halted, and in 1813 the naval battle on Lake Erie played a major role in the War of 1812 with Britain. In spite of several economic panics, a program of internal improvements created a network of land, water, and rail transportation. By 1850 Ohio was the third most populous state in the nation, with nearly 2,000,000 inhabitants, and the leader in diversified agriculture, but it was still an essentially rural, agricultural state.

Economic growth. The basis of Ohio's industrial structure was built between 1850 and 1880, when the value of its manufacturing grew to more than twice that of agriculture. A great stimulus was provided by the Civil War (1861–65), in which Ohio supported the North, though strong antiwar sentiment existed. After the war, the growth continued, notably in the northeast and around Lake Erie. This growth led to considerable economic and social dislocation. After 1900 considerable attention was given to municipal reforms in Cleveland, Toledo, and other cities and to statewide programs of social legislation that attempted to alleviate problems caused by industrialization. In 1920 two Ohioans, Warren G. Harding and James M. Cox, faced one another for the presidency, and, in following decades, Ohio continued to play a pivotal role in national political life. Ohio reflected the racial strife that was widespread in the United States in the summer of 1966, when disorders in the predominantly black Hough district of Cleveland cost four lives. In 1968 Carl B. Stokes became the first Negro ever elected mayor of a U.S. city as large as Cleveland, though in 1967 and 1968, respectively, Springfield and Dayton also came under the administration of Negro mayors. The state became the focus of national attention in May 1970, when four students were killed by national guardsmen, who had been called out as a result of demonstrations at Kent State University, near Akron. Tragically, Ohio had again proved itself the mirror of the troubled nation as a whole.

THE NATURAL LANDSCAPE

The topography, river systems and groundwater, and soils in most of Ohio are the product of past glacial activity. These factors have strongly influenced the patterns of human settlement and land use.

Physiographic provinces. Three larger physiographic provinces extend into Ohio. The Allegheny Plateau reaches westward from Pennsylvania and West Virginia into the counties along Ohio's eastern border, from near Lake Erie to the Ohio River. The northeast is only partially glaciated, while the unglaciated terrain in the south-

east has made it the one region in the state that has showed declining population and economic power for decades. Throughout the plateau, the land is dissected by rivers winding among steep hills, and many elevations reach 1,400 feet (425 metres).

The Lake Plains stretch along Lake Erie to the northwestern counties and the Michigan boundary, then irregularly to the south. These level to slightly rolling lands were once underwater, and the swampiness of the northwest, around Toledo, presented obstacles to settlement before drainage made it more arable. The Central Plains, which extend westward beyond the Mississippi, take in western and southwestern Ohio and provide a deep soil. Curiously, the state's highest and lowest points are found here: the former, at 1,550 feet, is near Bellefontaine; the latter, at 425 feet, lies at the confluence of the Miami and Ohio rivers, near Cincinnati.

Waters. The principal water sources are the rain-fed streams, lakes, and reservoirs. Floods, which were once prevalent, have been controlled by numerous state and federal dams and other conservation measures throughout the state. Groundwater is used widely for public supplies, though the major industrial and population centres have limited resources. Huge stores are buried in preglacial valleys in central and south central Ohio; cities, which require millions of gallons daily for industrial use alone, are contemplating the use of these sources.

Lake Erie and the river systems

Lake Erie has an average depth of only 62 feet, which makes it the shallowest of the Great Lakes. It is also the most tempestuous, with frontal storms often roaring across it from Canada, and the most liable to shoreline erosion, harbour silting, and filling of its bed. Its shallowness, coupled with the concentration of population and industrial plants in its watersheds, led to a pollution that drew wide attention beginning in the 1960s. Programs in various areas were begun to deal with the environmental problems of the lake, which continues to be the principal source of water for many lakeside cities.

A low watershed separates the 30 percent of Ohio drained by the Maumee, Cuyahoga, and other rivers into Lake Erie from the 70 percent drained by the Miami, Scioto, Muskingum, and others into the Ohio–Mississippi system. The Ohio, no part of which is under state jurisdiction, is canalized and channelled for its entire length, as is the Muskingum from Zanesville to Marietta. Of the 110 lakes in Ohio, 83 have been built for industrial, supply, recreational, and other purposes.

Soils. Most of Ohio's soils are well suited to agriculture. The Great Plains soils are mainly rich glacial limestones, but the Lake Plains are the most productive. The sandstone soils of central and northeastern Ohio are best adapted to pasturelands, while the thin and heavily eroded hilly areas of the southeast provide little basis for productive farming. State and federal conservation and reforestation programs are under way in that area.

Climate. Temperatures in Ohio are similar to those across the north central and eastern United States, with summer highs and winter lows only infrequently reaching 100° F (38° C) and −20° F (−29° C), respectively. It is open both to cold, dry fronts from Canada and to warm, moist fronts from the Gulf of Mexico. The frequent meeting of such fronts causes much of the precipitation, which totals about 38 inches (965 millimetres) a year, including an average annual snowfall of about 28 inches, which is about one-seventh of the total precipitation.

Vegetation and animal life. The great hardwood forests that covered 95 percent of Ohio prior to white settlement have been reduced to less than 20 percent. The forests in the glaciated areas have less forest but better stands of timber, which include oak, ash, maple, walnut, basswood, hickory, and beech. Both wild and domestic flowers abound, though the clover, wild rye, and bluegrass of early Ohio are gone.

About 350 bird species are found in Ohio, at least 180 of them native. Among the 170 fish species are bass, trout, and perch, while the 60 or so species of wild animals include deer, opossum, fox, skunk, groundhog, and rabbit.

THE PEOPLE OF OHIO

The urban areas of Ohio first exceeded the rural in population by 1910 and by 1970 included more than 75 percent of all Ohioans. Five counties surrounding the major cities contained 43 percent of the people. Nonetheless, Ohio's large cities are following the national pattern of losing population to surrounding suburban areas. The growth of Columbus proper is largely attributable to municipal annexation of the suburbs.

The early settlers. The people who laid the foundations of Ohio, most of them of English ancestry, came from the older seaboard states. The first permanent American settlement in Ohio and the Northwest Territory was at Marietta in 1788 by a company of New Englanders who had fought in the Revolution. In the same year a group from New Jersey established a settlement near Cincinnati, and in the next few years other villages sprang up. In the south, particularly in the Virginia Military District between the Scioto and Miami rivers, many of the settlers came from Virginia and Kentucky. In 1796 the Western Reserve in northern Ohio was first settled, mainly by New Englanders from Connecticut.

The northern European stock

Pioneers from other European nations appeared in the years before 1830. Germans and Swiss came from Pennsylvania into the east central area. Many settlers of Ulster Protestant background also came in from the Middle Atlantic and Southern states. After 1830, settlers came directly from Germany and Ireland. Many Irishmen came from the Erie Canal to work on the Ohio canals and stayed on, and, when the railroads were built, Irish and German workers remained as permanent settlers. Germans who drained the Black Swamp country of the northwest stayed on to develop the farmlands around such centres as New Bremen, New Bavaria, and Minster.

After 1830 immigrants from southern Ireland settled in such cities as Cleveland, Defiance, and Cincinnati, where by 1850 they were second to the Germans among the foreign-born. In the 1830s Columbus also became an important centre of German population, as did Lancaster, where immigrants from Württemberg joined the Pennsylvania Germans who had founded the city.

In northeast Ohio, such towns as Massillon, Alliance, and Canton were established at an early date by Pennsylvania Germans, as was Steubenville, farther to the east. German settlers were also attracted to the rolling surface and fertile soil of Wayne County, which became one of the top agricultural counties in the nation. German-speaking Moravian missionaries came to Christianize the Indians in the early 1760s under the leadership of John Heckewelder, David Zeisberger, and Frederick Post. In 1817 Joseph Bimeler founded an experimental communistic settlement in Zoar that lasted until 1898. The Swiss settled around Dover and Sugar Creek in Tuscarawas County, as well as in Monroe County. This general area is sometimes referred to as the "Little Switzerland of Ohio." In Holmes County, Amish immigrants from Germany and Switzerland established settlements that remain. Today there are more Mennonites in Ohio than in Switzerland.

The Welsh arrived in the early 19th century to develop the mineral resources in Jackson County, and, for a long period, Welsh was the only language spoken. The Eisteddfodd, a festival of Welsh bards, and other elements of Welsh culture and music flourished. The language persisted to the third generation in many communities, with old Welsh songs passed on from father to son.

Later waves of immigration. In 1850 the principal racial stock was Scotch–Irish, although the Germans and the English also were important elements. In 1870 nearly 14 percent of Ohio's and 40 percent of Cleveland's population were foreign-born. The New England character of northern Ohio's beginnings was changing. Each new group established its own newspapers, clubs, social life, and churches, as the state became increasingly industrialized and urbanized between 1870 and 1900.

Increasing numbers of immigrants from eastern and southern Europe came into Ohio after 1880. By 1920 great numbers of Italians, Poles, Hungarians, Russians, and others had come to Cleveland, Toledo, Youngstown,

Akron, Dayton, Middleton, and other industrial communities. No other Ohio city, however, acquired such a polyglot population as Cleveland. Cleveland's foreign-born population was supplemented between 1880 and 1890 by new arrivals from Austria, The Netherlands, Russia, Hungary, Portugal, Greece, China, Japan, Turkey, and Mexico. Eventually, the city's culture was enriched by over 48 groups with different languages and backgrounds. There were many new Roman Catholic and Eastern Orthodox churches, as well as synagogues. The Greeks brought their coffeehouses, and the Slovenes and Poles brought their social halls. By the 1970s, through the action of the "melting pot" character of Ohio, the descendants of these many ethnic groups had firmly established themselves in the social, economic, and political life of the state.

The diversity of Cleveland's population

The changing ethnic character of the state is shown as well in the growth of the black population. In 1843, when the Wyandot Indians left Ohio for new lands west of the Mississippi, the Negro population of Ohio numbered about 25,000. By 1870 it had risen to more than 62,000, most of which was in southern Ohio, where Wilberforce University, one of the first permanent Negro educational institutions, was established. By 1970 the black population had risen to about 970,000, most of it in the cities.

THE STATE'S ECONOMY

A good physical location, a rich store of natural resources, productive soils, and ample transportation facilities made Ohio one of the first great industrial states in the nation. Over one-half of the nation's population is within 500 miles of its borders; and coal, iron, water, salt, limestone, ferroalloys, chemicals, clays, and plastics are close at hand. In all, two-thirds of the raw materials processed in Ohio's factories come from its mineral, agricultural, and forest resources. Although well over one-third of its labour force is employed in manufacturing, its continuing activity in agriculture and mineral production provides economic balance and diversity.

Economic resources. *Agriculture.* In 1850 Ohio ranked first among the states in agricultural production, and in the 1970s it continued to rank near the top. Although its farming acreage and the number of farms and farmers had decreased, more than 60 percent of Ohio was still farmland. Ohio farms have become larger, supplying more food with fewer man-hours of work through the application of science and technology. In 1900 the farm was an almost self-sustaining unit, providing its own food, fertilizers, and fuel; today, however, it requires machinery, motor power, electricity, and tractors, as well as a substantial investment of money and skill.

One of the leading states in the production of corn, oats, and hay, Ohio maintains large marketing inventories of fruit, feed, and vegetables, as well as livestock and poultry.

Mineral production. Ohio's mineral resources include stones and clays for manufacturing and construction and such mineral fuels as coal, petroleum, and natural gas for heat, power, and transportation. In recent years the value of Ohio's mineral production has reached new highs. Coal production accounted for the highest return, followed by industrial minerals, oil, and gas.

Coal was discovered in Ohio as early as 1808. It was adapted for use with iron and limestone in the pioneer iron-making enterprises that sprang up in the eastern and southwestern parts of the state. Later, the discovery of deposits of iron ore in the upper Midwest gave rise to great iron and steel centres of northern Ohio. Today, good-quality coal is found in 32 eastern and southeastern counties. Remaining supplies are estimated at approximately 42,000,000,000 tons. More than half the coal is produced by strip-mining procedures, which have left a scarred landscape, polluted water supplies, and a generally damaged environment. Although Ohio has laws regulating strip mining and requiring restoration, the results were not satisfactory. Citizen groups led the battle for stronger safeguards.

Limestone is used in many construction and manufacturing processes. Ohio is first among the states in sandstone production, accounting for about two-thirds of the nation's building sandstone, and it is third in sand and gravel production. The abundance and quality of surface clays, plastic fireclays, shales, and some gypsum and peat have made Ohio a leader in the manufacture of ceramic products. Almost one-half of its extensive salt production comes from natural brine, the remainder from large rock-salt mines. Important industrial consumers of salt in Ohio are the soda-ash and chlorine industries. Geologists estimate that the state's salt deposits could supply the nation's need for centuries to come.

Petroleum and natural-gas production

Ohio has been a producer of oil and natural gas since 1860, but, by 1900, billions of barrels of crude oil had been taken, and production in the state declined. In the early 1960s, however, new oil and gas deposits were discovered, and the industry experienced a revival. In 1970 there were more than 21,000 producing gas and oil wells.

Other natural resources of importance to Ohio's economy are its forest products and water supplies. Climate and soil are conducive to the production of fine hardwood. Ohio leads the nation in the industrial use of water, with countless industries depending upon the groundwater supply.

Manufacturing. Ohio's manufacturing is its most important economic activity and represents the largest single segment of Ohio's employment. Measured by employment, the most important manufacturing enterprises are the nonelectrical-machinery industry and the automobile and aircraft industries. In spite of increased activity in other parts of the nation, the northeast-to-Middle West manufacturing belt that stretches through Ohio still constitutes the chief industrial strength of the country. The state leads all others in such diversified manufactures as rubber products and porcelain ware, electrical machinery and apparatus, pumps and plumbing equipment, steam shovels, and coffins, and it is listed among the leaders in almost all industrial classifications.

Economic regions. Governmental planners have identified eight regions throughout Ohio that possess distinctive economic characteristics in terms of their human and physical resources, landforms, transportation, communications, and other attributes.

The northern strip. The Maumee Valley region, comprising ten counties in the northwest, is primarily agricultural, though it lies in the path of industrial expansion from east and west. Corn and wheat, as well as hogs and dairy and poultry products, are important. Its largest city, Lima, has a growing manufacturing base. The Lake Plains region, comprising ten counties on the southwestern shores of Lake Erie, also have flat, fertile plains with highly productive soils. Toledo, the major city, is an important centre in the Great Lakes industrial belt and the leading coal-handling port in the United States. It supplies glass and transportation equipment to nearby Detroit, Michigan, and processes the farm products of the region.

Economic diversity

The Lakeshore and Uplands region comprises 16 counties in the north and northeast. This area contains Ohio's largest industrial concentration and, on 18 percent of the state's land, holds 42 percent of its people. Cleveland is the industrial, financial, and cultural centre. Akron is the rubber capital of the world, a trucking centre, and a large cereal producer. Youngstown, in the United States "Ruhr Valley," is a major metal producer and fabricator, and Canton specializes in roller bearings, bank vaults, and vacuum cleaners.

Central and southern regions. The Miami Valley region, with some 2,300,000 inhabitants in southwestern Ohio, centres on Cincinnati and Dayton. Cincinnati leads the world in the manufacture of machine tools, soaps and detergents, and playing cards. Dayton is the world's largest producer of cash registers and magazines and the nation's largest producer of putty and plastics; it is the home, as well, of Wright-Patterson Air Force Base, a major flight-testing and research centre. The Sandusky Valley, seven counties with fewer than 250,000 people, is basically agricultural, though the small cities of Marion, Galion, and Bucyrus have manufacturing.

The Scioto Valley region of rolling plains in central

Ohio has a diversified economic base. Columbus, its central city, is the home of state government and numerous educational institutions, including the Ohio State University. Two-thirds of the working force is in government, education, finance, and other service occupations. The sparse populations of the Tuscarawas Valley region of eastern Ohio, and of the Ohio valley region in the southern and southeastern portion of the state, are predominantly rural. Terrain limits the agricultural productivity of both regions. In the southwestern Ohio Valley, mining and lumbering are major activities. Stone, clay, glass, chemicals, and metal fabrication are major industries.

Transportation. Ohio's chief transportation system in the first years of statehood, as in the territorial period, was its water routes. Lake Erie and the Ohio River provided east–west passage for the Indian trader, pioneer, and settler, and many rivers provided access into the interior. Shortly after statehood, the development of land transportation facilities was begun, when, in 1811, the federal government began to build a major portion of the National Road running from Cumberland, Maryland, to Vandalia, Illinois. In the same year the first steamboats appeared on the Ohio River, and in the 1820s the era of canal building began and lasted for some 30 years. The first railroad was constructed in 1832, and in the 1850s the first great east–west rail lines were constructed across Ohio as the shortest and easiest route to the Western cities, prairies, and beyond.

Ohio's transportation facilities of the 1970s play a major role in moving the nation's people and goods from east to west and from north to south, by highway, railroad, river and lake, and air. The shipping to and from its lake ports is worldwide, and the Ohio River carries twice the tonnage of the Panama Canal. The railroad mileage is the nation's sixth largest, though Cleveland and other major cities lost interstate passenger service when the semipublic Amtrak system went into operation in 1971. The pioneer experiments of Dayton's Wright brothers, Orville and Wilbur, led to man's first successful powered flight, at Kitty Hawk, North Carolina, in 1903, and Ohio today is both a testing centre and a focus of commercial aviation. Among the many research and development facilities contributing to Ohio's economic life is a huge Transportation Research Center administered by the Ohio State University.

ADMINISTRATION AND SOCIAL CONDITIONS

Governmental structure. Ohio's government, which is more extensive than that of many small nations, is a multibillion-dollar enterprise. The state and local governments are similar to those in other states.

State level. The executive branch, which is headed by the governor and other elected officials, includes the heads of 22 agencies serving as an executive cabinet. The General Assembly comprises a Senate and a House of Representatives. It has broad powers in policy formulation and monetary appropriation. The judiciary, which interprets and applies the law in cases before it, comprises a Supreme Court, ten courts of appeals, courts of common pleas and of probate in each of the 88 counties, and such other inferior courts as the legislature may establish. All judges are elected for six-year terms.

Executive, legislature, and judiciary

Local government. Each county exists as a quasi-municipal corporation, an arm of state government, but without general authority of self-government in the legislative field. Various optional forms are available, and the increasing social and economic problems of the multicounty metropolitan areas may lead to newer forms. Most larger cities operate under home-rule charters that permit them to choose the form most suitable to their needs. The mayor–council type is most common, though Cincinnati operates under a city-manager–council plan. The township, Ohio's oldest form of government, remains a viable form, though the number is diminishing as they are annexed into municipalities or as newly incorporated villages assume their functions.

Political life. State laws carefully prescribe the rules for forming and running political parties, conducting elections, and balloting. The two-party system has prevailed generally, but Ohio has produced such minor-party leaders as Norman Thomas, many times a presidential candidate on the Socialist Party ticket; Victoria Claflin Woodhull, the first woman to run for president, with the Equal Rights Party; and such leaders of fringe political groups as Jacob Coxey, who led the march of "Coxey's Army" from Massillon, Ohio, to Washington, D.C., in 1894 to demand various economic reforms.

Since its inception the Republican Party has been slightly more successful than the Democratic in statewide elections. In national politics, the parties are evenly matched, making Ohio a focus of national electioneering. In addition to eight presidents, Ohio has supplied three vice presidents, three chief justices and eight associate justices of the U.S. Supreme Court, and 35 federal Cabinet officers. One of these chief justices was William Howard Taft, who also served as president. Although political dynasties have been rare in Ohio political life, Taft's father had been secretary of war and attorney general under Grant and was later U.S. minister to Russia and Austria-Hungary. His son, Robert A. Taft, served in the United States Senate from 1939 to 1953, and his grandson, Robert A. Taft, Jr., was elected to the U.S. Senate in 1970.

Social conditions. The major expenditures of state government are for education, health, welfare, and highways. A number of state executive agencies administer the programs in these areas, often in conjunction with federal bodies.

Education. More than 50 percent of the Ohioan's state-tax dollar is channelled into education. More than 3,000,000 students are accommodated at schools of all levels, which include 12 public and 65 private institutions of higher learning. Several municipal universities also receive state aid. A nine-member Board of Regents is responsible for the development of higher education in Ohio.

Higher education

Called a "land of schools and colleges" from its beginning, contemporary Ohio ranks about seventh in the United States in the number of accredited colleges it contains. Ohio University was established by Ohio's first legislature in 1804 as the first state institution of higher education west of the Alleghenies. In 1809 Miami University became the second state university. Both these institutions conduct important programs of work, have large enrollments, and maintain numerous branches. Ohio State University, which was founded in 1870, is the largest contemporary state-assisted university. It is also a land-grant college. A major graduate and professional centre, it also has one of the largest undergraduate enrollments in the nation. It maintains several regional campuses and a graduate program at Dayton. Many of Ohio's small independent colleges have made distinguished contributions to the state and have served as pioneers in education in various ways. Thus Oberlin College, founded 1833, became the first coeducational college in the United States and one of the first to admit Negroes. Antioch College, which was founded in 1852, is one of the nation's oldest experimental liberal arts colleges. Like several other institutions in the state, it has implemented innovative programs for advancing students from minority group backgrounds. The University of Cincinnati, established as Cincinnati College in 1819, is the nation's oldest municipal university.

Welfare and social services. The aged and poor, the blind and disabled, and crippled and dependent children are among the groups benefitting from the welfare activities of several state agencies. Other state bodies oversee programs in the prevention and cure of illness. A youth commission operates diagnostic and training centres, youth corps, and schools. State activities with labour and industry include programs in employment and unemployment services, industrial safety, and job-connected-injury compensation.

The development and use of natural resources fall under the jurisdiction of the Department of Natural Resources, which handles forestry, reclamation, parks, wildlife, and similar environmental concerns. The Department of Agriculture's responsibilities include economic and health problems of farm life and the regulation of foods through

pure-food-and-drug laws. Organized research is carried out at the Ohio State University and at the Ohio Agricultural Experiment Station in Wooster. A department of urban affairs cooperates with local and regional bodies and coordinates state programs bearing on community problems.

CULTURAL LIFE AND INSTITUTIONS

Early settlers of Ohio put the stamp of their former homes—New England, the Middle Atlantic states, Virginia, and Kentucky—upon certain sections of the state. Situated at the crossroads of the nation with an ever-changing society, Ohio never developed a distinctive state culture.

The arts. There has never been a clearly identifiable Ohio school in any of the arts, though there has been great activity in all of them. Such diverse Ohio writers as William Dean Howells, Ambrose Bierce, Paul Laurence Dunbar, Brand Whitlock, Charles F. Brown ("Artemus Ward"), David R. Locke ("Petroleum V. Nasby"), Sherwood Anderson, Louis Bromfield, and James Thurber drew upon their Ohio background. None of them, however, attempted to develop a distinctive regional character in his work.

When the log-cabin phase of early Ohio was ended, the settler was likely to follow the building styles he had known in his former home. In the Virginia Military District, where Southern influence was marked, the red-brick and stone houses were built in the Southern Federal architectural style. In the Western Reserve and the Marietta area, the New England influence was manifested in the austere lines of the Early American style and the later modified Georgian style. Later developments tended to follow the fashions of American architecture in general, most of them revivals of earlier European modes such as Greek, Gothic, and Romanesque. In recent years, in line with international trends, the emphasis has been on simplification and functionalism.

The performing arts: music and the theatre

The thousands of musical organizations in Ohio range from symphony orchestras to local choral societies. The symphony orchestra of Cleveland is among the finest in the world, and that of Cincinnati (which was considered the musical centre of the inland United States in the early days) is also renowned. Programs in music, theatre, dance, and the visual arts abound in Ohio's colleges and universities. With community theatres and arts centres, they serve as the cultural centres for many cities and towns. The Cleveland Play House (1915–16) and the Karamu Theatre (1915), an organization predominantly of black performers, also in Cleveland, have long enjoyed a national reputation. Cincinnati's Playhouse-in-the-Park (1960), noted for its experimentation, has become one of the United States' major regional professional theatres. The Ohio Arts Council, established in 1965 by the legislature, aids communities and arts organizations in stimulating greater interest as well as participation.

Libraries. Ohio has several hundred public libraries in addition to college and university facilities containing over 14,000,000 volumes and about 100 specialized libraries in many fields. The Ohio State Library, in Columbus, provides service to every Ohioan. Bookmobile service is a feature of rural areas.

Folk and popular culture. Reflecting Old World origins are the Welsh Eistedfodd festivities in Cleveland, Steubenville, Lima, Columbus, and Jackson and a German Saengerfest (Song Festival). More than 40 other nationality groups—including Hungarians, Bohemians, Swiss, Scandinavians, and Irish—present folk music and dances at festivals throughout the state. Other gatherings, in addition to the annual state and county fairs, include the Apple Festival in Jackson, the "River Days" Festival in Portsmouth, the Ohio Hills Folk Festival in Quaker City, and the Pumpkin Show in Circleville.

Sites of public interest. In the early 1970s Ohio had more than 100 museums of art, science, and history. The Cleveland Museum of Art ranked among the foremost art galleries in the nation, and those in Cincinnati, Toledo, Youngstown, and Columbus also held major art collections. In addition, many historical sites are maintained by state and local societies, including Indian mounds, old forts and battle sites, reconstructions of early settlements, and graves, homesteads, and similar memorials to Ohio's presidents and other leading citizens.

Recreational facilities include more than 125,000 acres of state park facilities, at 50 locations, for water sports, hiking and camping, or picnicking—in addition to numerous municipal recreational areas. Numerous public gardens, zoos, caves and caverns, and privately run amusement parks add to Ohio's recreational repertory.

Newspapers, broadcasting, and publishing

Communications. Ohio's newspapers have experienced numerous mergers and consolidations in recent decades, though some 450 of varying political views remain. They are more likely to carry syndicated columns expressing a spectrum of opinion than to be the organ of one party. Some newspapers have attempted to meet the competition of the broadcasting media by acquisition of stations. All of Ohio is well served by both commercial and noncommercial radio and television stations. Cleveland is a centre of publishing, and several of Ohio's universities publish through their own presses.

Research and development. A top-ranking state in scientific manpower, Ohio has a great number of laboratories maintained by specialized institutes, industries, educational institutions, and national and state agencies. Reflecting industrial concentrations, Akron is the world centre for rubber research, Cleveland for research in lighting. Glass, steel, soap, and electronics are among the many other products undergoing constant testing and evaluation. Federal centres include ones devoted to aviation medicine, aeronautics and space, atomic energy, agriculture, and forestry.

PROSPECTS

Ohio continues to reflect trends in the rest of the United States in its increasing concentration of population, in its loss of population from farms that are at the same time becoming vastly more productive, in the growing problems of the inner cities, and in the concern for the quality of the environment everywhere. Northern Ohio is already being considered as part of an urban agglomeration stretching west to Chicago from Buffalo, New York, and Pittsburgh, Pennsylvania. Similar concentrations have begun to develop through central Ohio around Columbus, along the Ohio River, and in the Miami Valley from Dayton to Cincinnati.

This demographic growth compounds both environmental and urban problems. Public attention has turned increasingly to programs to conserve natural resources, to reverse the pollution of rivers and lakes, to reclaim land scarred by strip-mining, to ensure clean air and water, and to rehabilitate the inner cities. In the southeast, relief programs have been undertaken for the rural poor. The effective treatment of problems of this character continue to challenge the best efforts of the people of Ohio, and of the nation.

BIBLIOGRAPHY. CARL F. WITTKE (ed.), *The History of the State of Ohio*, 6 vol. (1941–44), is the definitive work on this subject. EUGENE H. ROSEBOOM and FRANCIS P. WEISENBURGER, *A History of Ohio*, 2nd ed. (1967), the most authoritative one-volume account, contains a valuable annotated bibliography on Ohio. The *Ohio Historical Quarterly*, originally the *Ohio Archaeological and Historical Quarterly* (1887–), presents articles on a variety of Ohio historical subjects. Highly readable accounts of various aspects of Ohio history may be found in HARLAN HATCHER, *Buckeye Country*, rev. ed. (1947) and *Western Reserve: The Story of New Connecticut in Ohio* (1949); as well as in WALTER HAVIGHURST, *The Heartlands: Ohio, Indiana, Illinois* (1962). Ohio's economic geography and governmental processes are examined at length in ALFRED J. WRIGHT, *Economic Geography of Ohio*, 2nd ed. (1957), and in FRANCIS R. AUMANN and HARVEY WALKER, *The Government and Administration of Ohio* (1956). Current data on Ohio's economy, vital statistics, housing, education, science, religion, welfare, and the like may be obtained from a number of publications prepared by the ECONOMIC RESEARCH DIVISION OF THE DEVELOPMENT DEPARTMENT OF THE STATE OF OHIO, such as *Ohio Manufacturing* (1965); *Ohio's Economic Regions* (1964); *Ohio Population* (1968); *Ohio Economic Outlook* (1969); and the *Statistical Abstract of Ohio*, 2nd ed. (1969).

(F.R.A.)

Oils, Fats, and Waxes

Oils and fats are substances of plant or animal origin. They are nonvolatile, insoluble in water, and oily or greasy to the touch; they constitute, along with proteins and carbohydrates, one of the main kinds of foodstuffs, and they are widely distributed in nature. Chemically (see below for details), fats and oils are compounds produced by reaction between an alcohol called glycerol and certain members of a group of compounds called fatty acids. Waxes are the esters that result from a reaction between fatty acids and certain alcohols other than glycerol, either of a group called sterols, such as cholesterol, or an alcohol that contains 12 or more even numbers of carbon atoms in a straight chain, such as cetyl alcohol. Essential oils are volatile materials isolated by a physical process from an odoriferous plant of a single botanical species, such as oil of cloves. The dominant chemical compounds are terpenes, organic compounds consisting of multiples of isopentane units (*i.e.*, containing five carbon atoms), but many other classes of chemical constituents have been observed in various essential, volatile, or ethereal oils.

The term oils is used in a generic sense to describe all substances that are greasy or oily fluids at ordinary temperatures. The term also can refer (1) to the individual fatty or fixed (*i.e.*, nonvolatile) oils, such as olive or soybean oils; (2) to the hydrocarbon or mineral oils and derivatives, such as petroleum, shale oils, oils from the low-temperature distillation of coal, fuel oils, and lubricating oils; (3) to the odoriferous and volatile essential oils; and (4) to synthetic materials that possess the characteristic of oiliness or lubricity; for example, silicone oils. The mineral or hydrocarbon oils and synthetic oils do not come within the scope of this article.

This article is divided into the following main sections and subsections:

I. Fats and oils

To understand reactions between organic compounds, a general knowledge of the nomenclature is necessary. Organic molecules consist of carbon atoms (each having four bonds) linked with single, double, or triple bonds to one another in straight chains, in various kinds of branching chains, and with the ends of the chain bonded together to form what are called cyclic compounds. The number of atoms in an organic molecule may reach tens of thousands and contain many kinds of atoms other than carbon. Organic compounds are classified by various criteria into families with similar structures and properties, the largest being the hydrocarbons, which contain only hydrogen and carbon, but each member of which has thousands of derivatives containing other kinds of atoms as well and each recognizable fragment of which—called a radical—participates in countless reactions. The carbohydrates, another large family, always contain oxygen as well as hydrogen, and they, too, have an enormous number of derivatives and participate, as free radicals in reactions with one another, with hydrocarbons, and other types of organic molecules. The vast majority of organic compounds have what are called functional groups; that is, groups of atoms that always react in the same way whatever radical they are bonded to. Some of the most important functional groups, when attached to a radical, form the subfamilies of compounds called alcohols, carboxylic acids, ketones, amines, ethers, and esters. All alcohols have a functional group consisting of an oxygen atom, one of whose two bonds is linked with a hydrogen atom, the other to a carbon atom in the chain. This is called the hydroxyl group and in formulas it is written R—OH, R being the radical to which it is attached. A molecule may have many such hydroxyl groups. The group forming carboxylic acids consists of a carbon atom double-bonded to an oxygen atom and single-bonded to a hydroxyl group, leaving the carbon with one bond free to attach to any other carbon (or a hydrogen, as in formic acid, the simplest carboxylic acid) in a molecule. The formula is written $R'-\overset{\overset{O}{\|}}{C}-OH$, R′ being the radical. In a great many chemical reactions, it is the functional groups that react. An important type of reaction is called esterification; it results when any alcohol group reacts with any carboxylic acid group to produce a characteristic compound called an ester, which consists of the radical (R′) of a carboxylic acid joined to the radical (R) of an alcohol through the functional group of the carboxylic acid, with the carbon atom double-bonded to one oxygen and single-bonded to the oxygen of the alcohol thus:

$$R'-\overset{\overset{O}{\|}}{C}-O-R.$$

Oil and fat compositions and properties

Oils and fats consist primarily of glycerides, which are esters formed by the reaction of three molecules of fatty acids with one molecule of glycerol to yield one molecule of triglyceride and three molecules of water. When the three fatty acids are the same, the product is a simple triglyceride; when there are at least two different acids, the product is a mixed triglyceride. Simple triglycerides rarely occur in nature.

Oils are usually liquids at ordinary temperatures, such as 25° C (77° F); fats are solid at this temperature. There is no basis, however, for chemical distinction between the two. The difference is only a physical one that can be removed by changing the environment.

Use of the word fat to include both fats and oils is becoming common, especially among chemists. Thus, both soybean oil and tallow would be considered fats. This terminology avoids confusion with nonfatty oils. No sharp distinction can be made between fats and oils, because at sufficiently low temperature all oils will solidify, and at even moderately elevated temperatures all fats will liquefy. Moreover, a fat is never entirely solid except under unique circumstances. Fats have been defined as plastic solids consisting of a mixture of crystalline particles and liquid oil.

Lipid (lipide) is a much broader term that may be used to include all of the ether-soluble, water-insoluble substances obtainable from plant and animal sources. An abbreviated definition is: (1) simple lipids are esters of fatty acids and various alcohols, classified as (a) fats and oils—*i.e.*, glycerides (fatty-acid esters of glycerol); and (b) waxes—*i.e.*, fatty-acid esters of alcohols other than glycerol. (2) Compound lipids are esters of fatty acids and alcohols containing additional groups, divided into (a) phospholipids (phosphatides)—*i.e.*, fatty-acid esters containing a phosphoric acid group (they also usually contain a nitrogenous group); (b) glycolipids—*i.e.*, compounds that consist of fatty acids, a carbohydrate, and a nitrogen-containing compound but no phosphoric acid group; and (c) others, such as sulfolipids and amino lipids. (3) Derived lipids are compounds derived from the preceding groups and having the general properties of the lipids; the derived compounds include (a) fatty acids; (b) alcohols, including glycerol, sterols, and the long-

chain alcohols from the waxes; and (c) others, including hydrocarbons such as squalene found in fish livers, and nitrogenous bases from the phospholipids (see also LIPID).

The fats comprise one of the three classes of food. Nearly all cells contain fat, protein, and carbohydrate. Fat is sometimes called nature's storehouse of energy because on a weight basis it contains more than twice as much energy as does carbohydrate or protein. It is probably as storehouses or depots of concentrated energy that fats appear in plant reproductive organs such as pollen grains and seeds. It is this fat that man recovers from plants for use as food or in industry. The fat content of the nonreproductive tissue of plants is usually so low that recovery is impracticable. Yet much of man's dietary fat comes from natural foodstuffs without being separated from the other plant materials with which it occurs. The proportion of fat in these foodstuffs varies from 0.1 percent in white potatoes to 70 percent in some nut kernels.

More than 90 percent of the fat recovered in the world is obtained from about 20 species of plants and animals. Most of this separated fat is used eventually as human food. Consequently, fat technology deals largely with the separation and processing of fats into forms acceptable to the various dietary customs in the countries in which they are to be used.

HISTORY OF USE

Early food and non-food uses

Man has used many natural fats for both food and nonfood purposes since prehistoric times. The Egyptians used olive oil as a lubricant in moving heavy building materials. They also made axle greases from fat and lime, mixed with other materials, as early as 1400 BC. Homer mentions oil as an aid to weaving and Pliny talks about hard and soft soaps. Candles and lamps using oil, tallow, or beeswax have been used for thousands of years.

Waxes and oils that had dried to hard films were used in protective or decorative coatings on walls and mummy cases, and as waterproofing agents for wooden ships. The form of painting known as encaustic employed a mixture of pigments in natural waxes. Tempera, another early form of painting, can be considered a forerunner of the modern emulsion paints. It was a water emulsion of wax or oil and pigments, stabilized with vegetable gums or egg yolk (see also PAINTS, VARNISHES, AND ALLIED PRODUCTS).

Apparently the first mention of the use of a drying oil as a protective coating was made in about the 6th century AD, after which the art grew rapidly. Transparent varnishes were made of linseed oil and natural resins. Chemical driers, however, were not used until much later.

The primary uses of fats for nonedible purposes are much the same as they were centuries ago, although the efficiencies of use have improved. The commercial uses of fats have increased in number as the understanding of the chemical nature of fats has expanded. C.W. Scheele, a Swedish chemist, discovered in 1779 that glycerol could be obtained from olive oil by heating it with litharge (lead monoxide), but it was not until about 1815 that the French chemist Michel-Eugène Chevreul (1786–1889) demonstrated the chemical nature of fats and oils. A few years later the separation of liquid acids from solid acids was accomplished. Margarine was invented by the French chemist Hippolyte Mège-Mouriès, who in 1869 won a prize offered by Napoleon III for a satisfactory butter substitute. The modern hydrogenation process had its origin in research in the late 19th century that led to the establishment of the vegetable-oil-shortening industry and a variety of industrial applications.

After World War I, organic chemists gained extensive knowledge first of fatty-acid compositions and then of glyceride compositions. Growth of the chemical industry stimulated a simultaneous expansion of the use of fats as raw materials and as intermediates for scores of new chemicals. The modern application of many organic chemical reactions to fats and fatty acids formed the foundation of a new and rapidly growing fatty-chemicals industry.

FUNCTIONS IN PLANTS AND ANIMALS

Fats as reserves and agents

The universal distribution of fats in plant and animal tissues suggests physiological roles that go beyond their function as a fuel supply for the cells. In animals the most evident function of fats is that of a food reserve to supply energy (through subsequent enzymatic oxidation—that is, combination with oxygen catalyzed by enzymes). The storage of fat in vegetable seeds can be explained similarly on the basis that it is a food reserve for the embryo. It is not so easy, however, to account for the presence of large quantities of fat in such fruits as olives, avocados, and palms; much of this fat is probably lost or destroyed before the seed germinates. Fats, and especially the waxes, fulfill other valuable functions in plants and animals. Subcutaneous deposits of fat insulate animals against cold because of the low rate of heat transfer in fat, a property especially important for animals living in cold waters or climates; *e.g.*, whales, walruses, and bears. Beeswax prevents dilution or contamination of concentrated sugar solutions in the comb. Waxes, and in some cases fats, secreted on the surface of plant leaves protect the underlying tissues against loss of water by transpiration. Apples, citrus fruits, and melons have natural protective wax coatings.

Fats that have been separated from tissues always contain small quantities of closely associated nonglyceride lipids such as phospholipids, sterols, vitamins A, D, and E, and various carotenoid pigments. Many of these substances are vital emulsifying agents or growth factors. Others function as agents that prevent deterioration of fats in plant tissues and seeds caused by destructive combination with oxygen. These minor constituents probably are present in the fats as a result of their physical solubility, and thus fats serve as carriers for these substances in animal diets.

Many animals require some fat containing one or more of the essential fatty acids (linoleic, arachidonic, and to a limited extent linolenic) to prevent the physical symptoms of essential-fatty-acid deficiency manifested by skin lesions, scaliness, poor hair growth, and low growth rates. These essential fatty acids must be supplied in the diet since they cannot be synthesized in the body.

The prostaglandins, discovered by the Nobel laureate U.S. von Euler of Sweden, are hormones derived from arachidonic acid. Despite intensive international research, in the early 1970s the functions of prostaglandins were understood only vaguely. These biologically active fatty acids, which are present in very minute quantities in animal tissues, apparently are involved in contraction of smooth muscles, enzyme activity in lipid metabolism, function of the central nervous system, regulation of pulse rate and blood pressure, function of steroid hormones, fat mobilization in adipose tissue, and a number of other vital functions. A prominent exploratory area is the study of prostaglandins as a means of once-a-month birth control.

SYNTHESIS AND METABOLISM IN LIVING ORGANISMS

Formation of fats in seeds and fruits occurs late in the ripening process. Sugars and starches predominate in fruits, seeds, and sap in the unripe condition. These apparently are converted by enzymes during the maturing process to fatty acids and glycerol, which then form glycerides. Studies with radioactive-tracer techniques confirm the synthesis of fats from carbohydrates in both plants and animals. In fact, it has been shown by the use of labelled acetic acid, or acetate, ions that any food source from which acetate ions may form as an intermediate metabolite can be converted to fatty acids in at least some animal tissues. It has been further demonstrated that acetate can be converted to cholesterol in animal tissue. It is noteworthy that, almost without exception, natural fats contain only fatty acids with an even number of carbon atoms. These acids apparently are built up of two-carbon units. Although the preponderance of fatty acids with 18 carbon atoms has suggested the hypothesis that fats are derived from three molecules of glucose (a carbohydrate with six carbon atoms), later discoveries through tracer studies have indicated the buildup from

Buildup of two-carbon units

Table 1: Common Fatty Acids

common name	systematic name	formula	carbon atoms	double bonds	melting point (°C)
Caprylic	octanoic	$C_7H_{15}COOH$	8	0	16.5
Capric	decanoic	$C_9H_{19}COOH$	10	0	31.5
Lauric	dodecanoic	$C_{11}H_{23}COOH$	12	0	44
Myristic	tetradecanoic	$C_{13}H_{27}COOH$	14	0	58
Palmitic	hexadecanoic	$C_{15}H_{31}COOH$	16	0	63
Stearic	octadecanoic	$C_{17}H_{35}COOH$	18	0	72
Arachidic	eicosanoic	$C_{19}H_{39}COOH$	20	0	77
Oleic	*cis*-9-octadecenoic	$C_{17}H_{33}COOH$	18	1	13.4
Linoleic	*cis*-9, *cis*-12-octadecadienoic	$C_{17}H_{31}COOH$	18	2	−5
Linolenic	*cis*-9, *cis*-12, *cis*-15-octadecatrienoic	$C_{17}H_{29}COOH$	18	3	−11.3
Eleostearic	*cis*-9, *cis*-11, *cis*-13-octadecatrienoic	$C_{17}H_{29}COOH$	18	3	49
Ricinoleic	12-hydroxy-*cis*-9-octadecenoic	$C_{17}H_{33}OCOOH$	18	1+OH	16
Arachidonic	5,8,11,14-eicosatetraenoic	$C_{19}H_{31}COOH$	20	4	−49.5
Erucic	*cis*-13-docosenoic	$C_{21}H_{41}COOH$	22	1	33.5

the two-carbon acetate units. Since acetate can be formed from fats, proteins, or carbohydrates by reaction with oxygen, it is thus possible for fats to be synthesized indirectly from any of these sources. The formation of multiple linkages between carbon atoms (double bonds) in the fats synthesized from acetate is accomplished (probably in the liver) by addition or removal of hydrogen atoms through the action of enzymes.

Utilization of stored fat by plant embryos has not been entirely explained, but it is known that in germinating embryos the glycerides are hydrolyzed—that is, decomposed to glycerol and fatty acids—by lipolytic (fat-splitting) enzymes. These may pass through oxidative processes to form intermediate metabolic products that can be oxidized further to carbon dioxide and water or can be converted to carbohydrates, which may then pass through the many steps of carbohydrate metabolism.

In animal digestive tracts, the fats in foods are emulsified with digestive secretions containing lipase, an enzyme that hydrolyzes at least part of the glycerides. The glycerol, partial glycerol esters, fatty acids, and some glycerides are then absorbed through the intestine and are at least partially recombined to form glycerides and phospholipids. The fat, in the form of microscopic droplets, is transported in the blood to points of use or storage. The fat of an individual animal may vary somewhat according to the composition of fats in the food. The body fat of swine that are fed cod-liver oil, for instance, is softer than normal and has some long-chain fatty acids containing more than 20 carbon atoms, characteristic of the dietary fat.

Fats used by or stored in animal tissues come from two sources—enzymatic synthesis and diet. The fat synthesized from carbohydrates or proteins is characteristic for each animal species. Most dietary fat is broken up by enzymes to small intermediates followed by enzymatic resynthesis to form the fat characteristic of the animal, but some dietary fatty acids are absorbed directly and recombined in the body fat. Mammary glands apparently have enzyme systems that can produce fats quite different from those stored in adipose tissue, but even so, milk fats still can reflect unusual components of ingested fats. It has also been found that the fatty-acid composition of milk fat can be altered substantially by feeding cows or goats fat-containing products that have been coated to prevent their combination with oxygen and resynthesis in the first rumen. Thus, the feeding of encapsulated polyunsaturated fats (fats containing two or more double carbon–carbon bonds in their molecular structure) such as corn oil or soybean oil to ruminants greatly increases the polyunsaturated-fatty-acid content of the milk fat.

The manner in which fat reserves are circulated to the organs where metabolism occurs is incompletely understood. Radioactive-tracer studies provide some insight into this complicated process. It has long been established that when mobilization of reserve fat takes place the stream is directed primarily to the liver, where fatty acids may be partially desaturated; *i.e.*, hydrogen is removed from the fatty-acid chains to produce unsaturated or double bonds between carbon atoms. This apparently facilitates subsequent oxidation in other tissues.

Oxidation of fatty acids

Fatty acids also may be oxidized directly in the various tissues as well as in the liver. Fatty-acid metabolism is presumed to be by oxidation in successive two- and four-carbon stages. Intermediate products could be acetoacetate and acetate groups. If the mechanism is faulty, acetone is formed and excreted (acetonuria). The final products of normal metabolism are carbon dioxide and water.

In certain genetic diseases that cause mental retardation and early death, there is an unusual accumulation of phospholipids in tissues, especially in the brain. In these diseases (Gaucher's, Tay-Sachs, Fabry's, and Niemann–Pick) there is an inherited lack of certain enzymes involved in lipid metabolism.

The relationship between dietary fats and the disease atherosclerosis, a form of arteriosclerosis, has been much studied but is little understood. It was observed empirically that there was a correlation between the content of cholesterol in blood serum and the incidence of atherosclerosis. Individuals with high blood-cholesterol levels (hypercholesterolemia) were frequently subject to atherosclerosis, a cardiovascular disease characterized by the accumulation of fatty material in the arterial system. In such cases partial replacement of solid animal fats in the diet by liquid fats resulted in a lowering of the serum cholesterol level in short-term feeding studies. Dietary interest was focussed on the polyunsaturated fatty acids —those with two or more double bonds in each molecule. The chief polyunsaturated acids considered in this connection were linoleic (two double bonds), linolenic (three), and arachidonic (four). The oils with the highest ratios of polyunsaturated to saturated acids (*P/S* ratio) are, in decreasing order: safflower, sunflower, corn, soybean, and cottonseed. Cholesterol, an alcohol, is rarely present in the free state but is combined with fatty acids. It was postulated that the higher melting cholesterol esters of solid fatty acids would be more difficult to transport and metabolize than the lower melting cholesterol esters of liquid fatty acids, and, as a result, a change in the dietary fat should reduce the cholesterol level. Results of incorporating polyunsaturated fats as partial replacements for animal fats in the diet have been inconclusive. Some tests indicate possible merit, but one large-scale study extending over a period of three years and completed in 1969 failed to confirm any significant value of polyunsaturated fats in reducing atherosclerosis (see also METABOLISM).

CHEMICAL COMPOSITION OF FATS

Although natural fats consist primarily of glycerides, they contain many other lipids in minor quantities. Corn oil, for example, may contain glycerides plus phospholipids, glycolipids, phosphoinositides (phospholipids containing inositol), many isomers of sitosterol and stigmasterol (plant steroids), several tocopherols (vitamin E), vitamin A, waxes, unsaturated hydrocarbons such as squalene, and dozens of carotenoids and chlorophyll compounds, as well as many products of decomposition, hydrolysis, oxidation, and polymerization of any of the natural constituents.

Fatty acids contribute from 94 to 96 percent of the total weight of various fats and oils. Because of their prepon-

Table 2: Saturation and Unsaturation in Fatty Acids

Lauric acid	$CH_3–CH_2–CH_2–CH_2–CH_2–CH_2–CH_2–CH_2–CH_2–CH_2–CH_2–COOH$	a saturated fatty acid with 12 carbon atoms
Oleic acid	$CH_3(CH_2)_7CH=CH(CH_2)_7COOH$	an unsaturated fatty acid with one double bond and 18 carbon atoms
Linoleic acid	$CH_3(CH_2)_4CH=CHCH_2CH=CH(CH_2)_7COOH$	an unsaturated fatty acid with two double bonds and 18 carbon atoms
Linolenic acid	$CH_3CH_2CH=CHCH_2CH=CHCH_2CH=CH(CH_2)_7COOH$	an unsaturated fatty acid with three double bonds and 18 carbon atoms
Arachidonic acid	$CH_3(CH_2)_4CH=CHCH_2CH=CHCH_2CH=CHCH_2CH=CH(CH_2)_3COOH$	an unsaturated fatty acid with four double bonds and 20 carbon atoms

Effect of fatty acids on properties

derant weight in the glyceride molecules and also because they comprise the reactive portion of the molecules, the fatty acids influence greatly both the physical and chemical character of glycerides. Fats vary widely in complexity; some contain only a few component acids, and at the other extreme more than 100 different fatty acids have been identified in butterfat, although many are present in only trace quantities. Most of the oils and fats are based on about a dozen fatty acids (see Table 1). In considering the composition of a glyceride it is particularly important to distinguish between the saturated acids (acids containing only single bonds between carbon atoms, such as palmitic or stearic), with relatively high melting temperatures, and the unsaturated acids (acids with one or more pairs of carbon atoms joined by double bonds, such as oleic or linoleic), which are low melting and chemically much more reactive.

In the series of saturated acids, the melting point increases progressively from below room temperature for the acids of lower molecular weight to high melting solids for the longer chain acids. Unsaturated acids may contain up to six double bonds, and as unsaturation increases the melting points become lower. Glycerides based predominantly on unsaturated acids, such as soybean oil, are liquids; and glycerides containing a high proportion of saturated acids, such as beef tallow, are solids. The carbon atoms in fatty acids are arranged in straight chains, and the first site of unsaturation (double bond) in most of the unsaturated acids appears between the ninth and tenth carbon atoms, starting the counting from the terminal carboxyl group (see Table 2). The specificity of location of unsaturation in fatty acids from both plant and animal sources suggests formation by a related enzymatic dehydrogenation mechanism.

During the 1960s, several revolutionary separation techniques, especially gas–liquid chromatography and thin-layer chromatography (methods used to separate and identify mixtures of chemical compounds), improved the precision of analysis and led to the discovery and characterization of a number of new and unusual fatty acids. Natural fatty acids were found with such structures as triple (acetylenic) bonds; hydroxyl and ketone groups (RC=O, where C is a carbon atom, O an oxygen atom, and R an alkyl group); and epoxy (a closed ring group consisting of two alkyl groups and an oxygen atom), cyclopropane, and cyclopentane rings. For years it was thought that fats contained only straight-chain fatty acids with an even number of carbon atoms, but it was demonstrated that butterfat and the depot fats of cattle and sheep contain small quantities (1.5 to 2 percent) of saturated and traces of unsaturated straight-chain acids with odd numbers (11, 13, 15, 17, 19) of carbon atoms. The same studies uncovered both odd- and even-numbered fatty acids containing 13, 14, 15, 16, and 17 carbon atoms and a methyl branch as minor components (0.5 to 1 percent) of these fats.

Since the glycerides, which make up 90 to 99 percent of most individual fats or oils of commerce, are esters formed by three fatty-acid molecules combining with one molecule of glycerol, they may differ not only in the fatty acids that they contain but also in the arrangement of the fatty-acid radicals on the glycerol portion. Simple triglycerides are those in which each molecule of glycerol is combined with three molecules of one acid; *e.g.*, tripalmitin, $C_3H_5(OCOC_{15}H_{31})_3$, the glyceryl ester of palmitic acid, $C_{15}H_{31}COOH$. Only a few of the glycerides occurring in nature are of the simple type; most are mixed triglycerides; *i.e.*, one molecule of glycerol is combined with two or three different fatty acids. Thus stearodipalmitin, $C_3H_5(OCOC_{15}H_{31})_2(OCOC_{17}H_{35})$, contains two palmitic acid radicals and one stearic acid radical. Similarly, oleopalmitostearin, $C_3H_5(OCOC_{15}H_{31})(OCOC_{17}H_{33})(OCOC_{17}H_{35})$, contains one radical each of oleic, palmitic, and stearic acids. Each mixed triglyceride containing three different acid radicals may exist in three different isomeric forms (molecules having the same composition but different structures), because any of the three can be linked with the centre carbon of the glycerol molecule. A mixed triglyceride containing two radicals of the same acid and one radical of another acid has only two isomeric forms.

Monoglycerides and diglycerides are partial esters of glycerol and have one or two fatty-acid radicals, respectively. They are seldom found in natural fats except as the products of partial hydrolysis of triglycerides. They are easily prepared synthetically, however, and have important applications mainly because of their ability to aid in the formation and stabilization of emulsions. As constituents of shortening in baked products they increase product volumes, improve tenderness, and retard staling. They also have technical importance as intermediates in the manufacture of coatings and resins.

Glyceride distribution in natural fats

The glyceride composition of natural fats tends to follow the rule of even distribution, which proposes that each of the individual fatty-acid radicals in a fat is apportioned evenly among the different glyceride molecules. Consequently, specific acids present in quantities less than one-third of the total acids are inclined to appear singly in the glycerides, and it is necessary to have an acid present in a quantity greater than two-thirds of the total acids before any simple triglycerides would be present. Vegetable-seed fats appear to follow this rule much more closely than do animal fats.

Some fats tend to follow the pattern of random, or statistical, distribution of fatty acids in their glycerides. In this pattern, the acids are distributed among the glycerides as if there were no directing influences and they had organized themselves by chance alone. In beef tallow, for example, which contains slightly more than 50 percent of saturated acids, the chance of a saturated acid appearing on any glyceryl group is ½, and the chance of three saturated acids combining with glycerol is $½ \times ½ \times ½ = ⅛$. The observed quantity of trisaturated glycerides is fairly close to the 12.5 percent level calculated for random distribution.

With the development of more sophisticated analytical tools, including specific enzymatic hydrolysis, it was found that fats deviate significantly from both even and random distribution patterns, and several "restricted random" theories have been proposed; these relate the fatty-acid distribution to enzyme-controlled biosynthesis.

Fats that are heated in the presence of certain alkaline catalysts, such as sodium methoxide, "rearrange" themselves to follow more closely the system of random distribution. Physical properties are often altered significantly. Some manufacturers in the United States, for example, rearrange the molecules in lard and other fats to make improved shortenings and margarines.

Table 3: Sources, Iodine Value, and Uses of Principal Oils and Fats

	principal sources of raw material	iodine value	principal uses
Vegetable oils			
Drying oils			
Parilla	China, Korea, Japan, India	193–208	paint, varnish
Linseed	U.S., Argentina, India, Canada, U.S.S.R.	175–205	paint, varnish, linoleum, printing ink
Tung	China, Japan, U.S.	160–175	paint, varnish
Oiticica	Brazil	139–155	paint, varnish
Semidrying oils			
Poppyseed	Levant, India	132–143	salad oil, artists' oil, soft soap
Safflower	U.S., India	130–150	salad oil, paints, resins
Soybean	U.S., China, Manchuria	125–140	food, paint, resins, chemicals
Corn (maize)	U.S., Argentina, Europe	115–130	food
Sunflower	U.S.S.R., South America, Eastern Europe	125–136	food, resins
Cottonseed	U.S., India, Egypt, Mexico, U.S.S.R.	103–116	food, soap
Sesame	India, Egypt, Levant	103–118	food, soap
Rape (colza)	India, Europe, Canada, Pakistan	94–105	food, lubricant
Nondrying oils			
Almond	southern Europe, North Africa	93–100	perfumery, pharmacy, food
Arachis (peanut, groundnut)	India, West Africa, China, U.S.	85–100	food
Olive	Mediterranean countries, U.S.	75–90	food, soap, lubricating, pharmacy
Castor	India, Mediterranean, Brazil, U.S.	80–90	medicine, lubricant, chemicals
Animal and marine oils			
Marine oils			
Anchovy	Peru	170–190	resins, leather currying, paints, food
Sardine	west coast of North America, Japan	160–190	resins, leather currying, paints, food
Menhaden	Atlantic coast of North America	150–180	resins, leather currying, paints, food
Herring	North Sea, Japan	120–145	leather currying, paints, food
Cod liver	North Sea, east coast of North America	120–180	vitamins, leather currying
Shark liver	coasts of North America	100–115	vitamins, leather currying
Seal	Arctic and Antarctic seas	130–150	food, leather currying, soap
Whale	Arctic and Antarctic seas	110–150	food, soap, fibre dressing, leather currying, greases
Sperm whale	west coast of South America	65–90	lubricating oil for delicate machinery
Terrestrial animal oils			
Neat's-foot	U.S., South America, Europe	67–75	lubricating, high-grade leather dressing
Vegetable fats			
Mahua (Illipe) butter	India, Malaysia	53–67	food, soap, candles
Shea butter	West Africa, The Sudan	53–65	food, soap, candles
Palm oil	West Africa, Indonesia, Malaysia, Brazil	50–60	soap, candles, tinplate industry
Cacao (cocoa) butter	Indonesia, Malaysia	32–42	chocolate, pharmacy, perfumery
Babassu oil	West Africa	12–16	food, soap
Coconut oil	Philippines, Indonesia, India, Sri Lanka, South American coasts	8–10	food, soap, chemicals
Japan wax	China, India, Japan	5–17	polishes
Animal fats			
Lard	U.S., central Europe	46–66	food, soap, pharmacy, chemicals
Bone	U.S., India, Europe	43–56	soap, candles
Tallow, beef	Argentina, U.S.	30–45	food, soap, candles, chemicals
Tallow, mutton	Australia	31–45	food, soap
Butter	U.S., Europe, Australia, Canada, U.S.S.R.	25–40	food

PHYSICAL AND CHEMICAL PROPERTIES

Fats and oils may be divided into animal and vegetable fats according to source. Further, they may be classified according to their degree of unsaturation as measured by their ability to absorb iodine at the double bonds. This degree of unsaturation determines to a large extent the ultimate use of the fat, as shown in Table 3. Oils having iodine values higher than about 150 (and therefore a high degree of unsaturation) are generally called drying oils and are used primarily in protective coatings. Those having iodine values of 100 to 150 are considered as semidrying and may be used either for food or in protective coatings. The nondrying oils, with the lowest amount of unsaturation, have iodine values generally below 100 and are used mainly in foods, soaps, chemicals, and specialty products.

Liquid fats (*i.e.*, vegetable and marine oils) have the highest degree of unsaturation, while solid fats (vegetable and animal fats) are highly saturated. Solid vegetable fats melting between 20° and 35° C (68° and 95° F) are found mainly in the kernels and seeds of tropical fruits. They have relatively low iodine values and consist of glycerides containing high percentages of such saturated acids as lauric, myristic, and palmitic. Fats from fruits of many members of the palm family, notably coconut and babassu oils, contain large amounts of combined lauric acid. Most animal fats are solid at ordinary temperatures; milk fats are usually characterized by the presence of short-chain carboxylic acids (butyric, caproic, and caprylic); and marine oils contain a large number of very long chain highly unsaturated acids containing up to six double bonds and up to 24 or even 26 carbon atoms.

Composition of vegetable oils

The fatty-acid composition of vegetable oils is strongly influenced by temperature during the period of seed maturation. Sunflower-seed oils from plants grown in the cool climate of the central Soviet Union or the plains area of the northern United States and southern Canada contain about 70 percent of linoleic acid, but oils from the same seed variety grown in the southern United States have half as much linoleic acid, with a corresponding increase in oleic acid. Even animals show effects of environmental temperature, and tallow from Indian cattle may have as much as 70 percent of saturated acids, compared with the normal 50 to 55 percent range for tallows from Europe, Argentina, and North America.

The plant breeder is beginning to play an important part in determining the chemical constitution of vegetable oils. Historically, the role of the plant breeder has been to improve resistance of plants to diseases and insect pests and to increase the yield of crops or of specific components of crops. As an outstanding example of this historical function, Soviet scientists, under the leadership of academician V.S. Pustovoyt, have increased the oil content of sunflower seed from 28 to 50 percent, a change that catapulted sunflower oil to second place behind soybean oil in worldwide importance. By genetic manipulation, breeders have created essentially new oilseed crops. Safflower contains about 11 percent of saturated acids, 12 percent oleic, and 77 percent linoleic. In the United States, a new crop, "high-oleic safflower," was developed; it has a composition of 80 percent oleic and 10 percent linoleic acids in the unsaturated fraction, a complete reversal of the normal pattern. Similarly, in Canada a new type of rapeseed, Canbra, was created in which oleic

acid replaces the 10 to 15 percent of eicosenoic acid and the 40 to 50 percent of erucic acid present in conventional rapeseed oil.

Specific gravities of oils and fats range from 0.913 (rapeseed oil) to 0.975 (Japan wax, myrtle wax); for most fats the value is 0.915 to 0.945.

Fats are practically insoluble in water and, with the exception of castor oil, are insoluble in cold alcohol and only sparingly soluble in hot alcohol. They are soluble in ether, carbon disulfide, chloroform, carbon tetrachloride, petroleum benzin, and benzene. Oils and fats have no distinct melting points or solidifying points because they are such complex mixtures of glycerides, each of which has a different melting point. Glycerides, further, have several polymorphic forms with different melting or transition points. The freezing points of the oils range from a few degrees above zero to about 30° C below zero (−22° F). At low temperatures (*e.g.*, 12° C, or 54° F, for cottonseed oil), solid portions, termed stearine, separate from many oils. Formerly, commercial separation was carried out simply by allowing oils such as cottonseed oil or fish oils to stand in outside storage tanks during the winter until higher melting components had settled out, leaving a clear supernatant layer of winter oil. Such oils remain limpid at low temperatures and are especially valued for use in salad dressings and mayonnaise. Mechanical refrigeration is now used in the preparation of winterized oils.

Effects of low temperatures

Fats can be heated to between 200° and 250° C (392° and 482° F) without undergoing significant changes provided contact with air or oxygen is avoided. On being heated above this temperature, the more unsaturated oils gradually polymerize and become considerably more viscous. When this is done commercially in the protective-coating industry, the process is called bodying. Castor oil, when heated to high temperature or under suitable conditions in the presence of catalysts, loses a molecule of water from each ricinoleic acid radical to form what is called dehydrated castor oil. This modified oil is used in the protective-coating industry for manufacture of light-coloured finishes that retain their colour during the service life of the coating. Above 300° C (572° F) fats may decompose, with the formation of acrolein (the decomposition product of glycerol), which has the pungent odour of burning fat. Hydrocarbons also may be formed at high temperatures.

On exposure to air, oils and fats gradually undergo certain changes. The drying oils absorb oxygen (dry) and polymerize readily; thin layers form a skin or protective film. The semidrying oils absorb oxygen more slowly and are less useful as paint oils. Still, sufficient oxygen is absorbed in time to produce distinct thickening and some film formation. Oxidation of the drying and semidrying oils is accelerated by spreading the oil over a large surface. On greasy cloths, for example, oxygen absorption may proceed so rapidly that spontaneous combustion ensues. The nondrying oils, of which olive oil is typical, do not oxidize readily on exposure to air, although changes do take place gradually, including slow hydrolysis (splitting to fatty acids and glycerol) and subsequent oxidation. This slow oxidation causes a disagreeable smell and taste described by the term rancidity.

The chemical reactions involved in oil oxidation have been studied widely. When oils and fats are exposed to air, little change takes place for a period of time that varies from oil to oil depending upon the amount and type of unsaturation and the content of natural antioxidants. During this so-called induction period, there is virtually no change in either odour or chemistry of the oil because of the protective effect of natural antioxidants, especially tocopherol. Gradually, the effectiveness of the antioxidant is overcome and there is an accelerating rate of oxidation of unsaturated acids, called autoxidation. Chemically, the first identifiable oxidation products are hydroperoxides. These break down into a large variety of low-molecular-weight aldehydes, esters, alcohols, ketones, acids, and hydrocarbons, some of which possess the pungent, disagreeable odours characteristic of rancid fats. In soybean oil exposed to air to the point of incipient rancidity, more than 100 different oxidation products have been identified. Natural oils such as coconut oil, with very low levels of unsaturation, are very stable to flavour deterioration, but the more highly unsaturated oils such as soybean oil or safflower oil lose their flavour more quickly. Sesame oil is unique in its flavour stability because of the presence of several natural antioxidants (sesamin, sesamolin, sesamol). Synthetic antioxidants such as propyl gallate, butylated hydroxyanisole (BHA), and butylated hydroxytoluene (BHT) have been used to retard the onset of rancidity and increase the storage life of edible fats.

When air is bubbled through heated drying or semidrying oils, oxygen is absorbed, and the oils polymerize to viscous products. These "blown oils" are used in various ways in the protective-coating industry. Oils blown at lower temperatures are used as plasticizers for nitrocellulose lacquers, and products of higher temperature processing are useful in caulks, putties, and paints.

Fats are hydrolyzed readily. This property is used extensively in the manufacture of soaps and in the preparation of fatty acids for industrial applications. Fats are hydrolyzed by treatment with water alone under high pressure (corresponding to a temperature of about 220° C [428° F]) or with water at lower pressures in the presence of caustic alkalies, alkaline-earth metal hydroxides, or basic metallic oxides that act as catalysts. Free fatty acids and glycerol are formed. If sufficient alkali is present to combine with the fatty acids, the corresponding salts (known popularly as soaps) of these acids are formed, such as the sodium salts (hard soap) or the potassium salts (soft soaps). Soaps of certain heavy metals, especially the lead, cobalt, and manganese salts, are called driers and are used to speed the rate of film formation of paints based on drying and semidrying oils.

Hydrolysis of fats

GENERAL METHODS OF EXTRACTION

The raw materials for the fat and oil industry are animal by-products from the slaughter of cattle, hogs, and sheep; fatty fish and marine mammals; a few fleshy fruits (palm and olive); and various oilseeds. Most oilseeds are grown specifically for processing to oils and protein meals, but several important vegetable oils are obtained from by-product raw materials. Cottonseed is a by-product of cotton grown for fibre, and corn oil is obtained from the corn germ that accumulates from the corn-milling industry, whose primary products are corn grits, starch, and syrup.

Fats may be recovered from oil-bearing tissues by three general methods, with varying degrees of mechanical simplicity: (1) rendering; (2) pressing with mechanical presses; and (3) extracting with volatile solvents.

Rendering. The crudest method of rendering oil from oleaginous fruits, still practiced in some countries, consists of heaping them in piles, exposing them to the sun, and collecting the oil that exudes. In a somewhat improved form, this process is used in the preparation of palm oil; the fresh palm fruits are boiled in water and the oil is skimmed from the surface. Such processes can be used only with seeds or fruits (such as olive and palm) that contain large quantities of easily released fatty matter. The rendering process is applied on a large scale to the production of animal fats such as tallow, lard, bone fat, and whale oil. It consists of cutting or chopping the fatty tissue into small pieces that are boiled in open vats or cooked in steam digesters. The fat, gradually liberated from the cells, floats to the surface of the water, where it is collected by skimming. The membranous matter (greaves) is separated from the aqueous (gluey) phase by pressing in hydraulic or screw presses; additional fat is thereby obtained. The residue is used for animal feed or fertilizer. Several centrifugal separation processes were developed in the 1960s. Cells of the fatty tissues are ruptured in special disintegrators under close temperature control. The protein tissue is separated from the liquid phase in a desludging type of centrifuge, following which a second centrifuge separates the fat from the aqueous protein layer. Compared with conventional rendering, the centrifugal methods provide a higher yield of better

quality fat, and the separated protein has potential as an edible meat product.

Tall-oil manufacture

In the manufacture of tall oil an unusual isolation technique is used that is somewhat related to rendering. In the sulfate pulping process for the manufacture of paper from pinewood, a dark-coloured, strong-odoured soap mass separates as a by-product. This separated soap is acidified with sulfuric acid, and crude tall oil, a mixture of about 60 percent of fatty acids and 40 percent of resin acids, separates and is skimmed off. Commonly, the crude product is distilled to yield the distilled tall oil of commerce, or it may be fractionally distilled, in which case two products are obtained. The fatty acids derived from tall oil consist of 55 to 60 percent of oleic acid, 35 to 40 percent linoleic acid, and from 1 to 3 percent of resin acids. The tall-oil rosin is a mixture of various cyclic resinous acids related to abietic acid. United States production of tall oil exceeds 1,200,000,000 pounds, and both the distilled product and the various fractionated grades find use in coatings manufacture, paper sizes, and miscellaneous industrial products.

Pressing. With many oil-bearing seeds and nuts, rendering will not liberate the oil from the cellular structures in which it is held. In these cases the cell walls are broken by grinding, flaking, rolling, or pressing under high pressures to liberate the oil. Many different mechanical devices have been used. The most primitive method is to crush seed in mortars until the oil exudes. In India, the ghani, a bullock-powered mortar and pestle, is used to extract oil from rapeseed and various local nuts and seeds as a cottage-industry type of operation. After the seeds are ground for several hours, the liberated oil is soaked up in rags and removed by hand wringing. There are more than 100,000 ghanis, supplementing the output of modern mills that use mechanical presses or solvent extraction. The lever and wedge, one step more advanced and employed for centuries, is still used in some places. The Romans developed a screw press, described by Pliny, for the production of olive oil. Centuries ago, the Chinese employed the same series of operations followed in modern pressing mills; viz., bruising or grinding the seeds in stone mills, heating the meal in open pans, and then pressing out the oil in a wedge press. The Dutch, or stamper, press invented in the 17th century was used almost exclusively in Europe for pressing oilseeds until the early part of the 19th century, when the hydraulic press was developed. The yield of oil from the hydraulic press was considerably higher than that from earlier processing methods because of the much higher applied pressures. In open presses, the ground seed material was confined in cloths of human hair or, less commonly, camel hair. Pressures on the cake varied from 1,000 to 2,000 pounds per square inch, and in the closed-type press, in which the oil-containing material was confined in a strong perforated steel cage during the pressing operation, pressures of 6,000 pounds per square inch or more were attained. Under ideal conditions the oil content of the hydraulic-press cake can be reduced to about 3 percent, but in practical operation a 5 percent level is average. The modern screw press replaced many of the hydraulic presses because it is a continuous process, has greater capacity, requires less labour, and will generally remove more oil. As ground seed is fed continously into the mechanical press, a worm screw increases the pressure progressively as the material moves through a slotted barrel. Pressures from 10,000 to 30,000 pounds per square inch are attained, and the oil is squeezed out through the slots, leaving a cake containing 3 to 3.5 percent oil under optimum processing and 4 to 5 percent oil under average conditions.

The general sequence of modern operations in pressing oilseeds and nuts is as follows: (1) the seeds are passed over magnetic separators to remove any stray bits of metal; (2) if necessary, the shells or hulls are removed; (3) the kernels or meats are converted to coarse meal by grinding them between grooved rollers or with special types of hammer mills; and (4) they are pressed in hydraulic or screw presses with or without preliminary heating, depending on the type of oil-bearing material and the quality of oil desired. Oil expressed without heating contains the least amount of impurities and is often of edible quality without refining or further processing. Such oils are known as cold-drawn, cold-pressed, or virgin oils. Pressing the coarse meal while it is heated removes more oil and also greater quantities of nonglyceride impurities such as phospholipids, colour bodies, and unsaponifiable matter. Such oil is more highly coloured than cold-pressed oils. Residual meals are concentrated sources of high-quality protein and are generally used in animal feeds. Because certain meals (*e.g.*, castor-bean meal and tung-nut meal) are toxic, they are generally used in fertilizers. In some primitive areas most press cakes are used as fertilizers because of decomposition before or after processing.

Quality variations in pressed oils

Solvent extraction. Cakes obtained by pressing operations still retain 3 to 15 percent of residual oil. When the value of the oil is considerably greater as oil than as a part of the meal, it is desirable to obtain more complete extraction with solvents. Modern commercial methods of solvent extraction use volatile purified hydrocarbons, especially the various grades of petroleum benzin (commonly known as petroleum ether, commercial hexane, or heptane). In large-scale operations, solvent extraction is a more economical means of recovering oil than is mechanical pressing. In the United States and increasingly in Europe, there are many instances of simple petroleum benzin extraction of seeds, mainly soybeans. For seeds or nuts containing a higher oil content than soybeans it became customary to press the material in screw presses to remove a large proportion of the oil before extraction. Since this prepressing also ruptures the cellular structures of oil-bearing materials, most of the residual oil is easily removed with solvents.

A typical extraction system consists of (1) cleaning to remove tramp iron, dirt, foreign weed seeds, and stones; (2) removing hulls or cortex in cracking, aspirating, or screening operations; (3) cracking or rough grinding the kernels, meats, or prepressed cake; (4) steaming (tempering or cooking) of the meats; (5) flaking the small pieces between smooth flaking rolls; (6) extracting the oil with solvent; (7) separating the meal, or marc, from the oil-solvent solution, called miscella; and (8) removing the solvent from both the miscella and the marc. The marc may be toasted or pelletized, or both, for use in animal feeds. Most extracted meals contain less than 1 percent of residual oil. The amount varies depending upon the amount of prepressing, the type of material being extracted, and the efficiency of the extracting system.

An interesting combination of solvent extraction and rendering is the process in which a chlorinated solvent is mixed with wet fresh tissue and the mass is heated to the boiling point of the mixture of solvent and water. When condensation of the vapour occurs, the mixture separates; water is drawn off, and the solvent is returned to the extraction system. The resulting product consists of a dry (low-moisture) extracted meal and miscella containing most of the fat. The high cost of chlorinated solvents is a disadvantage, but the process has promising potentialities because of the feasibility of preparing extracted animal or fish meals that have not been subjected to high temperatures. With a rapidly rising world population there has been concern about protein malnutrition, particularly among children in the developing countries. Fish protein concentrate, FPC, a product prepared by solvent extraction and moisture removal from whole fish, has been suggested as a possible source of highly nutritious protein that could be made available from underutilized fisheries resources. In one process, isopropyl alcohol is used to extract oil from chopped or minced fish; the alcohol simultaneously dehydrates the flesh and extracts the oil; a second process uses a chlorinated solvent (ethylene dichloride) to accomplish oil extraction and dehydration; and a third makes use of petroleum naphthas rather than chlorinated solvents. Acceptance of FPC as a protein supplement would increase production of fish oils.

Fish protein concentrate

Solvent extraction was first practiced in Europe, using batch extractors for the recovery of additional oil from the residues obtained from mechanical pressing. The

greater efficiency of solvent extraction encouraged direct application to oilseeds, and the batch extractor gradually gave way to continuous units in which fresh flakes are added continuously and subjected to a counterflow of solvent. One of the earliest continuous extractors, and a type still considered to be one of the best, was the Bollman or Hansa–Mühle unit from Germany, in which solvent percolates through oilseed flakes contained in perforated baskets moving on an endless chain. After the extraction cycle is complete, the baskets of extracted flakes are dumped automatically and then refilled with fresh flakes to initiate another cycle. Many extractor designs have been proposed, but only a few have found wide acceptance. In the DeSmet extractor, popular in Europe and in a number of developing countries, a bed of flakes on an endless horizontal travelling belt is extracted by solvent percolation. The Blaw–Knox Rotocell has become the most popular extractor in the huge United States soybean industry. The flakes are conveyed into wedge-shaped segments of a large cylindrical vessel. Solvent percolating through the cells falls into the bottom of the extractor housing, where it is picked up by a series of pumps and recirculated countercurrent to the flakes.

Most European extraction plants have been smaller than those in the United States. They are usually multipurpose and can alternatively process several types of oilseeds, depending on the availability of raw materials and the market demands. In contrast, most extraction plants in the United States process only one kind of seed. Historically, oilseed processing was carried out by squeezing or crushing; today processors are still called "crushers," even though almost all processing is by solvent extraction. The size of the crushing plant is increasing rapidly. In 1951 there were 193 soybean extraction plants in the United States with an average capacity of 160 tons per day; in 1969 there were 132 units, with an average capacity of 580 tons, and new installations usually have capacities of 1,000 to 2,000 or more tons per day. World trends indicate a gradual transition to fewer mills with larger individual capacities.

PROCESSING OF EXTRACTED OIL

The extent of processing applied to fats depends upon their source, quality, and ultimate use. Many fats are used for edible purposes after only a single processing step; *i.e.*, clarification by settling or filtering. Most cold-pressed oils (for example, cold-pressed olive, peanut, and some coconut and sunflower oils) can be used in food products without further processing. Tremendous quantities of butter and lard are used without special treatment after churning or rendering. The growing demand for bland-tasting and stable salad oils and shortening, however, has led to extensive processing techniques. But in the less industrialized countries, processing often is limited by lack of facilities, a shortage of technically trained engineers, and added costs for chemicals, which are often in short supply.

Refining. The nonglyceride components contribute practically all of the colour and flavour to fats. In addition, such materials as the free fatty acids, waxes, colour bodies, mucilaginous materials, phospholipids, carotenoids, and gossypol (a yellow pigment found only in cottonseed oil) contribute other undesirable properties in fats used for edible purposes and, to some extent, for industrial applications. Many of these can be removed by treating fats at 40° to 85° C (104° to 185° F) with an aqueous solution of caustic soda (sodium hydroxide) or soda ash (sodium carbonate). The refining may be done in a tank (in which case it is called batch or tank refining) or in a continuous system. In batch refining, the aqueous emulsion of soaps formed from free fatty acids, along with other impurities (soapstock), settles to the bottom and is drawn off. In the continuous system the emulsion is separated with centrifuges. After the fat has been refined, it is usually washed with water to remove traces of alkali and soapstock. Oils that have been refined with soda ash or ammonia generally require a light re-refining with caustic soda to improve colour. After water washing, the oil may be dried by heating in a vacuum or by filtering through a dry filter-aid material. The refined oil may be used for industrial purposes or may be processed further to edible oils. Usually, the refined oils are neutral (*i.e.*, neither acidic nor alkaline), free of material that separates on heating (break material), lower in colour, less viscous, and more susceptible to rancidity.

Batch and continuous refining

Other refining techniques that have been employed include the use of sulfuric acid instead of alkaline agents. This technique removes or destroys many of the impurities in the oils without removing the free fatty acids. For industrial oils, such as linseed oil, that are not required to be low in fatty acids, acid refining is sometimes advantageous. Steam refining, consisting of blowing clean steam through oil at high temperature and under vacuum, may sometimes be used on oils (*e.g.*, coconut oil) that contain few phospholipids or other impurities; or it may be used on oils that have been treated with acetic anhydride to remove the phospholipids and other impurities. Steam blowing removes most of the free fatty acids. Other refining agents that have been used include ammonia, magnesium and calcium oxides, ion-exchange resins, and certain organic bases.

Water refining, usually called degumming, consists of treating the natural oil with a small amount of water, followed by centrifugal separation. The process is applied to many oils that contain phospholipids in significant amounts. Since the separated phospholipids are rather waxy or gummy solids, the term degumming was quite naturally applied to the separation. The separated phospholipid emulsion layer from oils such as corn (maize) and soybean oils may be dried (commercially, these products are called lecithin) and used as emulsifiers in such products as margarine, chocolate products, and emulsion paints. The degumming of crude soybean oil, which has an average phospholipid content of 1.8 percent, provides the primary source of commercial lecithin. To obtain products of lighter colour, hydrogen peroxide may be added as a bleaching agent during the drying of lecithin. The degummed oil may be used directly in industrial applications, such as in paints or alkyd resins, or refined with alkalies for ultimate edible consumption.

Bleaching. If further colour removal is desired, the fat may be treated with various bleaching agents. Heated oils are treated with fuller's earth (a natural earthy material that will decolorize oils), activated carbon, or activated clays. Many impurities, including chlorophyll and carotenoid pigments, are adsorbed onto such agents and removed by filtration. Bleaching often reduces the resistance of oils to rancidity, because some natural antioxidants are removed together with impurities. When many oils are heated to more than 175° C (347° F), a phenomenon known as heat bleaching takes place. Apparently the heat decomposes some pigments, such as the carotenoids, and converts them to colourless materials.

Destearinating or winterizing. It is often desirable to remove the traces of waxes (*e.g.*, cuticle wax from seed coats) and the higher melting glycerides from fats. Waxes can generally be removed by rapid chilling and filtering. Separation of high melting glycerides, or stearine, usually requires very slow cooling in order to form crystals that are large enough to be removed by filtration or centrifuging. Thus linseed oil may be winterized to remove traces of waxes that otherwise interfere with its use in paints and varnishes. Stearine may be removed from fish oils in order to separate the solid glycerides that would detract from its use in paints and alkyd resins. At the same time, fish stearine is more suitable than whole oil for edible purposes. Cottonseed and peanut oils may be destearinated to produce salad oils that remain liquid at low temperatures. Tallows and other animal fats may be destearinated for simultaneous production of hard fats (high in stearic acid content for special uses such as in making candles) and of liquid oil called oleo oil.

Removing traces of wax

For many industrial purposes, fatty acids are required rather than glycerides. Saponification of tallow yields the mixed fatty acids. If these are crystallized and then pressed, a crude grade of oleic acid, termed red oil, is removed, leaving a solid residue termed single-pressed stearic acid. Repetition of melting–crystallization cycles

yields additional red oil and higher quality saturated acids termed double-pressed and triple-pressed stearic acids. Red oil is impure (approximately 75 percent) oleic acid, and despite the name "stearic acid," the solid product contains a ratio of approximately 60 parts of palmitic acid to 40 parts of stearic acid.

Hydrogenation. For many edible purposes and for some commercial applications it is desirable to produce solid fats. Many shortenings and margarines contain hydrogenated (hardened) oils as their major ingredients. The development of margarine and shortening products resulted from the invention of a successful method for converting low-melting unsaturated fatty acids and glycerides to higher melting saturated products. The process consists of the addition of hydrogen in the presence of a catalyst to the double (unsaturated) bonds. Thus oleic or linoleic acid (or their acid radicals in glycerides), which are normally liquid at room temperature, can be converted to stearic acid or the acid radical by the addition of hydrogen.

Limited use was made of this hydrogenation technology in Europe; the greatest potential use for the process lay in the United States, where a vast production of cottonseed oil, a by-product of the Southern cotton industry, awaited developments that would permit its conversion to a plastic fat. The hardening of cottonseed oil in the early 1900s gave birth to the shortening industry. Practical hydrogenation then spread to all countries where margarines and shortenings are produced from liquid oils.

In commercial practice, hydrogenation is usually carried out with vigorous agitation or hydrogen dispersion with a narrow range of catalyst concentration (about 0.05 to 0.10 percent of finely divided nickel suspended on kieselguhr, or diatomaceous earth) in a steel pressure-reaction vessel. The ordinary ranges of temperature and pressure are from 100° to 200° C (212° to 392° F) and from atmospheric pressure to 60 pounds per square inch, respectively. These conditions can be controlled to make the hydrogenation reaction somewhat selective; *i.e.*, to add hydrogen to the linolenic (three double bonds) and linoleic (two double bonds) acid radicals before adding to the oleic (one double bond) acid radicals. The most unsaturated fatty-acid groups are most easily hydrogenated and thus react first with the hydrogen if conditions are right. Copper-containing catalysts are especially selective in the hydrogenation of vegetable oils. If very hard fats with low amounts of unsaturation are desired and selectivity is unimportant, higher temperatures and pressures are employed to shorten the reaction time and to use partially spent catalyst that would otherwise be wasted. After hydrogenation, the hot oil is filtered to remove the metallic catalyst for either reuse or recovery.

Rearrangement of molecular structure

During the catalytic treatment another reaction also takes place—isomerization (rearrangement of the molecular structure) of unsaturated fatty-acid radicals to form isooleic, isolinoleic, and similar groups. Because these isomers have higher melting points than do the natural acids, they contribute to the hardening effect. The unsaturation of natural oils has the *cis* configuration, in which hydrogen atoms lie on one side of a plane cutting through the double bond and alkyl groups lie on the other side. During hydrogenation some of the unsaturation is converted to the *trans* configuration, with like groups on opposite sides of the plane. The *trans* isomers are much higher melting than the natural *cis* form. Simultaneously with the change of some of the unsaturation to the *trans* configuration there is a migration of double bonds along the chain. Thus isomers of oleic acid may be formed with the double bond in any position from carbon atom 2 to carbon atom 17. Many of these isomerized acids are higher melting than the natural oleic acid. Infrared analysis is useful for quantitative measurement of changes occurring during hydrogenation.

Partial hydrogenation of soybean oil selectively converts the highly unsaturated linolenic acid to linoleic and oleic acids, which are more stable to oxidation and better in flavour stability. Simultaneously, isomerization of double bonds coupled with limited increases in the saturated-acid content increases the melting point, so that the product becomes semisolid. Liquid oils possessing much better flavour stability than the original oils can be separated by winterizing such partially hydrogenated products. During the late 1960s hydrogenated–winterized soybean oil became an important commercial product.

Deodorization. Odourless and tasteless fats first came into high demand as ingredients for the manufacture of margarine, a product designed to duplicate the flavour and texture of butter. Most fats, even after refining, have characteristic flavours and odours, and vegetable fats especially have a relatively strong taste that is foreign to that of butter. The deodorization process consists of blowing steam through heated fat held under a high vacuum. Small quantities of volatile components, responsible for tastes and odours, distill, leaving a neutral, virtually odourless fat that is suitable for the manufacture of bland shortening or delicately flavoured margarine. Originally, deodorization was a batch process, but increasingly, continuous systems are being used in which hot fat flows through an evacuated column countercurrent to the upward passage of steam. In Europe, a deodorization temperature of 175°–205° C (347°–401° F) is common, but in the U.S., higher temperatures of 235°–250° C (455°–483° F) are usually employed. About 0.01 percent of citric acid is commonly added to deodorized oils to inactivate trace-metal contaminants such as soluble iron or copper compounds that otherwise would promote oxidation and the development of rancidity.

Characteristics of cooking oils

Olive oil is invariably marketed in undeodorized form. The natural flavour is an important asset, and olive oil, as is true of butter, commands a premium in the market because of its distinctive and prized flavour. The common cooking oils of Asia—soybean, rapeseed, peanut, sesame, and coconut—are consumed in their crude form as expressed from oilseeds. In contrast, deodorized oils are in particular demand in the United States and Europe. For many years the only important vegetable oil consumed in the United States was cottonseed oil, which in its crude form has such a strong and unpleasant flavour that further processing was an absolute necessity in order to render it edible. Because of widespread sale of neutral-flavoured cottonseed oil products over many years, a general preference was developed for odourless and tasteless fats.

Another reason for the practice of deodorizing edible oils in Europe and America relates to differences in oil quality by occidental and oriental extraction techniques. In China and Southeast Asia, edible oils have been produced principally by small, relatively crude equipment. The yield of oil is relatively low, and a minimum amount of nonglyceride substances is expressed from the seed, with the result that the flavour of the oil is fairly mild. In Europe and the United States, oil extraction is carried out in large factories that operate on an extremely competitive basis. Very high pressure expression or solvent extraction is used, and in order to improve yields the seeds are heat-treated prior to extraction. Oils obtained in high yield under such conditions are stronger in flavour than oils prepared by low-pressure expression, and the refining and deodorizing steps are required to improve palatability. The improvement in yields more than compensates for the added costs of refining and deodorizing.

When fats are hydrogenated for manufacture of margarine and shortening, they develop a characteristic sweet, but rather unpleasant, "hydrogenation odour" that must be removed from edible fats by deodorization.

EDIBLE FAT AND OIL PRODUCTS

The oil and fat products used for edible purposes can be divided into two distinct classes: liquid oils, such as olive oil, peanut oil, soybean oil, or sunflower oil; and plastic fats, such as lard, shortening, butter, and margarine. The physical nature of the fatty material is unimportant for some uses, but the consistency is a matter of consequence for other products. As a dressing on green salads, for example, a liquid oil is used to provide a coating on the ingredients; a plastic fat such as lard or butter would be unsuitable. Spreads for bread, foods that require a highly developed dough structure, or icings and fillings

Table 4: Fat Content of Typical Foods

food	fat (%)	food	fat (%)	food	fat (%)
Fruits	0.1–20	Cereals	0.2–7	Fish	0.2–30
Banana	0.2	Rice		Haddock	0.2
Pineapple	0.2	Polished	0.4	Tuna	4
Orange	0.2	Brown	2.0	Salmon	10–15
Apple	0.6	Barley	1.0	Cod	18
Avocado	17	Rye		Meat	5–60
Olive	20	Flour	1.0	Beef	10–40
Vegetables	0.1–2.0	Whole	1.7	Lamb	10–40
Potato	0.1	Wheat		Pork	15–60
Beet	0.1	Flour	1.0	Poultry	2–35
Asparagus	0.2	Whole	2.0	Nuts	40–70
Cabbage	0.2	Corn		Peanuts	45
Carrot	0.2	Meal	1.0	Almonds	55
Tomato	0.2	Whole	4.0	Pecans	70
Onion	0.3	Oats, meal	7.5	Miscellaneous	
Peas		Dairy, eggs	3–85	Cocoa	12–25
Green	0.4	Milk, whole	3.5	Chocolate	50
Dry	1.3	Cheese, cottage	4	Mayonnaise	75–80
Beans		Cheeses (cheddar, Parmesan, Swiss)	25–35	Dressings	25–60
Green	0.3	Eggs		Margarine	80
Lima	0.5	White	0.6		
Navy	1.8	Yolk	33		
		Whole	10		
		Butter	80		

with a plastic structure, require plastic fats rather than liquid oils.

Sources of edible products

For reasons related both to history and climate there are pronounced geographical patterns of consumption of fats and oils. The ancestors of the present inhabitants of central and northern Europe obtained their edible fats almost exclusively from domestic animals. The food habits and the cuisine depended on the availability of plastic fats; and butter, lard, margarine, and shortening continue to be their primary fatty-food materials. In contrast, population pressures in the older civilizations of the Orient and the Mediterranean countries of southern Europe, northern Africa, and the Near East have long since made extensive raising of livestock impractical, necessitating that the edible oils of these regions be derived primarily from intensively cultivated vegetable crops. In the tropics, conditions are relatively unfavourable for livestock but are well suited to culture of a variety of oil-bearing plants, many of which flourish in the wild state. In contrast to most high-population-density tropical areas, cattle abound in India.

Dehydrated butter or ghee is an important item of Indian cookery, and a hydrogenated shortening called vanaspati is designed to reproduce the coarsely crystalline plastic texture of ghee.

More than 90 percent of the world production of fats and oils is used in edible products, and the objective of most processing steps is to convert crude fats of low palatability or undesirable physical form into refined products that meet the regional requirements for food fats. In the highly industrialized areas of Europe and the United States, the annual consumption of visible fats—such as lard, butter, shortening, or salad oils that have been separated from the original animal or plant source—varies from 40 to 55 pounds per person in various European countries to 52 pounds per person in the United States. For the world as a whole, the average available supply is 23 pounds per person; and in many areas of South America, Africa, and Southeast Asia, the annual consumption is ten pounds or less per person.

About 40 percent of the dietary fat in the developed countries comes from isolated fats and oils, with 60 percent obtained from basic foods, whereas in the less developed countries most of the dietary fat is obtained from fruits, cereals, vegetables, dairy products, and meats, and relatively little is consumed in the form of isolated fat products. The quantities of fats and oils in conventional food supplies vary over wide ranges. Most fruits and vegetables have from 0.1 to 2.0 percent of fat, with the exception of avocados and olives, which are exceptional in their high fat content (Table 4). Cereals range from 1 to 7 percent, and nuts may contain as much as 70 percent of fat.

INDUSTRIAL USES

Historically, soap has been the largest industrial outlet for fats and oils. Leaching of wood ashes provided the alkaline agent for the first soaps, but as the chemical industry developed, such soft potassium soaps gradually gave way to hard soaps based on sodium hydroxide. Modern soap processing is a sophisticated industry making use of carefully proportioned blends of animal and vegetable fats to achieve the proper balance of foaming ability, detergency, and desirable physical form. After World War II, synthetic detergents manufactured from petroleum products supplanted much of the soap, especially in the United States.

Some hard fats, tallows, and beeswax are still used in candle manufacture, but petroleum waxes also took over a large share of candle production. Considerable quantities of specialty oils and sulfonated oils are used in leather dressing and textile manufacture.

Some oils have properties of medicinal value. Castor oil, for example, has a strong purgative action; fish-liver oils are sources of vitamins A and D; and others such as lard, olive oil, and almond oil serve as vehicles in pharmaceutical preparations. Chaulmoogra oil, which contains unique fatty acids with a cyclic (cyclopentenyl) structure, has been used in treatment of Hansen's disease (leprosy).

Linseed, tung, and other drying oils and large quantities of soybean oil, sunflower oil, and safflower oil are used in paints, varnishes, and alkyd resins. The latex-emulsion paints containing products derived principally from petroleum oil began to find widespread acceptance for interior finishes in the United States in the late 1940s. This trend, which spread to other countries in the 1950s and included emulsion paints for outside use, displaced large volumes of the drying oils in protective coatings. Competition of products from petroleum encouraged (and forced) the development of new products and new markets for glyceride oils. Among the new derivatives were drying-oil products that could be thinned with water. A new industry based on the manufacture of chemicals from fats was started just before World War II and grew rapidly after the war. Soapstocks from the refining of fats and oils were acidified to split out a dark-coloured crude grade of fatty acids, termed acidulated soapstock. Distillation of such soapstock acids provided large quantities of distilled fatty acids with a light colour. Supplementing these acids, both animal and vegetable fats, with and without hydrogenation, were hydrolyzed with water in high-pressure autoclaves. The separated acids were fractionated by vacuum distillation, by solvent segregation, or by both. This made available commercial quantities of relatively pure, single fatty acids. From the fatty acids dozens of products, such as long-chain alcohols,

Development of new glyceride oil products

amines, amides, esters, nitriles, and ketones, were made. These chemicals could be used for many applications directly or as chemical intermediates in the manufacture of other products, such as detergents, plasticizers, special lubricating oils, polyamide resins, and special thixotropic (gel) paints. Modifications of the glycerides themselves, through such chemical processes as epoxidation, copolymerization, rearrangement, chlorination, vinylation, and acetylation, led to more new industrial products.

WORLD PRODUCTION

The principal countries and regions producing various oil-bearing materials (seeds, nuts, and the like) are listed in Table 3. In many cases the fats are not extracted in the countries of origin; instead, the raw materials are exported. Thus Marseille and Rotterdam became centres for copra processing, and Liverpool, Hull, and Hamburg-Harburg developed into centres of oilseed expression. The development of industry in tropical countries and the high cost of transportation, however, have led to the increased production of oils and fats in the countries of origin.

During the two decades from 1950 to 1970 there was a marked increase in the world production of fats and oils (Table 5). Population increased by 44 percent, from 2,517,000,000 in 1950 to an estimated 3,630,000,000 in 1970, but during the same 20 years production of fats and oils increased by 73 percent. Production of edible vegetable oils more than doubled, led by huge increases in soybean, sunflower seed, and rapeseed oils. Palm oils showed a modest increase, but large plantings in the 1960s suggest an increase in production in the early 1970s. Production of industrial oils was static and animal fats increased about proportionally to the increase in human population. During the 1960s a large-scale fish-processing industry developed in Peru, based on the abundant anchoveta (a small anchovy) of the highly nutrient coastal waters. Whale-oil production is decreasing because of overfishing. In 1970, the U.S. Department of the Interior placed the sperm whale on the endangered-species list and prohibited sperm-oil imports.

Table 5: Estimated World Production of Fats and Oils
(000,000 short tons)

	1950	1960	1970*
Edible vegetable oils			
Cottonseed	1,565	2,450	2,655
Peanut	1,835	2,770	3,435
Soybean	1,640	3,685	6,710
Sunflower	915	1,835	4,210
Rapeseed	990	1,215	2,090
Sesame	690	595	635
Olive	1,267	1,300	1,380
Total	8,902	13,850	21,115
Palm oils			
Coconut	1,975	2,160	2,210
Palm kernel	455	480	490
Palm	1,240	1,330	1,885
Babassu kernel	41	64	90
Total	3,711	4,034	4,675
Industrial oils			
Linseed	1,210	1,055	1,230
Castor	230	305	350
Oiticica	14	22	14
Tung	115	134	129
Total	1,569	1,516	1,723
Animal fats			
Butter	3,520	4,250	5,050
Lard	3,680	4,000	4,310
Tallow and grease	2,350	3,440	4,700
Total	9,550	11,690	14,060
Marine oils			
Whale	425	418	88
Sperm	55	122	145
Fish	375	509	1,147
Total	855	1,049	1,380
World total	24,587	32,139	42,953

*Estimated.
Source: U.S. Dept. of Agriculture, *World Agricultural Production and Trade.*

II. Waxes

Waxes differ from fats chemically in that they are fatty-acid esters of monohydroxy alcohols instead of glycerol. A few are esters of dihydroxy alcohols. Moreover, their physiological function as a protective coating on cuticles of leaves and fruit appears to be different from that for fats, because waxes rarely occur as cell constituents. The waxes are difficult to saponify, in contrast with the relative ease of saponification or hydrolysis of glycerides. Many of the waxes melt at high temperatures (*i.e.*, between about 35° and 100° C, or 95° and 212° F) and form hard films that can be polished to a high gloss; therefore, they are used in many kinds of polishes. Their other physical properties are similar to those of the fats. They are soluble in the same solvents and leave grease spots on paper. Latex emulsions consisting of polymers of vinyl and acrylic compounds have partially replaced polishes based on natural waxes, particularly for use on floors.

Constituents of waxes

The fatty acids found in waxes are almost always saturated. They vary from lauric to octatriacontanoic acid ($C_{37}H_{75}COOH$). Saturated alcohols from C_{12} to C_{36} have been identified in various waxes. Several dihydric (two hydroxyl groups) alcohols have been separated, but they do not form a large proportion of any wax. Also, several unidentified branched-chain fatty acids and alcohols have been found in minor quantities. Several cyclic sterols (*e.g.*, cholesterol and analogues) make up major portions of wool wax.

Only a few vegetable waxes are produced in commercial quantities. Carnauba wax, which is very hard and is used in some high-gloss polishes, is probably the most important of these. It is obtained from the surface of the fronds of a species of palm tree native to Brazil. A similar wax, candelilla wax, is obtained commercially from the surface of the candelilla plant, which grows wild in Texas and Mexico. Sugarcane wax, which occurs on the surface of sugarcane leaves and stalks, is obtainable from the sludges of cane-juice processing. Its properties and uses are similar to those of carnauba wax, but it is normally dark in colour and contains more impurities. Commercial production of other waxes such as rice wax and grain-sorghum wax proved unsuccessful. Other cuticle waxes occur in trace quantities in such vegetable oils as linseed, soybean, corn (maize), and sesame. They are undesirable because they may precipitate when the oil stands at room temperature, but they can be removed by cooling and filtering. Cuticle wax accounts for the beautiful gloss of polished apples.

Beeswax, the most widely distributed and important animal wax, is softer than the waxes mentioned and finds little use in gloss polishes. It is used, however, for its gliding and lubricating properties as well as in waterproofing formulations. Wool wax, the main constituent of the fat that covers the wool of sheep, is obtained as a by-product in scouring raw wool. When it is purified it is called lanolin and is used as a pharmaceutical or cosmetic base because it is easily assimilated by the human skin. Sperm oil and spermaceti, both obtained from sperm whales, are liquid at ordinary temperatures and are used mainly as lubricants.

III. Essential oils

The volatile oils of the plant kingdom are called essential oils because they were considered to represent the very essence of odour and flavour. As a practical definition, an essential oil is a more or less volatile material isolated by a physical process from an odoriferous plant of a single botanical species. The oil bears the name of the plant from which it is derived; for example, rose oil or peppermint oil.

Distillation is the most common method for isolation of essential oils, but other processes—including enfleurage (extraction by using fat), maceration, solvent extraction, and mechanical pressing—are used for certain products. Younger plants produce more oil than older ones, but old plants are richer in more resinous and darker oils because of the continuing evaporation of the lighter fractions of the oil.

Function of essential oils

The function of the essential oil in a plant is not well understood. Odours of flowers probably aid in natural selection by acting as attractants for certain insects. Leaf oils, wood oils, and root oils may serve to protect against plant parasites or depredations by animals. Oleoresinous exudations that appear when the trunk of a tree is injured prevent loss of sap and act as a protective seal against parasites and disease organisms. Few essential oils are involved in plant metabolism, and some investigators maintain that many of these materials are simply waste products of plant biosynthesis.

Out of the vast number of plant species, essential oils have been well characterized and identified from only a few thousand plants, and of these only 150 to 200 have been processed commercially for their fragrant products. The oils are stored as microdroplets in glands of plants. After diffusing through the walls of the glands, the droplets spread over the surface of the plant before evaporating and filling the air with perfume. Characteristically, the most odoriferous plants are found in the tropics, where the solar energy is greatest. Many of the essential oils are processed in remote tropical areas under exceptionally primitive conditions.

HISTORY

The history of essential oils is closely intertwined with that of flavours and spices. The first records of essential oils come from ancient India, Persia, and Egypt; and both Greece and Rome conducted extensive trade in odoriferous oils and ointments with the countries of the Orient. Most probably these products were extracts prepared by placing flowers, roots, and leaves in fatty oils. In most ancient cultures, odorous plants or their resinous products were used directly. Only with the coming of the golden age of Arab culture was a technique developed for the distillation of essential oils. The Arabs were the first to distill ethyl alcohol from fermented sugar, thus providing a new solvent for the extraction of essential oils in place of the fatty oils that had probably been used for several millennia.

The knowledge of distillation spread to Europe during the Middle Ages, and isolation of essential oils by distillation was described during the 11th to 13th centuries. These distilled products became a specialty of the medieval pharmacies, and by about the year 1500 the following products had been introduced: oils of cedarwood, calamus, costus, rose, rosemary, spike, incense, turpentine, sage, cinnamon, benzoin, and myrrh.

Distillation was considered as serving to separate the essential from the crude and nonessential with the help of fire, and it met almost ideally the definition of a "chymical" process as expounded by the noted Swiss physician and alchemist Paracelsus (Theophrastus Bombastus von Hohenheim, 1493–1541). Paracelsus maintained that the objective of alchemy was the development of medicines, rather than transmutation of base materials to gold. It

The *quinta essentia*

was his theory that there is a most sublime extractive, the *quinta essentia* (quintessence), which represents the effective part of every drug, and that isolation of this extractive should be the goal of pharmacy. His theories stimulated physicians and pharmacists to seek essential oils by distillation of a range of aromatic leaves, woods, and roots; and the very name essential oil recalls the Paracelsian concept of the *quinta essentia.*

Starting from the time of Marco Polo, the much-prized spices of India, China, and the Indies served as the impetus for trade with the Orient. Quite naturally, such spices as cardamom, sage, cinnamon, and nutmeg were subjected to the pharmacists' stills. By the middle of the 18th century about 100 essential oils had been introduced, although there was little understanding about the nature of the products. As chemical knowledge expanded in the late 1800s and early 1900s, many well-known chemists took part in the chemical characterization of essential oils. Improvement in knowledge of essential oils led to a sharp expansion in production, and use of the volatile oils in medicine became quite subordinate to uses in foodstuffs, beverages, and perfumes.

In the United States, oils of turpentine and peppermint were produced before 1800; within the next several decades oils of four indigenous American plants became important commercially; namely, sassafras, wormwood, wintergreen, and sweet birch.

Since 1800 many essential oils have been prepared and characterized, but only a few have attained commercial significance.

METHODS OF PRODUCTION

The first step in the isolation of essential oils is crushing or grinding the plant material to reduce the particle size and to rupture some of the cell walls of oil-bearing glands. Steam distillation is by far the most common and important method of production, and extraction with cold fat (enfleurage) or hot fat (maceration) is chiefly of historical importance (see below).

Steam distillation. Three different methods of steam distillation are practiced. In the oldest and simplest method a vessel containing water and the chopped or crushed plant material is heated by a direct flame, and the water vapour and volatile oil are recovered by a water-cooled condenser. This original method is being replaced by a process in which the plant material is suspended on a grid above the water level, and steam from a second vessel is introduced under the grid. The volatiles are condensed and the oil is separated. In the third process, the vessel containing the plant material on a grid is heated to prevent condensation of steam, so that dry distillation is attained.

The first and, to a lesser extent, the second methods are subject to significant oil losses because of water solubility of some components. To avoid such losses, the distillate remaining after the essential oil is skimmed off may be redistilled or, alternatively, the water distillate may be added back to the next charge in the distillation vessel. Yields in a single distillation vary, depending on the plant material. The yield from most flowers is less than 0.05 percent; many roots such as angelica or grasses such as citronella have yields in the 0.1 to 1.0 percent range; yields from seeds may be as high as 16 percent (nutmeg); and oleoresins may produce from 10 percent (olibanum, the Biblical frankincense) to 75 percent (gurjun balsam).

In many tropical areas, aromatic plants grow wild or are cultivated as a garden crop. Distillation of the oil is usually a family industry and most often a side occupation. The harvesters produce small quantities of an oil that they then sell to brokers or village buyers who consolidate these products and forward them to exporters. For such processing the simple water stills are used because they are low in cost, easy to operate, and portable, so that they can be moved readily to follow the plant material. Oil of star anise in Vietnam, oil of lemongrass in India, or oil of cananga in Java are produced in this way. In contrast, the highly organized essential-oil industry of southern France uses more complex and permanent facilities. With better transportation the plant materials can be gathered from wider areas, and the lack of portability is more than compensated for by the higher yields obtained from scientifically monitored equipment.

Enfleurage. In the Grasse region of southern France, flowers were extracted with cold fat (enfleurage) long before the introduction of extraction with volatile solvents. This process is applied to flowers that do not yield an appreciable quantity of oil by steam distillation or whose odour is changed by contact with boiling water and steam. A highly purified mixture of tallow and lard is spread on both sides of the chassis, a rectangular wood

Use of the chassis

frame (about 5 × 50 × 40 centimetres, or 2 × 20 × 16 inches) containing a glass plate. When piled one above the other the chassis form airtight compartments with a layer of fat on the upper and lower surface of each glass plate. Each morning, flowers are carefully spread over the fat surface, and the chassis are assembled in a cool, dark room and left for a period varying from 24 hours (for jasmine) to 72 hours (for tuberose). During this time most of the flower oil is absorbed by the fat. The petals are then removed (defleurage), and the process is repeated until the fat is saturated with oil. The final product is called pomade (*e.g.*, pomade de jasmine, pomade

Table 6: Some Important Essential Oils

essential oil	source	production (tons; 1970, est.)*	price ($/lb)†
Abies sibirica (Siberian fir)	U.S.S.R.		4
Almond, bitter	U.S., Morocco, Spain, France, Algeria		3
Angelica root	Belgium, France, The Netherlands, Germany	1	140
Anise and star anise	U.S.S.R., eastern Europe, China, Vietnam	200	3.50
Bay	West Indies	50	8
Bergamot	Sicily and the Italian Peninsula, France, Switzerland	200	11
Bois de rose	Brazil, Peru	500	2.50
Camphor	China, Japan, Taiwan	5,000	0.30–1.15
Cananga	Java		3–6
Caraway	The Netherlands, U.S.S.R.		4–5
Cardamom	Sri Lanka, India, Central America		60–70
Cedarleaf (thuja)	northeastern U.S.	30	8
Cedarwood	U.S., Africa, India	500	1–1.50
Celery seed	India, France, U.S.	50	35–70
Chenopodium (American wormseed)	U.S.	3	5
Cinnamon (and cassia)	China, Burma, Sri Lanka	250	6
Bark oil	China, Burma, Sri Lanka		64
Leaf oil	China, Burma, Sri Lanka		2
Citronella	Sri Lanka, Java, China, Guatemala	3,000	1
Clove			
Bud	Zanzibar, Madagascar	400	14
Leaf	Zanzibar, Madagascar	1,000	2
Copaiba	South America	50	2
Coriander	eastern Europe, U.S.S.R.	25	9
Dillweed	central Europe, U.S.	150	5
Eucalyptus	Australia, Spain, South America		0.75
Fennel	central Europe, U.S.S.R., Italy, France		2.50
Geranium	Réunion, Morocco, U.S.S.R., Madagascar	200	16
Ginger	Jamaica, India, West Africa		45
Grapefruit	U.S.	300	1
Jasmine	France, Italy, Morocco	10	2,000
Juniper berry	central Europe		3
Lavandin	France	700	3
Lavender	France, U.S.S.R., U.K.	50	5–8
Spike	Spain, France	100	3–4
Lemon	U.S., Italy, Brazil, Israel, Greece	1,500	6–9
Lemongrass	Southeast Asia, Central America, West Africa	2,000	3
Lime	Mexico, Haiti, West Indies	400	11
Mandarin	Italy, Spain, Cyprus	80	8
Florida tangerine	Florida		7
Neroli	France, Italy, Spain, Lebanon, Egypt	1	300–425
Nutmeg	Indonesia, West Indies	75	6
Ocotea cymbarum	Brazil, Colombia, Paraguay	1,000	1
Orange	U.S., Italy, West Indies	1,500	0.75
Bitter	West Indies, Italy, Brazil, Spain		6
Origanum	Spain, Morocco		4.50
Palmarosa	India, Java	30	14
Patchouli	Indonesia, Malaysia, Philippines	150	5
Pennyroyal	Spain, Morocco	50	3
Peppermint	U.S., U.K., India, U.S.S.R., Japan	4,000	5
Pine	U.S.	100,000	0.30
Rose			
Bulgarian		5	1,150
Turkish			700
French			145
Moroccan			190
Rosemary	Spain, Italy, eastern Europe, U.S.S.R.	200	1.80
Sage			
Clary	U.S.S.R., France, Hungary	5	42
Dalmatian	Yugoslavia	25	4
Sandalwood	India	125	16
Spearmint	U.S., Germany, U.K., China	1,500	5
Spruce	U.S., Canada	2	3.50
Turpentine	U.S., Sweden, Greece, U.S.S.R.	300,000	0.17
Vetiver	Réunion, Java, Brazil, Haiti	100	18
Wintergreen	U.S.	1,000	0.60
Wormwood	France, U.S.	20	5
Ylang-Ylang	Madagascar, Réunion, Philippines	25	8

*In part from C. Donnarumma, International Flavors and Fragrances, Inc. †*Oil, Paint and Drug Reporter*, January 1971.

de violet), the most saturated being Pomade 36, because the chassis has been treated with fresh flowers 36 times during the process.

Originally the pomades were used directly; later they were extracted with alcohol. The alcoholic washings are called Extrait 36 when made from Pomade 36, and they reproduce to a remarkable degree the natural flower perfumes. If the alcohol is evaporated, the oil residue, or absolute, is called *absolu de pommade* or *absolu d'enfleurage*.

The process is very expensive, and it is now limited to a few flowers, such as jasmine or tuberose, that continue their physiological activity after picking and thus give higher yields by the cold processing of enfleurage than they would in hot extractions, which stop perfume production.

Maceration. The physiological activity of most flowers stops soon after picking. In this case, it is possible to shorten the long enfleurage process by extracting the flowers by molten fat for one to two hours at a temperature ranging from 45° to 80° C (113° to 176° F). The fat is filtered after each immersion, and after ten to 20 extraction cycles the pomade is sold as such, or it may be extracted with alcohol to yield the extract or the absolute.

Extraction. Since both enfleurage and maceration are rather expensive processes, the essential-oil specialists of Grasse shifted almost completely to extraction for the recovery of essential oils from plant materials that could not be processed by steam distillation. Petroleum naphthas, benzene, and alcohol are the primary solvents. The solvents dissolve not only essential oils but also waxes, fats, pigments, and other cell components. Re-

moval of solvent leaves a semisolid concrete. The concrete is used directly for some applications, but more commonly it is extracted with alcohol, followed by chilling to precipitate waxes, and filtration. Finally, the wax-free alcohol solution is desolventized under reduced pressure to yield the essential oil, the so-called absolute.

Expression. This procedure is applied only to citrus oils. The outer coloured peel is squeezed in presses, and the oil is decanted or centrifuged to separate water and cell debris. The method is used for oil of sweet and bitter orange, lemon, lime, mandarin, tangerine, bergamot, and grapefruit. Much oil is produced as a by-product of the concentrated-citrus-juice industry.

CHEMICAL COMPOSITION

The development of chromatographic procedures, especially gas–liquid chromatography and thin-layer chromatography, as routine analytical techniques during the 1950s and 1960s facilitated the precise identification and determination of the chemical constituents of essential oils.

Constituent compounds of essential oils

Terpenes are by far the most important components, but individual oils may contain appreciable quantities of straight chain, aromatic, or heterocyclic compounds. Thus allyl sulfides are characteristics of oil of garlic, traces of indole and anthranilic acid esters are found in orange oil, straight chain alcohols and aldehydes are recognized in oil of violets, and phenols and other aromatic compounds are common to many oils.

Terpenes, the dominant constituents of essential oils, are built up from units of the simple five-carbon molecule isoprene. Both hydrocarbons and oxygenated compounds such as alcohols, aldehydes, ketones, acids, esters, oxides, lactones, acetals, and phenols are responsible for the characteristic odours and flavours.

In some oils one or only a few components predominate: thus oil of wintergreen contains about 98 percent of methyl salicylate; orange oil, about 90 percent of *d*-limonene; bois de rose, 90 percent of linalool; and cassia, up to 95 percent of cinnamaldehyde. In most oils there is a mixture of anywhere from a few dozen to several hundred individual compounds. Trace components are very important, since they give the oil a characteristic and natural odour.

Essential oils are generally expensive, with prices ranging from several dollars per kilogram on the low side to several thousand dollars per kilogram. The high price of the natural oils coupled with their limited availability has encouraged a search for substitutes. Great progress has been made in the synthesis of individual components such as geraniol, citral, linalyl acetate, and the like. These synthetics have been combined with natural oils to extend supplies, and they have also been blended together in an attempt to duplicate the oils themselves. Such reconstituted oils usually lack certain of the odour notes of the natural products, because of absence of trace ingredients, often unidentified, that may be present in the natural oils. They also tend to have a more "chemical" odour, because of trace impurities in the synthetics that are different from the components of natural oils.

PRODUCTION AND ECONOMICS

Essential oils are used in three primary ways: as odorants they are used in cosmetics, perfumes, soaps, detergents, and miscellaneous industrial products ranging from animal feeds to insecticides to paints; as flavours they are present in bakery goods, candies, confections, meat, pickles, soft drinks, and many other food products; and as pharmaceuticals they appear in dental products and a wide, but diminishing, group of medicines.

The world production of most essential oils is in the range from a few tons per year to a few thousand tons. Pine oil from steam distillation of pine stumps and turpentine from distillation of pine oleoresin are exceptional, with annual volumes of 100,000 and 300,000 tons, respectively; but these products are used primarily as solvents, and neither is used directly by the essential-oil industry, although oil of turpentine is a starting point for the manufacture of synthetic camphor, terpineol, geraniol, nerol, citronellol, linalool, citral, neral, and menthol, all of which are used in flavours or perfumery.

Annual exports by France are approximately 4,000 metric tons, with a value of $40,000,000–$50,000,000. Imports into the United States during 1969 were obtained from 74 countries and had a value of $56,000,000, almost balanced by exports of essential oils from the U.S., with a value of $51,000,000. Table 6 lists the more important essential oils together with the source and approximate world production.

BIBLIOGRAPHY

Oils and fats: T.P. HILDITCH, *The Chemical Constitution of Natural Fats,* 3rd ed. rev. (1956); AMERICAN OIL CHEMISTS' SOCIETY, *Official and Tentative Methods of the AOCS,* 3rd ed. with supplements (1971); A.W. RALSTON, *Fatty Acids and Their Derivatives* (1948); P. KARRER and E. JUCKER, *Carotinoide* (1948; Eng. trans., *Carotenoids,* 1950); F.D. GUNSTONE, *An Introduction to the Chemistry of Fats and Fatty Acids* (1958); K.A. WILLIAMS (ed.), *Oils, Fats, and Fatty Foods,* 4th ed. (1966); D. SWERN (ed.), *Bailey's Industrial Oil and Fat Products,* 3rd ed. (1964); H.J. DEUEL, *The Lipids,* 3 vol. (1951–57); E.W. ECKEY, *Vegetable Fats and Oils* (1954); M.E. STANSBY, *Fish Oils* (1967); T.J. WEISS, *Food Oils and Their Uses* (1970); J.H. VAN STUYVENBERG, *Margarine* (1969); B. LEVITT, *Oils, Detergents and Maintenance Specialities,* 2 vol. (1967); A. JART (ed.), *Fat Rancidity* (1958); W.O. LUNDBERG (ed.), *Autoxidation and Antioxidant,* 2 vol. (1961); M.K. SCHWITZER, *Continuous Processing of Fats* (1959); E.S. PATTISON (ed.), *Fatty Acids and Their Industrial Applications* (1960); H.V. PAREKH, *Solvent Extraction of Vegetable Oils* (1959); H.W. BRACE, *History of Seed Crushing in Great Britain* (1960); H.G. KIRSCHENBAUER, *Fats and Oils,* 2nd ed. (1960); V.C. MEHLENBACHER, *Analysis of Fats and Oils* (1960); H.A. BOEKENOOGEN (ed.), *Analysis and Characterization of Oils, Fats and Fat Products* (1964); L.V. COCKS and C. VAN REDE, *Laboratory Handbook for Oil and Fat Analysts* (1966); DONALD J. HANAHAN, *Lipide Chemistry* (1960); J. DEVINE and P.N. WILLIAMS (eds.), *The Chemistry and Technology of Edible Oils and Fats* (1961); K.S. MARKLEY (ed.), *Fatty Acids: Their Chemistry, Properties, Production and Uses,* 2nd rev. ed., 5 vol. (1960–68); A.J.C. ANDERSON, *Refining of Oils and Fats for Edible Purposes,* 2nd rev. ed. (1962); *Journal of the American Oil Chemists' Society* (monthly); *Fette-Seifen-Anstrichmittel* (monthly); *Lipids* (bimonthly); R.T. HOLMAN, W.O. LUNDBERG, T. MALKIN (eds.), *Progress in the Chemistry of Fats and Other Lipids* (annual); UNITED STATES DEPARTMENT OF AGRICULTURE, BUREAU OF AGRICULTURAL ECONOMICS, *Fats and Oils Situation* (bimonthly).

Waxes: H. BENNETT (ed.), *Commercial Waxes* (1956); *Industrial Waxes,* 2 vol. (1963); A.H. WARTH, *The Chemistry and Technology of Waxes,* 2nd ed. (1956); L. ROTH and J. WEINER, *Waxes, Waxing and Wax Modifiers* (1961).

Essential oils: E. GUENTHER, *The Essential Oils,* 6 vol. (1948–52); M. STOLL, "Essential Oils" in *Kirk-Othmer Encyclopedia of Chemical Technology,* 2nd rev. ed. vol. 14 (1967); E. GILDEMEISTER and F. HOFFMANN, *Die Ätherischen Öle,* 7 vol. (1968); P.Z. BEDOUKIAN, "Progress in Perfumery Materials," in *American Perfumer and Cosmetics* annual reviews (1944–70).

(A.R.B./M.W.F.)

Oil Shales

During the several stages of formation of sedimentary rocks following the deposition of sediments in a swampy or marine environment, there is much interaction of organic and inorganic substances. When the influence of the organic substances becomes pervasive, the resulting sedimentary rock is called organogenic, meaning that it is in large part of organic origin. The organogenic sediments, or bioliths as they are termed, encompass the oil shales or kerogenites. These are layered, cleavable (fissile), dark shales that are rich in hydrocarbons and fossil organic substances.

Processing and uses

Though the inflammable character of oil shale has been known for centuries, its use as industrial fuel was long discouraged by the lower costs of coals and petroleum. The industrial processing of oil shales goes back 100 years all over the world, but the processing of shale oil took on economically important dimensions only in the 1920s. The production of oil from oil shale is costly; first, because shale oil cannot be refined by conventional means, and second, because oil from shale is deficient in hydrogen and hydrogen therefore must be added during the

refining process. Shale oil also contains relatively large amounts of sulfur and nitrogen, the removal of which during the refining process is quite expensive. With the introduction of special shale-processing methods, however, the economic efficiency of both crude-oil processing and that of shale oil becomes almost identical.

Oil shales are worth processing for three purposes: fuel-gas production; the development of a chemical industry based on shale-oil refining; and the extraction of the fissionable materials uranium and thorium and other trace elements (elements that occur in oil shales in very small amounts) from shale. Uranium and thorium together with other such industrially important metals as germanium, vanadium, cobalt, nickel, copper, silver, chromium, molybdenum, and others can be found in the largest quantities in sediments derived from swamp environments and in oil shales.

It was estimated by Victor Goldschmidt in 1933 that the quantity of sedimentary rocks in the earth's crust amounts to 300,000,000 cubic kilometres. The oil-shale and asphaltite reserve is about 10^{13} tons, and their uranium concentration has been estimated as 0.01 percent. As for crude oil, the present world stock is 50×10^9 tons, with 250×10^9 tons more estimated to be in an unproved crude-oil reserve. The quantity of oil contained in oil shale is believed to amount to another 300×10^9 tons.

One part of the trace elements molybdenum, vanadium, copper, zinc, boron, manganese, and arsenic in organic sediments is of biogenic origin; another part accumulated during the diagenesis (chemical and physical changes that accompany the transformation of sediments to solid rock) of organic substances, partly due to sorption and partly by chemical precipitation. In humic rocks, those metals that migrate as cations in the aqueous solutions of the soil or that are reduced to cations by the reductive effect of humic acids become concentrated. Because of the cation sorption property of the humic acids, the concentration of metals in humic rocks can be 10^4 times as great as metal concentrations in subsoil water. A uranium concentration of this magnitude can be observed in the United States and in Sweden even today, derived from the natural waters that originate in granitic detrital areas and that seep through peat layers rich in humic acids.

This article treats the nature and distribution of oil shales, their probable genesis, and the environmental factors of greatest significance. For further information on lithology, see SEDIMENTARY ROCKS and SHALES; for discussions of environments of deposition and diagenetic changes see also articles on CLAY MINERALS; COALS; PETROLEUM; and NATURAL GAS; for information on the status of associated technologies see PETROLEUM REFINING and PETROLEUM AND GAS EXTRACTION.

ORGANIC CONTENT OF OIL SHALES

Bitumens and bituminization

A predominant part of the organic content of oil shales is formed by bituminous materials. The word bitumen is a collective term denoting the mixture of liquid and solid hydrocarbons that either occur in nature or can be extracted from natural substances without decomposition. Bitumens are colloidal systems, sols, that contain resinous or carbonaceous particles of an average molecular weight of 300 to about 3,000. Most oil-shale bitumens, also called pyrobitumens and kerogenes, are insoluble. The term kerogen was applied for the first time in the description of the shales in Scotland, when it was used to refer to the organic mineraloid of indefinite chemical composition that is the source of the oil when distilled. The term kerogenite is applied not only for kerogen-bearing sedimentary formations of shaly structure but also for other ones containing kerogen. The formation of bitumen takes place in nature in the metamorphism of fossil organic substances under definite chemical and physical circumstances; the term bituminization is used in geochemistry for the designation of the progressive development of hydrocarbons in organogenic materials.

On land or in areas periodically flooded by water the organic sediment can, either continuously or periodically, get into contact with the oxygen of the atmosphere. The end-products of this limited oxidation (*i.e.*, carbonization) are humus or coal (carbonaceous rocks). Humite coals are the products of the peaty sediments of accumulated terrestrial vegetation. Water plays an important part in material transport as well as in covering the organic sediment with layers of mineral sediment. The organic material of sapropel coals (formed from thick deposits of the organic remains of marine plants and animals) is of aquatic origin. They accumulate as underwater sediments in larger pools of stagnant water that are overgrown with coal-forming vegetation and inhabited by algal colonies. In dry seasons, when shallow waters dry up, they perish, leaving behind yellowish or brown elastic remains. Sediments of this kind can be found in the southern part of Lake Balkhash, in the U.S.S.R.; in the coastal regions of Lake Coorong in southern Australia; in Mozambique; and in the area north of Lake N'hangella.

Today carbonic sedimentation takes place in areas where the oxidation of organic substances is limited and the rate of bacterial decomposition is low. These formations are characteristic of cold and moderate climatic conditions. In the tropics, however, where temperatures are high, intensive bacterial activity and a rapid oxidation of organic substances are dominant.

Generation of crude oil

A large part of the living organisms of the oceans can assimilate hydrocarbons photosynthetically from dissolved CO_2, carbon dioxide; and on their perishing and sedimentation, these hydrocarbons form the starting material of sapropel oil and bitumen (oil algae). Oxygen is excluded from the very beginning of the metamorphism of aquatic sediments. A maximum bacterial activity is characteristic of the bituminization process under anaerobic (lacking in oxygen) conditions (enrichment in hydrocarbons) and the end-product of the process is petroleum or crude oil. The "pure" processes of carbonization and bituminization depend on the environmental conditions and also may be present in a complex form. The mixed appearance of inorganic sediment-forming components and the end-products of coalification and bituminization is representative of this fact. Figure 1 shows the main carboniferous sediment types with different concentrations of these three components and their nomenclature.

From F. Pettijohn, *Sedimentary Rocks* (1949); Harper & Row

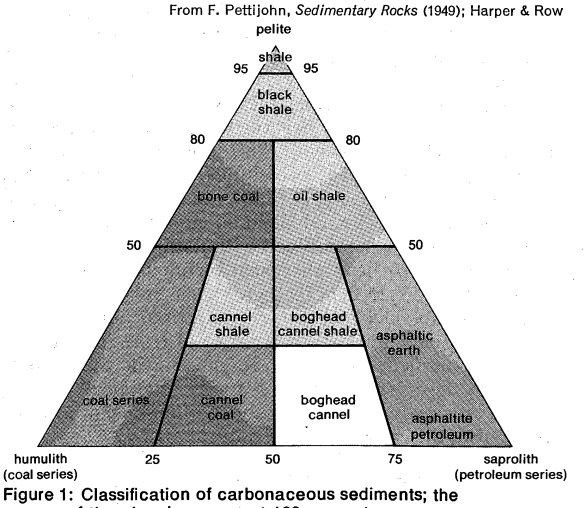


Figure 1: Classification of carbonaceous sediments; the apexes of the triangle represent 100 percent.

PROPERTIES OF OIL SHALES AND RELATED COALS

Kerogenic shales, which may be regarded as the "parent rocks" of crude oil, contain transitional bituminous substances that possess certain qualities of both petroleum and coal. They are brown to black, of low specific weight, inflammable, and burn with a sooty flame; if heated above 350–400° C, they yield oil. Their external structure is laminar and in a stratigraphic section, alternating darker and lighter strata correspond to the periodic changes of organic content. They are quite resistant to the oxidizing effect of air. Chemically, they are various mixtures of clay, shale, limestone, and organic substances. From the point of view of mineral composition, they mainly consist of silica, iron, aluminum, calcium, magnesium, and sodium carbonates, silicates, oxides, and sulfides.

Organic substances are contained in shales in the following main forms: soluble liquid hydrocarbons, soluble asphalt, and insoluble kerogen, or pyrobitumen. The organic content appears mainly in the form of mineraloid kerogen, which is insoluble. In superior shales the kerogen content can amount to 20 or 30 to 50 percent. The exact chemical composition of kerogen still has not been established. It is certain, however, that the kerogen substance of the various shales is polymerized in the same way; they are stereopolymers in which carbonyl, carboxyl, and phenolic hydroxyl groups can be found. The reason for the high insolubility can be explained by the giant-sized macromolecules of kerogen. The elemental composition of kerogen is as follows: carbon 69–80 percent; hydrogen 7–11 percent; nitrogen 1.2–2.5 percent; sulfur 1–8 percent; oxygen 9–17 percent. In a thin section, kerogen can be recognized through the presence of colloidal globules or macerated plant or algae remains.

The chemical composition of shale oil is still controversial. Kerogens probably consist of aliphatic groups, which can be regarded as ethers, formed by the combination of hydrocarbons and perhaps alcohol derived fatty acids.

Figure 1 demonstrates that with an increase of the kerogen content of oil shales, they turn first into torbanites and boghead coals, and then gradually into cannel coal. In this way, these formations can be conveniently distinguished from the members of the peat–anthracite series.

Boghead coals

Boghead coals differ from normal coal because, upon destructive distillation, they yield paraffins and olefins whereas coal yields aliphatic hydrocarbons. There are abundant torbanite occurrences in France, New South Wales, South Africa, and in the northern part of the United States. The torbanite of Scotland seems to be quite homogeneous and compact and seldom reveals interlayers. The organic mineraloid in it shows yellowish transparency and can be regarded as the remains of *Pila scotica* algal colonies. The Permo-Carboniferous kerogen shale of New South Wales is largely made up of the alga *Reinschia australis*. At Satmar, South Africa, where a large torbanite occurrence was under exploitation in the 1970s, the microscopically recognizable fossil algal remains are known as *Botriococcus brauni*. This formation is a good example of the close association of torbanite with a different kind of coal. Torbanites generally contain no massive plant remains, which indicates that they have not been formed from normal peat. Their high ash content indicates their aquatic origin. The organic content of a highly pure boghead coal occurrence in Australia known as coorongite is provided exclusively by *Elaeophyton coorongiana*, oil algae.

Sapropelic coals (algal coals) can be the component parts of coal formations of any geological age. The oil algae recognized in Tertiary bogheads created large one-celled colonies with very thick, bituminous cell walls. In coastal regions these are usually known as balkasite and coorongite. Marahunite, named for Marahu, Brazil, its place of occurrence, is also a kind of brown coal, occurring primarily in Cretaceous or Pliocene formations. A marahunite colony consists of a soft streak of several metres width. The Australian wollongongite and hartleyite, as well as the Siberian closterite, the latter having received its name from the *Closterium* algae, also belong to the populous category of the bogheads. Boghead coals occur in France, Saxony, Poland, and North America in Carboniferous rocks. In the boghead of Poland, eualginite dominates. In shallow and frequently drying waters, mushroom colonies develop and turn alginite into eualginite. It is a common peculiarity of the bogheads that they are rich in nitrogen and pyrite sulfur and that no testacean fossils occur in them. Boghead coals yield dark green shale oil, rich in paraffins, when they are heated up to 460° C in closed retorts.

Cannel coals

A characteristic type of Carboniferous coal is cannel coal, also referred to as spore-pollen coal. It burns in a long, soft, sooty flame and its name refers to the fact that it will even catch fire from the flame of a candle. To the naked eye, cannel coal resembles boghead. Antraxylon predominates among its components, and microscopic spores, algal remains, or even huminitic components may occur. In Carboniferous cannels the quantity of soluble bitumen and distillable tar content is very high, and these give off numerous gaseous products during combustion. Tertiary spore-pollen coals do not have the same tar-yielding ability. It is thought that the origin of cannel begins with the sinking of spores in shallow pools and small still lakes at neutral pH (pH = 6–8) and under aerobic conditions. Following the filling of these pools and lakes, anaerobic conditions probably prevailed. Although the spore content is characteristic of cannels, in certain boghead cannels a considerable amount of algae is present. Algae are completely missing or are insignificant in common cannels. No sharp distinction can be made between cannel and boghead coals on the basis of algal or spore content, because they are extremely similar with respect to their elemental chemical composition.

Dysodile, or paper coal, constitutes a typical transitory stage between algal coal and spore-pollen (boghead-cannel) coal. It occurs in Germany, Brazil, in the Etna region of Italy, and in Bohemia (Cypres Shale). It is grayish yellow, laminated, thin, and flexible and contains the remains of animal organisms. Its composition is like that of oil shales.

Fimmenite may be regarded as an even purer kind of spore-pollen formation, dating back to the Eocene period. In the Geisel Valley of Germany, it is known as pollen coal. Both carbonization and bituminization occur during the formation of the oil-shale–boghead-cannel coal sequence.

ORIGIN OF OIL SHALES

The diagenesis of coal formation is relatively well known because the consecutive members of the carbonization process in the peat–anthracite sequence are well illustrated in nature. As far as the diagenesis of petroleum is concerned, however, the several empirical rules available primarily refer to the final kerosene state and have little to do with the process of crude-oil formation. In the early stage of the diagenesis of organic substances, bacterial activity is most important because it eliminates most of the nitrogen and oxygen from the organic substance and thus transforms it to a chemical composition closely connected with crude oil. It has been proved that methane in recent sapropelic sediments, for example, is the result of anaerobic bacterial activity. At present, there is no proof that the formation of the higher members of the hydrocarbon series is due to anaerobic bacterial activity.

Diagenetic factors

The reactions that produce liquid hydrocarbons from organic substances contained in shale are somewhat better known. The catalytic effect of clay minerals (*q.v.*) plays an important role. Controlling this process in natural environments is difficult. It is suspected that natural radioactivity may be instrumental in oil formation. Laboratory experiments have been carried out to reconstruct the diagenesis of natural oil. Assuming that change of temperature is the main diagenetic factor, a sample of brown coal was heated slightly in order to expedite bituminization. Afterwards, the end-products of the laboratory experiment were successfully identified with the paraffins of the Gippsland crude oil. Thus it was proved that the crude oil of the Gippsland Basin is likely to have been formed *in situ* from low ranking Cretaceous or Eocene coals.

General opinion holds the crucial differences among the various kinds of organic sedimentary formations to be not so much the result of different mechanisms of sedimentation but rather of the chemically distinctive features of the organic ingredients, distinctive character of their decomposition, and the conditions of their diagenesis. The most significant parameters of diagenesis of organic substance exercise their influence during the entire period of metamorphism. In the depositional period, a rise in temperature causes the dehydration of an organic substance. The effect of dehydration on the increase of reaction speed is generally known in terms of the speed of bacterial effects and of catalytic reactions. Heat effects often will render the bitumen content of oil shales into oil even under natural conditions because of the cracking of hy-

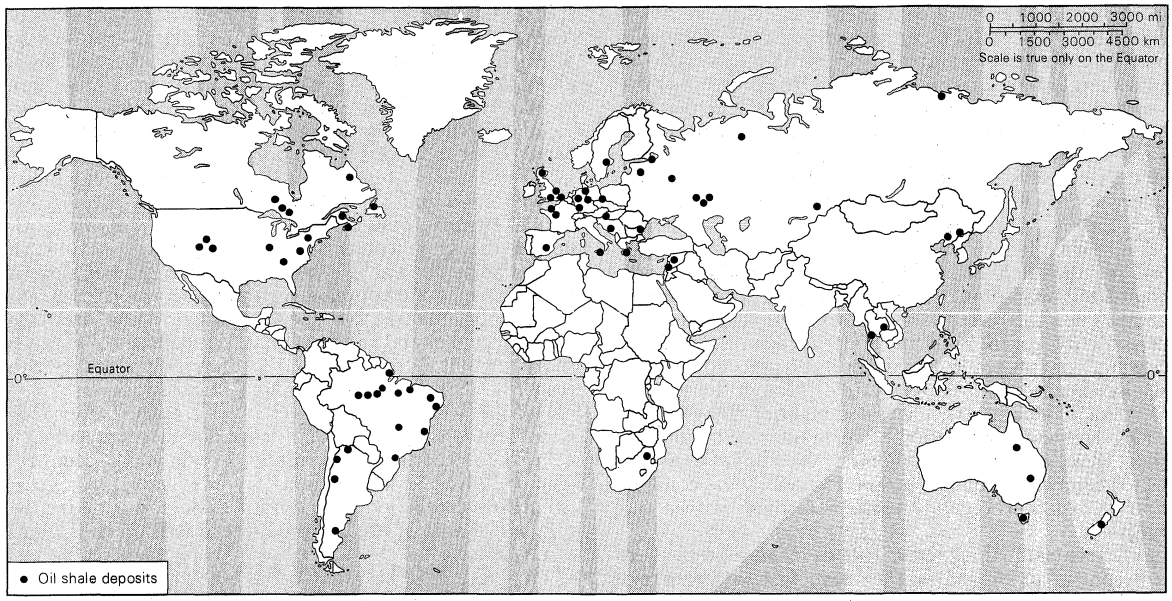


Figure 2: World distribution of oil-shale deposits. No deposits are known in Antarctica.

drocarbon chains. At temperatures above 350° C, oil shales yield gaseous products. The presence of porphyrins in kerosene and bituminites, however, makes it likely that temperatures could not have exceeded 200° C in the course of their diagenesis. In geochemical processes, the temperature and the time factor can replace each other to a certain degree on the basis of their interdependence in reaction speed.

This rivalry of the temperature and time factors is the reason ancient organic sediments can occur in the earth's crust with a lower degree of carbonization than more recent organic sediments. In the course of organic diagenesis, in the processes of bituminization and carbonization, certain age features can be recognized, but as far as the mechanism of the processes is concerned time and temperature are interdependent.

The role of pressure cannot be dismissed in the transformation of organic substances into coal and bituminous rocks. The uninterrupted, joint presence of all diagenetic factors, combined with environmental influences of organic and inorganic origin, lead to the formation of kerogenites of extremely varied nature. The exceptional feature of the Green River Shales in North America, for instance, is that they are not laminated. Because it is principally dolomitic limestone, the designation limestone kerogenite would describe it better. The structure of the rock is characterized by consecutive darker and lighter bands, which accord with the seasonal stratification of organic and inorganic substances. In periods of excessive rainfall, the dissolved carbonates formed highly calcareous layers, which straddle the oil-shale layers that were formed in the basins of Eocene Lake Uinta, Lake Goshute, and others (see TERTIARY PERIOD). The basins underwent considerable horizontal as well as vertical changes.

Environmental influence on oil-shale characteristics

The rocks of the Green River Formation are studied extensively because of their hydrocarbon content, especially ozokerite, albertite, gilsonite, and wurtzilite, which are naturally occurring solid or asphaltic pyrobitumens. Gilsonite forms thick branched veins that emanate from oil shale over long time periods. The migration of liquid gilsonite can be observed even today, where this bituminous substance diffuses into sandstone in the southern part of the Uinta Basin, Colorado. In other places, ozokerite has been found in brecciated (fractured and broken) inorganic sediments. The asphaltic sediments of the Uinta Basin show a crucial difference from the younger sediments of the basin, a difference related to different conditions of sedimentation.

The lower strata of the water of modern seas are progressively less rich in oxygen, particularly in places that are free from turbulences and currents. Life is absent in such zones, and the organic component of the sediments sinks down from the upper, better-aerated layers. The Paleozoic black shales containing graptolites (ancient marine organisms) of such origin are partly oil bearing and date back to the Devonian, Carboniferous, and Lower Jurassic periods; they are of essentially identical quality. Posidonia Shales, lithologically and from the point of view of the character of their fauna, represent equally anaerobic sedimentary conditions. Bottom-living forms are characteristically absent, and the fauna consists principally of cephalopods, together with lamellibranchs such as Posidonia, believed to be a form living attached to free-swimming cephalopods or to floating seaweed. Jet is often recognized in them as an isolated carbonic substance, proved to be a material of woody texture. These sediments, rich in floating or swimming organisms and essentially devoid of bottom dwellers, suggest that in Württemberg anaerobic conditions prevailed on the Jurassic sea bottom.

The environmental factors of recent sediment formation can be studied in the sediments of the Mexican seacoast. Both sapropelic and humitic sediments are formed, the further destiny of which is determined in nature by the wide-ranging variation of carbonization and bituminization, as these processes are determined by different environmental conditions.

OCCURRENCE AND DISTRIBUTION OF OIL SHALES

Some of the most significant oil-shale occurrences and their basic characteristics are described below. This survey is not exhaustive, as indicated by the world distribution map (Figure 2).

Lothian Shale. Important shale kerogenite beds can be found in the calcareous sandstone formations of West Lothian, in Scotland. The exceedingly thin stratification of the shale is not discernible to the naked eye.

The rock comprises finely dispersed spore and cuticle remains; yellowish or reddish ball-like forms in the shale have been verified as unicellular algae embedded in an argillaceous (clay-rich) matrix, which is impregnated with bituminous substances. This oil shale most likely accumulated in either shallow water or well-oxygenated lagoons.

Kolm Shale. Spore coal with an ash content of about 30 percent occurs in Cambrian shales. It is abundant in bituminous substances and exemplifies the cellular occurrence of bitumens. The shale layer is radioactive, the uranium content is 50–100 parts per million, and at some places is as much as 200 grams per ton.

The stratigraphic age of the lenticular shales containing the Kolm lenses is Upper Cambrian, based on the presence of the trilobite *Peltura scarabaeoides.* In Swe-

den, oil shale is the raw material base of the uranium industry. The most significant centre of processing is to be found in Göteborg. The shale comprising the Kolm lenses contains an average 10 percent of kerogen and 3 percent of the clay-limestone component.

Chattanooga Shale. A highly radioactive black shale is found in the Antrim-Chattanooga-Woodford area of the United States. Its bitumen content is usually high. In some areas intense gasification has been detected. It is thought by some workers that the formation of natural gas may be due to radioactivity (as a heating mechanism), although the irregular occurrence of the gas is not compatible with the general, evenly distributed, high radioactivity of the material.

Copper Shale. Black clay shale with high copper, sulfur, and bitumen content containing 0.5–3 percent copper, as well as silver, zinc, lead, iron, molybdenum, vanadium, and a small amount of other metals, occurs at several places in Germany, where it formed in shallow bays during the Zechstein Period. River-carried sulfates were changed into sulfides by reduction processes and then precipitated from the solution. The thickness of the layer is 25 to 30 centimetres. Its most important occurrences are at the foot of the Harz Mountains, at the fringes of the Thuringian Forest, in Hessen and Baden-Württemberg, and in Upper Bavaria and Westphalia in the Posidonian Shales. The sulfide minerals are finely dispersed in the clay that is saturated with bitumen. Of the German copper production, 90 percent is covered by the oil-shale raw material. Molybdenum also is extracted from it. Its kerogen content is relatively low (10 percent); its inorganic component consists of 30 percent clay, 25 percent limestone, and 6 percent pyrite, and the average silicate content is 28 percent.

Tasmanite. A distillable, shaly coal found in Tasmania in Carboniferous coal beds with an average thickness of 1–2 m (3.3–6.6 ft) mainly consists of carbonized polliniferous grains. Tasmanite contains a resinlike polycyclic compound with a molecular weight of about 1,400: $C_{90}H_{134}O_{15}(OH)_2$. The pyrolitic products of this compound are liquid hydrocarbons with 72.62 percent C, 9.18 percent H, and 18.21 percent O. The components of tasmanite are, respectively: kerogen 25 percent, clay 9 percent, carbonate 1 percent, pyrite 0.1 percent, and silicates 65 percent.

Green River Oil Shale. The oil-shale layers in the basins of the Piceance, Uinta-Washakie, and the Green Rivers cover exactly 42,000 square kilometres (16,216 square miles). The sedimentary layers extend over the states of Wyoming, Utah, and Colorado. At present they form the world's largest known natural hydrocarbon reserve.

The oil reserve in this territory is an estimated 270 $\times 10^9$ tons. Particularly exploitable layers, rich in oil, can be found in the Piceance Basin in northwestern Colorado, where a ton of oil shale yields 95 litres (.60 barrels) of oil. In the oil-shale layers of the Green River, the organic substances, varying from zero to 80 percent, consist mainly of fungus spores, but there are also spore and pollen grains of higher order plants. The Green River Formation itself is situated on a lacustrine bed, following a lenticular pattern, in a layer of 460 to 610 metres (1,500–2,000 feet) thick, flanked by marly sediments of the Wasatch and Bridger formations. The formation is of Eocene age, and the shape of the layers suggests that the lake underwent major vertical and horizontal changes during this time interval.

In the U.S., apart from oil shale, other oil-impregnated sediment reserves are estimated at 500,000,000 tons.

Satmar Torbanite. There is a torbanite occurrence at Satmar, South Africa, eight miles north of Ermelo. The layer is 135 centimetres (53 inches) thick. The upper 51 centimetres (20 inches) make up a coal layer of moderately laminated structure, which changes vertically into torbanite. The torbanite layer proper is only 36 centimetres (14 inches) thick. The microscopic investigation of the torbanite layer revealed that the layer proper is made up of the fossil remains of huge algal colonies of *Botryococcus braunii* or *Elaeophyton coorongiana.*

Kuckersite. A special boghead coal on the territory of the Estonian S.S.R., termed kuckersite, dates back to Ordovician-Silurian time. It is a relatively soft, light brown rock of maritime origin; its algal material is *Gloecapsomorpha prisca,* but its fauna numbers 230 maritime species. Kuckersite usually occurs together with bituminous shaly limestone. Its kerogen content is of uniform composition, with an 18–20 percent content in the upper layers and 40–50 percent in the lower ones. The Baltic oil-shale field extends over several basins with its layers almost completely horizontal and at a depth of 5–10 (16–33 feet) or 100–200 metres (330–660 feet). The total thickness of its six occurrences is 4–5 metres (13–16 feet), which, on the territory of the Estonian S.S.R., is made up of six shale layers (about 3.0–3.2 m [9.8–10.5 ft]), and four levels at Leningrad and Gdov (about 1.8–2.0 m [5.9–6.6 ft]). Its heating value approximates 3,500 calories per kilogram. At distillation it yields 18–20 percent crude oil. The Baltic oil-shale reserve amounts to an estimated 14 $\times 10^9$ tons, of which 10.5 $\times 10^9$ tons are in the Estonian S.S.R. The growth of the Estonian oil-shale production for a recent 25-year period is as follows (in millions of tons): 1940, 1.9; 1950, 3.5; 1955, 7.0; 1960, 9.2; 1965, 15.8.

There are additional important oil-shale occurrences in some other areas of the U.S.S.R.: in the eastern part of the Kazakh S.S.R. (Zaysan), in the territories to the east of Lake Ladoga and Lake Tsud, in the Caucasus, in the Ukraine and in western Siberia, and also in the Volga Basin.

Chinese deposits. The largest oil-shale occurrences are in Fushun and Huatian, in northeastern China, where the largest coal basins also are located. The oil-shale deposits contain 6–10 percent oil. In Manchuria, large-scale coking of bituminous coals has been started. In a similar way, oil shales are also cokified in other parts of the country, too. The quantity of exploitable oil in the Chinese oil fields amounts to an estimated 1 $\times 10^9$ tons. In 1955 the amount of processed shale oil was less than 1,000,000 tons; in 1965 it increased to 9,300,000 tons.

Brazilian deposits. Brazil has tremendous oil-shale reserves in the states of São Paulo and Paraná. These oil shales, of Tertiary formations, rank among the most significant occurrences in the world. The investigation of the Brazilian deposits and the final estimation of their contained reserves was underway in the 1970s.

BIBLIOGRAPHY. A great deal of information concerning the characteristic properties of oil shales, their sources, and special problems of exploitation may be found in the proceedings of meetings, such as the "Proc. Fifth Symposium on Oil Shale," *Colo. Sch. Mines Q.,* vol. 63, no. 4 (1968); and *Oil Shale and Cannel Coal,* vol. 2 (1951), devoted to the proceedings of the Second Oil Shale and Cannel Coal Conference. The transitional character of kerogene rocks, their limnological and their stratigraphical properties are treated, together with the general features of coal sediments and petroleum, in the following works: B. NAGY and U. COLOMBO (eds.), *Fundamental Aspects of Petroleum Geochemistry* (1967); A.I. LEVORSEN, *Geology of Petroleum,* 2nd ed. (1967); W. FRANCIS, *Coal: Its Formation and Composition* (1954); and D.W. VAN KREVELEN, *Coal: Typology, Chemistry, Physics and Constitution* (1961).

Data on world distribution, exploitation, and technology, with many references, are included in H.S. BELL, *Oil Shales and Shale Oils* (1948); and N.I. ZELENIN *et al.,* "Chemistry and Technology of Oil Shales," in Russian (1968); and F. MAYER, *Erdöl Weltatlas* (1966). A large source of modern information on oil shales is presented periodically in the following journals: *Oil and Gas Journal* (weekly); *World Oil* (14/yr.); *Zeitschrift für Angewandte Geologie* (monthly); *Erdöl-Erdgas-Zeitschrift* (monthly); and *Erdöl und Kohle* (monthly). Each year maps, production figures, and geological data for all known oil and gas pools in the entire world are published in August by *World Oil* and in December by the *Oil and Gas Journal.*

(M.Sz.)

Okavango River

The Okavango River, the fourth longest river in southern Africa, runs for 1,000 miles (1,600 kilometres) from Angola, where it is known as the Cubango, to the Kal-

ahari. Flowing generally southeastward from the point where it rises, it forms part of the Angolan–South West African border before it crosses South West Africa's narrow Caprivi Strip to enter Botswana, where it empties into the vast Okavango Swamp in the desert. The river—formerly sometimes called the Okovango—derives its name from the Okavango people of South West Africa. David Livingstone, the Scottish missionary and explorer, and the first European known to have seen the Okavango, reached the swampy delta in 1849.

Although it often brings water to a parched land, the river's resources remain largely unused, and its banks are only sparsely settled. In the early 1970s various plans for the utilization of its immense water resources were being considered, in part with the assistance of the United Nations Development Programme. They included irrigation schemes for the river's middle course, the diversion of water from the swamps for industrial development, and the establishment of the Okavango Swamp as a wildlife sanctuary and tourist attraction. (For a related physical feature, see KALAHARI.)

The river's course. *The upper course.* The river rises as the Cubango just south of Vila Nova, Angola, on the Bié Plateau, at an altitude of 5,840 feet. The turf-clad granite plateau surface receives an annual rainfall of 52 inches and is also the source of other smaller headstreams, including the Cuchi River, an important tributary.

The Cubango River

For the first 230 miles of its course, the Cubango flows over hard granite, gneiss (a coarse-grained rock in which bands containing granular minerals alternate with bands in which schistose materials predominate), and metamorphosed rocks that were formed during the Precambrian Era, more than 570,000,000 years ago. The river is frequently broken by rapids. Its generally uneven and occasionally rugged course suggests that in this area the river flows over underlying older rocks, whereas at an earlier time it may have traversed an overlying, younger surface. Below the Cuchi confluence near the town of Caiundo, the riverbed narrows abruptly from about 900 feet to about 300 feet. There, the fury of the falling waters gives the rapids their name of Richo Riacantento, or Eye of the Rock.

The middle course. To the south of Caiundo, the river passes from the hard-rock surface to the Kalahari sand and the character of its valley changes. The river channel cuts through the western side of a valley floor less than a mile wide and bounded by cliffs of compacted sand capped by ancient river gravels. As it flows toward the South West African frontier, the river continues to erode its valley to depths occasionally exceeding 200 feet. Its channel is often bounded by almost vertical cliffs in which layers of hard silcrete (a quartzite) and calcrete (a limestone) are exposed. An interesting and somewhat puzzling feature in this stretch is the Tandaué Chana, a dry channel that runs parallel to, and some 10 to 15 miles to the west of, the Cubango. The channel is a relic of former flow patterns during times of flood. It turns abruptly eastward and joins the river before it reaches the international boundary.

Below its confluence with the Cuito River, the Cubango has removed the sand cover from parts of the underlying rocks to expose great areas of lavas between 190,000,000 and 136,000,000 years old. Near Andara, in the Caprivi Strip, ancient quartzites also have been exposed. At the end of this rocky stretch the river, again flowing to the southeast, once more passes onto the Kalahari sand surface at Popa Falls, after which it is known as the Okavango.

The lower course. Below the falls, the river enters its swamp tract in Botswana territory. For the first 70 to 80 miles it flows between a well-defined sand scarp on the northeast and a low rise in the sandveld surface to the southwest. During this stretch, the width of the swamp increases from a minimum of about one mile in the north to a maximum of seven miles in the south. This relatively narrow section then comes to an abrupt end, after which the river spreads out to form a triangular-shaped delta, the base of which extends for about 150 miles. It forms a drainage line—its parts known as the Thamalakane River and the Ngabe River—that drains southwest toward Lake Ngami and northeast to the Mababe Depression. Flow from the Thamalakane occurs sporadically through the Boteti (or Botletle) River to Lake Xau (Dow) to the southwest and to the Makgadikgadi Depression in the west; this flow may not occur for years at a time.

The delta. The delta region is known as the Okavango Swamp and covers an area of about 6,500 square miles (16,835 square kilometres). It is divided into two relatively distinct sections by Chief's Island, which stretches for about 40 miles in a southeasterly direction and is about ten miles wide. The main channel issuing from the river is called the Ng-gokha River and divides at the island into an eastward continuation of its own channel to the north and the Taokhe (or Taoge) and Boro channels to the west. The Taokhe, the westernmost channel, is almost entirely choked by a dense growth of papyrus, while the Boro is the main distributary to the Thamalakane River. To the northeast of Chief's Island, the Ngokha divides into the Borokha-Santantadibe Channel to the east and the Moanachira Channel that forms the northeastern limit of the swamp and flows into the Mababe Depression.

The Okavango Swamp

Peoples. In Angola, between Caiundo and Cuangar, the dry, sandy thornbush country through which the river flows is virtually uninhabited, and vast areas are uncharted. The riverbanks, however, are thinly settled by the Kwangare tribes, who live mainly by fishing and subsistence agriculture. A few small bands of Bushmen roam the territory between the Cuito and Cuando rivers. The South West African territory on the river's banks is inhabited by the Okavango people, who engage in agriculture and stockbreeding. The Kaba (Bayei), Bagereku, and Mbukushu tribes live on the islands in the swamp and practice hunting and some agriculture.

Since 1884, the delta has been recognized as part of the Batawana Tribal Preserve. An important administrative problem has been the control of the tsetse fly. Infesting the southern delta, the fly has pushed the pastoral Batawana southwest into Ngamiland.

Hydrology. The river is subject to changes in flow along its length because of seasonal variations, the reception of tributaries, evaporation, and absorption into the riverbed. In its upper course, the rate of flow varies between 700 and 3,000 cubic feet per second. At Caiundo and Cuangar, both situated on the middle course, the rate of flow varies from 1,765 to 2,120 cubic feet per second in the dry season to 10,600 in the wet season. Its annual average discharge is 105,930,000,000 cubic feet. At Runtu (Rundu) above the Shimpuru Rapids, the average annual discharge increases to between 141,260,000,000 and 459,100,000,000 cubic feet. Below the confluence with the Cuito River, the maximum annual average discharge again increases to 635,670,000,000 cubic feet. Along the river's lower course at Shakawe, however, the rate of flow averages about 13,000 cubic feet per second and the annual discharge decreases to 388,460,000,000 cubic feet. It takes four months for the flood peak to travel from Shakawe, through the swamp, to Maun, the administrative centre on the Thamalakane River. The average rate of outflow at Maun represents about 5 percent of the inflow at Shakawe.

Approximately 770 square miles of the swamp's total area of 6,500 square miles are permanently flooded, and between 3,800 and 5,400 square miles are inundated during the rainy season. The remainder of the delta is no longer affected by floodwaters.

The delta channel system is continually changing. The dense growth of papyrus reeds throughout the region is continually blocking channels, thus changing patterns of flow, while older blockages are occasionally removed by the movement of herds of hippopotamuses or other aquatic animals. The flow pattern is also changed by earth movements, or seismic shocks, apparently caused by the unstable nature of the bedrock. Because the gradient from north to south is insignificant, changes in levels of only a few inches can result in altered channels.

Vegetation and animal life. In its headwater course, the river flows through woodland country, known in the

vernacular as *myombo*, and containing mainly dominant species of bark cloth trees. On the Kalahari sand, to the south of Caiundo, the woodland is less dense and is characterized by the Rhodesian teak tree (*Baikiaea plurijuga*); the area is called the Baikeaea Woodlands or the Dry Forest. Little is known of the vegetation to the north of the river in this zone: the western area, called Chiculecandi Sandveld, and the eastern area, to the north of the delta, known as the Mbunda, are said to be occupied by dry savanna (grassland with scattered acacia trees).

In the delta the vegetation is generally considered to be of two main types—the riverine and flood plain vegetation, which is typical of the swamps, and the woodland or savanna, which occupies the slightly higher parts of the delta and the marginal areas. The riverine vegetation includes aquatic communities of such plants as dense papyrus, which grows to a maximum height of about 15 feet above the water level. Among the reeds and sedges are reeds which grow to between 10 and 15 feet above water level and smaller bulrushes of about five to 12 feet. On the seasonally flooded areas are many species of grasses. On the higher ground the mopane tree is generally found on clay soils, and camel thorns and palms are more common on the sandy soils. The Moremi Wildlife Reserve covers 700 square miles of the northeastern corner of the Okavango Swamp. Its teeming wildlife is representative of that found throughout the delta region. Game animals include the lions, cheetahs, buffalo, and wildebeests, and there are herds of hippopotamuses, zebras, and wild dogs. Crocodiles also are to be found. Birds include storks, ibis, herons, egrets, cranes, and weaver birds. There are also numbers of ducks, geese, and quail. Varieties of fish include bream, pike, barbels, and tiger fish.

Economic development. *Transportation.* The river is unnavigable except for small craft. Papyrus canoes are used for transport in the swamp area. At Jangada de Cahoco a ferry crosses the river between Angolan and South West African territory.

Plans for utilizing the river's waters

Agriculture. In South West Africa plans were made in the late 1960s to increase agricultural production in the river basin by irrigation. Seven separate schemes are proposed, but all are on a small scale; since the river gradient is small, water cannot be obtained by gravity but has to be pumped.

Industry. The discovery of diamonds, near Orapa, in Botswana, to the east of Lake Xau, has raised the possibility of using the waters of the delta for industrial purposes. If this project is approved by the Botswana government, it will require the clearance of papyrus and the maintenance of a regular flow of water into the Thamalakane and Boteti rivers. Water can then be stored in a reservoir near Lake Xau, and pumped some 30 miles to Orapa.

Tourism. Botswana is developing the Okavango Swamp and the Makgadikgadi Depression as a major wildlife area. Lodges have been built, and the government hopes to attract more tourists to the area.

BIBLIOGRAPHY. C.J. ANDERSSON, *The Okavango River: A Narrative of Travel, Exploration, and Adventure* (1861, reprinted 1968), an early account by a game hunter; A.G. STIGAND, "Ngamiland," *Geogrl. J.*, 62:401–419 (1923), an early description of the delta region; A.L. DU TOIT, *Report of the Kalahari Reconnaissance of 1925* (1926), an investigation of a scheme to direct the Okavango waters to South Africa; L.A. MACKENZIE, *Report on the Kalahari Expedition, 1945* (1946), a summary of previous investigations; J.H. WELLINGTON, *Southern Africa*, vol. 1 (1955), general background; H.W. STENGEL, *Water Affairs in South West Africa* (1963); K.L. TINLEY, *An Ecological Reconnaissance of the Moremi Wildlife Reserve, Northern Okavango Swamps, Botswana* (1966); A.T. GROVE, "Landforms and Climatic Change in the Kalahari and Ngamiland," *Geogrl. J.*, 135:191–212 (1969), new material on the delta, also deals with the associated Boteti River and the Makgadikgadi Depression.

(J.H.We.)

Okhotsk, Sea of

The Sea of Okhotsk (Okhotskoye More in Russian; spelled Okhotskoje More in the transliteration system of the Akademiya Nauk), which borders the northwest Pacific Ocean, is enclosed by the east coast of Asia from Cape (Mys) Lazarev to the mouth of the Penzhina River, by the Kamchatka Peninsula and the Kuril Islands to the east, by Hokkaido (Japan) to the south, and by the island of Sakhalin to the southwest. Both Japan and the Soviet Union are washed by the Okhotsk, which covers 611,000 square miles (1,583,000 square kilometres), has an average depth of 2,549 feet, a maximum depth of 11,069 feet, and a volume of 294,500 cubic miles.

Physical characteristics. For the most part, the continental shore is high and rocky and interspersed with the mouths of large, regional rivers that flow into the sea—the Amur, Uda, Okhota, Gizhiga, and Penzhina. In the northern areas of the Sakhalin and Hokkaido islands, the shores are predominantly lower. The Gulfs of Aniva and Terpeniya are found on the southeastern coast of Sakhalin. Nearly all islands—Shantar, Zavyalov, Spafaryev, Yam, and others—are situated close to the shore; only the island of Ion is in the open sea.

Geologic history

The present formations of the Okhotsk Sea Basin were evolved in the Quaternary Period (2,500,000 to 10,000 years ago). The geological history of the sea shows that the forming of the present hollow was accompanied by repeated glacial transgressions and regressions. The sea bottom has a slope from north to south and southwest. The northern and northwestern parts constitute a continental shallow up to 650 feet in depth, while the remaining area (about 70 percent of the area) is a continental slope from 650 to 5,000 feet deep, on which isolated underwater heights, depressions, and troughs stand out. The deepest part of the sea, the southern part west of the Kuril Islands, is 8,200 feet and deeper. Bottom deposits in this deep basin are a clay–diatom silt, whereas approaching the shore there are fine, silt-covered sands, coarse sands, and pebbles mixed with mussel shells.

Climate. The northeastern, northern, and western regions of the Okhotsk experience severe weather during the winter, because of the influence of the Asiatic continent; from October through April these areas experience very cold air temperatures, are constantly covered with ice, and have very little precipitation. In short, a continental climate pervades these parts of the Okhotsk Sea. To the south and southeast, the proximity of the Pacific results in a softer marine climate. The coldest months in the Okhotsk Sea are January and February; the warmest are July and August. In the northeastern part the average monthly air temperature during February is −4° F (−20° C), while in August the average is 54° F (12° C). To the north and west of the sea, average monthly air temperature is −11° F (−24° C) in February and 57° F (14° C) in August. In the southern and southeastern parts, the average monthly air temperature is 19° F (−7° C) in February and 64° F (18° C) in August. The yearly precipitation averages in the north 16 inches, in the west, from 28 inches; and, in the south and southeast, about 41 inches.

Hydrology. The water of the Okhotsk Sea consists of continental drainage, precipitation, and waters flowing from the Pacific Ocean through the Kuril Straits and from the Sea of Japan through the La Perouse and Nevel straits. During the summer months the Okhotsk is warmed to a depth of 100–165 feet. The water temperature on the surface reaches 46° to 54° F (8° to 12° C) and the salinity 32.5 parts per thousand and lower. Deeper water has a temperature of 30° to 29° F (−1° to −1.8° C) and salinity of up to 33.75 parts per thousand. The thickness of the cold-water layer fluctuates from a few feet in the southeastern part of the sea to 245–525 feet in the northwest.

Water currents

The general movement of water in the Okhotsk is counterclockwise. Water flows from the Sea of Japan into the Okhotsk, accounting for the comparative warmth of its southwestern part. Warm water is also carried into the Okhotsk by Pacific currents. Because of their influence, the waters of the eastern half of the sea are warmer than those of the western part. For the most part, the currents flow clockwise around the Kuril Islands; in the northern half of the straits they flow into the Okhotsk but in the

southern half return into the Pacific. The Gulf of Shelikhov experiences the strongest tides (42.3 feet); the weakest tides occur at southeastern Sakhalin (2.6 feet). Ice cover appears at the end of October and reaches its greatest extent in March. In the coastal areas it welds to the shore, but in the open sea there is floating ice. The ice vanishes in June, with the exception of the Sakhalin Gulf and the Shantar Island region where ice floes are not uncommon in July and sometimes even in August.

Marine life. The river drainage, the intensive intermingling of waters by straits and wind, and the surge of deep ocean waters are all favourable to marine life. In months when it is warm enough, there occurs an extraordinarily rapid spread of life. The flora is represented by algae and seaweed and the fauna by crawfish, sea mussels, crabs, sea urchins, polyps, and various types of fish. Salmon, herring, pollack, flounder, cod, capelin, and smelts (or frostfish), as well as crab and shrimp, are all commercially important. The Okhotsk is also inhabited by mammals—whales, seals, sea lions, and fur seals. In 1965 record catches of fish in the Okhotsk totalled 1,500,000 tons.

Navigation. Regular navigation connecting the ports of the Soviet Far East is conducted through the Okhotsk. On the continental coast the most important ports are Nagayevo and Okhotsk. Korsakov on the island of Sakhalin and Severo-Kurilsk on the Kuril Islands are also important. During the winter, ice floes are an impediment to sea navigation, and dense fog is a hindrance during the summer. Strong currents and submerged rocks are other perils. The Okhotsk Sea region is playing an important role in the economic development of the Soviet Far East.

(T.Y.S.)

Oklahoma

In its land and its people, Oklahoma is a state of contrast and of the unexpected. The terrain varies from the rolling, timbered hills of the east, where it borders Missouri and Arkansas, to the treeless high plains that run on into Texas and New Mexico to the west. Its east central region is dominated by the lowlands of the Arkansas River, sweeping in from Colorado and Kansas on the north, and its tributaries and by the Red River, which forms nearly all of its southern border with Texas. Once basically agricultural—and the dust-bowl locale of John Steinbeck's famous novel *The Grapes of Wrath*—the state of Oklahoma now has hundreds of lakes and a diversified economy.

The word Oklahoma is derived from two Choctaw Indian words: *okla*, "people," and *humma*, "red." During the 19th century the future state was a symbol of one of the least glorious chapters in American history, becoming known as Indian Territory, the dumping ground for Indian tribes displaced by the white man's ever-increasing hunger for land. Since its admission in 1907 as the 46th state of the Union, however, Oklahoma has achieved an integration of its Indian citizens into modern economic and social life that probably is unmatched by any other state. There is no reservation in the usual sense for the Indian population. Though numbers of "blanket Indians" possess no more than their bedroll, others have risen to positions of distinction. Many share with their fellow Oklahomans the great wealth that oil resources have brought to the state.

In the census of 1970, Oklahoma's 69,919 square miles (181,089 square kilometres) held a population of 2,559,229, about 7 and 4 percent of whom were, respectively, black and Indian and about 70 percent of whom lived in urban areas. The state sent a considerably greater proportion of its high school graduates to college than the national average in the early 1970s, and the University of Oklahoma had become second only to Harvard in the number of Rhodes scholars among its graduates. The customs of the Deep South are reflected in the habits and attitudes of southern Oklahoma—"Little Dixie"—where cotton production, but not loyalty to the Democratic Party, has declined. The wheat growers in the north, however, show their largely Kansan origins by their dedication to Republican politics. (For further information about Oklahoma within larger geographic and political contexts, see UNITED STATES; UNITED STATES, HISTORY OF THE; NORTH AMERICA; and GREAT PLAINS.)

THE HISTORY OF OKLAHOMA

Early habitation and European exploration. Of the newer states, Oklahoma is one of the oldest in terms of human occupation. The abundant game of its plains attracted hunters of the Clovis and Folsom cultures 10,000 to 15,000 years ago. Others followed, producing between AD 500 and 1300 a golden age of exquisite pottery, textiles, sculpture, and metalware. Evidence indicates a widespread system of trade and communication. This high culture apparently fell before the onslaught of primitive people from the western plains, and until the expedition of Francisco Vásquez de Coronado in 1541 the region's population included representatives of at least three major Indian language groups.

Coronado claimed the area for Spain, but it became little more than a highway for wide-ranging Spanish explorers. In 1714 Juchereau de Saint Denis visited Oklahoma, and subsequent Frenchmen established a fur trade with the Indians. France and Spain struggled for control until 1763, leaving only the natives to contest Spanish authority until the return of the French flag in 1800. Three years later, through the Louisiana Purchase, Oklahoma was acquired by the United States.

American dominion. As one of the purchase's most attractive parts—because of trade opportunities—the area might well have become one of its first states; but it was, in fact, the last. Because of hostile Indians, Spanish intrigue, the mislabelling of its treeless plains as the American Desert, and the pressure for removal of the Indians from the settled East, Congress in 1828 reserved Oklahoma for red men and required all whites to withdraw. By 1880 more than 60 tribes had joined the local ones in Indian Territory. Some were sedentary, peaceful, agricultural, and semicivilized, others were migratory, belligerent, and barbaric. Indian sovereignty remained unchallenged until early in the Civil War when unhappiness with the federal government caused leaders of the Five Civilized Tribes—the Cherokee, Choctaw, Chickasaw, Creek, and Seminole—who had come from Southern states and some of whom owned plantations and held slaves, to sign treaties annexing Indian Territory to the Confederate States of America.

The Indian Territory

Not only did Indian land become a battleground during the war but defeat also brought other losses. The Reconstruction treaties required, among other things, land cessions to the former slaves, the resettlement of additional outside tribes, and railroad rights-of-way. Although a scheme to colonize black freedmen in Oklahoma never materialized, the weakness of the Indian governments encouraged both blacks and whites from adjoining states to trespass illegally. Thus, the territory again became a dumping ground for Indians and an even greater cultural hodgepodge of red, white, and black people.

White settlement and statehood. Railroads seeking revenue and land-hungry whites, both inside and outside the territory, coveted the Indian's land. By 1879 organized bands, the Boomers, were moving in despite federal law. Although most were ejected, pressure continued until Congress opened some 2,000,000 acres of western Indian Territory, bringing on the famous land run beginning at noon on April 22, 1889. Known as Oklahoma Territory, the new area came to include, through further land runs, about half of the former Indian domain. Then its settlers, many called Sooners for entering the area before official permission, sought union of the two territories in statehood. The remaining Indian Territory was dissolved by assignment of lands to the various tribes, and the Indians joined in approving the constitution of the proposed state in 1907.

The drought years of the 1930s blighted many rural areas of Oklahoma, driving thousands of farmers into long migrations in search of some form of livelihood. The economic boom of World War II allowed the economy to diversify. The major continuing event of the 20th century probably was the continuing exploitation of the state's vast petroleum reserves.

THE NATURAL AND HUMAN LANDSCAPE

Lying in a transitional zone in topography, climate, and other features, both east to west and north to south, Oklahoma comprises a jumble of environments.

Physical regions. Three of the nation's large physical regions extend into or across the state. The Interior Highlands is in the east; the Coastal Plain, extending through Texas to the Gulf of Mexico, is in the south; and the Interior Plains, including the Central Lowland and Great Plains, covers the remainder. Ten subregions lie within Oklahoma. Three are mountainous and in the south—the Ouachita, Arbuckle, and Wichita mountains—and are characterized by rough topography and thin soils. They have lumbering, grazing, some farming, and mining as their principal income. The northeastern Ozark Plateau, most of which lies in Missouri and Arkansas, has rough terrain and small fields devoted primarily to fruit and vegetables. Once important as a lead and zinc producer, it has a Cherokee heritage and beautiful rivers that make it a major recreation and tourist attraction.

The Sandstone Hills, a wide band stretching through the east central portion between the Red River and the Kansas border, is poor in agriculture and timber but important for oil, gas, and coal deposits. The region is sprinkled with deserted or dying oil-boom towns, with Tulsa a prosperous exception. The sparsely populated Gypsum Hills section of western Oklahoma is devoted largely to grazing and farming, with large wheat acreages in the north and smaller cotton farms in the south.

The remaining four areas are flat to rolling and agricultural. The Red River Plains, once the area of the best farmlands in the state, has been depleted by cotton. Its agriculture has been diversified by the addition of peanuts, melons, and vegetables grown on medium-sized plots. Its population is relatively dense, with many small towns serving as trade centres. The Prairie Plains region in the northeast is marked by grazing in its rougher portions and vegetable farms in the river valleys. Oil and gas fields are common, as is strip-mining for coal. It contains a number of middle-sized towns, some of which have small manufacturing plants. The Red Beds region is the largest, running through the middle of the state. The greatest population density is here, as are most of the larger towns. Oil provides much of the income. Although cotton rules in the south and wheat in the north, corn, watermelons, sorghum, alfalfa, vegetables, and livestock are common. The High Plains region of the northwest and the Panhandle offers a marked contrast. With the highest elevation and least moisture, wheat dominates the eastern portion and grazing the western.

Rivers and climate. Oklahoma's drainage pattern, consisting of the Arkansas and Red rivers and their tributaries, slopes from an elevation of 5,000 feet in the northwest to 500 feet in the southeast. Rainfall varies from more than 50 inches (1,270 millimetres) annually in the Ouachitas to less than 15 inches (380 millimetres) in the western Panhandle. Wheat and sorghum predominate in the drier western sections, whereas corn (maize), vegetables, and berries grow in the damper east. Virtually all of the ten regions have enough water for grass; hence, ranching is common.

Moisture and ecology

Oklahoma has a southern humid belt merging with a colder northern continental one and humid eastern and dry western zones that cut through the state. The result is normally pleasant weather and an average annual temperature of about 60° F (15.5° C), increasing from northwest to southeast. No region is free from wind; and, as the collision point for warm and cold air masses, with sudden rises and falls in temperature, the state has heavy thunderstorms, blizzards, and tornadoes.

Vegetation and animal life. Oklahoma is a transitional area for plant and animal life. More than 130 trees are native: the eastern forest of maple, sweet gum, hickory, oak, and pine phases into the cottonwood, elm, hackberry, and blackjack and post oaks of the grasslands. The arid zone plants are chiefly mesquite, sage, and cacti. There are deer, elk, antelope, rabbits, coyotes, wolves, foxes, prairie dogs, and the American bison. Native fish include bass, perch, catfish, and buffalo, and virtually every bird common to the land between the Mississippi and the Rockies is found. Horned toads, lizards, many varieties of nonpoisonous snakes, and the rattlesnake and the cottonmouth moccasin are native.

Human settlement. The outlines of roads and farms generally produce a pattern of unusual symmetry in the landscape, revealing the original survey divisions into townships, sections, and quarter sections. Small squares predominate where small-scale farming is common and very large ones where wheat and ranching prevail. As elsewhere in the nation, however, the trend has been toward urbanization. The Red Beds in the centre of the state grew most rapidly, and three of the state's four largest cities are found here, the exception being Tulsa.

Oklahoma City, the capital, is near the centre of the state and in area comprises one of the largest metropolitan areas in the nation. Banking, insurance, manufacturing, trade and transportation, state and federal installations, and educational facilities have made it the commercial and industrial heart of the state. Tinker Field, an Air Force base and the state's major employer, is located in nearby Midwest City. Tulsa, a former Creek Indian village in the Sandstone Hills region, grew slowly until the discovery of oil in the vicinity. Its refineries and facilities for manufacturing and distributing oil-field supplies make it the logical headquarters for numerous oil companies, and it has numerous other financial and industrial functions. Lawton is a centre for the Ft. Sill army reservation, the Wichita wildlife recreational centre, and the rural population of the area. Norman, the seat of the University of Oklahoma and the major state mental hospital, is also a "bedroom" city for commuters to Oklahoma City and Midwest City.

THE PEOPLE OF OKLAHOMA

Composition. During the 1960s, Oklahoma's Indians, most of whom live in the former Indian Territory in the eastern part of the state, increased in greater numbers than other races. The Plains tribes remain in western Oklahoma. Most of the blacks are descended from slaves of the Five Civilized Tribes, although some migrated from the South after 1865 and others came during the land runs. The majority live in urban centres or in the southern and eastern parts of the state, and several towns have entirely black populations.

Ethnic and religious groups

A wide variety of other racial and ethnic strains have contributed to Oklahoma's population. The original French claimants left their names and bloodlines, usually in conjunction with Indian families, and a mining boom in the 1870s brought an influx of Europeans into the Choctaw Nation. Descendants of these Italian, Slavic, Greek, Welsh, Polish, and Russian miners still live in Little Dixie. The land rushes brought homesteaders from China, Japan, Mexico, England, France, and Canada, and the spread of wheat farming attracted German Mennonites and Czechs to the northwest. German was the language of instruction in some public schools as late as World War I. By the 1970s, however, nearly all of Oklahoma's residents reflected a typically Midwestern American culture.

Religion. The state's religious sects bear out this trend toward uniformity. Of the Protestant majority, the Southern Baptists and the Methodists predominate, and the resultant conservatism has placed Oklahoma in the "Bible Belt." (This fundamentalism was a primary cause for Oklahoma's retention of the prohibition of liquor sales until 1959.) Other leading denominations include the Disciples of Christ, Presbyterians, Congregationalists, and Episcopalians. Roman Catholics and Greek Orthodox are represented throughout the state, but Jewish synagogues are limited to the cities. Most Indians have adopted some form of the white man's religion, although the Native American Church—in which use of the drug peyote is a part of the worship—is recognized by state charter. The Sun and Ghost dances of the western tribes reflect more primitive religious practices as well as reactions to the white man.

Demography. Though the average Oklahoman has kept his religious ties, he has changed his place of resi-

dence. In 1890, fewer than 4 percent of the population lived in towns, but by 1950 50 percent were urban, half of them living in the metropolitan areas of Oklahoma City, Tulsa, and Lawton. In the 1960s more counties lost population than gained, though the gains were far greater than the losses. Of the 26 cities of 10,000 or more in 1960, 21 continued to grow. Part of this willingness to move may be explained by the youthfulness of the people, one-third of whom are under 18 years of age.

It is predicted that Oklahoma's rate of growth below the national average will continue, with the growth continuing most heavily in the north central and northeastern areas, with Oklahoma City and Tulsa as focal points.

The state's vital statistics approximated the national averages, including birthrates and deathrates and causes of death. Infant mortality was higher among nonwhites by a ratio of 33 to 21 per 1,000 live births. The state has a below-average number of physicians per 100,000 but is above average in number of dentists and nurses.

THE STATE'S ECONOMY

At the start of the 1970s, Oklahoma had a per capita income comparable to that of the neighbouring states of Arkansas, Louisiana, and Texas as a group, but more than 15 percent below the nation as a whole; its unemployment rate, however, was lower.

Agriculture. Traditionally, agriculture has furnished an important part of Oklahoma's income, though Oklahoma's more than 90,000 farms and ranches are slightly larger than the nation's average in size but of slightly less value per acre. The United States Department of Agriculture forecasts that, in line with national trends, the averages will remain the same, but the number of units will continue to decline. In commercial agriculture, livestock ranked first, followed by grain, dairy products, cotton, and other field crops and general produce.

Major nonagricultural employment. Oklahoma remains somewhat of an economic satellite of the industrial North and East, furnishing food, raw materials, and fuels. Despite great efforts to diversify, the state still has far to go. In 1970 only 13 percent of its employed were in manufacturing, about half of the national average. Wholesale and retail trade employed the greatest number of persons, followed by government, services, transportation and public utilities, finance, insurance and real estate, construction, and mining. Union membership has not kept pace with the increases in nonagricultural employment. In 1968 nearly 30 percent of all such workers in the nation, were unionized, but Oklahoma's percentage was only about 17. Man-days lost by labour stoppages were about half of the national average.

Manufacturing and labour organization

Mining and forestry. In 1968, Oklahoma ranked fourth nationally in mineral production, with an income of more than $1,000,000,000 from petroleum, natural gas, natural-gas liquids, and stone, in that order. One year later its more than 81,000 oil wells, topped by about 10 percent of the nation's active drilling rigs, produced 224,000,000 barrels annually, with a proven reserve of 1,390,000,000. Timber is also important. In 1963 Oklahoma had about 5,300,000 acres of commercially exploitable timber divided almost equally between softwoods and hardwoods. In 1970, the first major commercial effort was made to establish a pulp and paper industry.

Transportation. Oklahoma's transportation facilities partially account for its favourable record in attracting new industry in recent years. Tulsa and Oklahoma City act as the major collection and distribution points. Railroads operate about 5,600 miles of track within the state. Though the inbound rail traffic exceeds the outbound by two to one, in 1968 well over 4,000,000 tons of agricultural produce alone were shipped by train. Seven airlines provide direct flights for passengers and freight to most cities, while interstate and auxiliary state and federal highways support an excellent trucking system that employs nearly 200,000 people. Intricate networks of pipelines move the petroleum products, with the newest addition to transportation a barge traffic linking Tulsa to the Gulf of Mexico by way of locks and dams on the Arkansas River.

ADMINISTRATION AND SOCIAL CONDITIONS

Territorial governments. Oklahoma's political identity began with the establishment of autonomous tribal governments in the 1820s. Indian Territory consisted of five republics, or nations, with fixed boundaries, written constitutions, courts, and other governmental apparatus similar to those of the Eastern states. The major difference was that in each all land was held jointly, or in severalty, by the individual tribes. The first major threat to these governments came when, as former allies of the South, they were placed under military rule during the Reconstruction period. The treaties of 1866 committed the tribes to eventual union under a single government, but this had not been accomplished when Congress, in its haste to open the unassigned lands, admitted whites in 1889 without making any provision for government.

In this interval Congress steadily encroached on Indian sovereignty by extension of federal law and the demand that landholding practices be brought into conformity with United States custom. Finally, on May 2, 1890, the Oklahoma Organic Act established a territorial government and provided that all reservations in the western Indian Territory, when opened for non-Indian settlement, would be annexed automatically to the new Oklahoma Territory. The last step came when the remaining tribal lands in eastern Oklahoma were distributed among the tribes, and it was agreed that tribal governments would cease in 1906.

Constitutional framework and politics. That same year the Oklahoma Enabling Act set the formula for statehood: the two territories—Oklahoma in the west, Indian in the east—would unite and draw up a single constitution. The resulting document reflected the influence of the progressives and reformers who dominated American politics. It became inordinately long in an effort to protect against corporations, bosses, and future redefinitions of intent by judicial interpretation. The general structure is common to other states, but Oklahomans strengthened the legislature by limiting the governor's appointive powers and ability to succeed himself, the latter prohibition removed in 1970, and making the judiciary elective. Also unusual was the right of initiating legislation by popular initiative and referendum. The governor is elected for four years, whereas the 48 senators serve similar but staggered terms, and the 99 house members serve two years. Each of the original 75 counties, later increased by two, has at least one representative. Constitutional provision was also made for township and city governments, though the former was abolished in 1913. This constitution, often amended, is still in force.

Eight of nine territorial governors were Republicans, but after statehood Oklahomans favoured the Democrats. Even when the state supports Republican presidential nominees, normally that party can hope for but one or two congressional seats, and it was not until 1962 that it won the governorship. Nevertheless, Oklahomans have a history of giving strong support to third parties; in 1914 the Socialists received 53,703 votes, and in 1968 the Southern states' rights candidate, George Wallace, received more than 20 percent of the total vote.

Judiciary and law enforcement. A major governmental change was revision of the state's court system in 1967, which abolished justices of the peace and established selection of major judgeships according to what has become known as the "Missouri Plan." Under this plan judges are nominated by a joint commission chosen by the governor and the state bar association rather than by the political parties. The state Supreme Court has exclusive appellate jurisdiction in civil cases, while the Court of Criminal Appeals has exclusive appellate jurisdiction in criminal cases. The Court of Appeal, with a judge elected from each congressional district, hears only cases assigned it by the Supreme Court, and there is no appeal from its decisions to other state courts. Lower courts include 24 district courts, with judges elected on nonpartisan ballots. In 1966 county attorneys were replaced by 27 district attorneys.

Social issues and state involvement. One of Oklahoma's major sociopolitical issues since statehood has

Education, health, and welfare

involved civil rights. Although its black population is small, as a border state Oklahoma was one of the first to feel the demand for equal education. All public education was segregated, and blacks who wanted advanced or professional degrees were furnished tuition for out-of-state institutions until 1948, when the United States Supreme Court ruled that a black must be admitted to the University of Oklahoma Law School. In 1949 all graduate work and in 1955 all higher education was opened to any qualified applicant. Economic, social, and public school integration moved forward with a minimum of violence.

Supervision of public schools is under elected state and county superintendents, and higher education is coordinated by the regents for higher education, appointed by the governor. The state university system is often regarded as overextended in relation to the state's needs and resources. Exceptions are the University of Oklahoma (founded, 1890) in Norman and Oklahoma State University (1890) in Stillwater. Both have an unusual number of graduate departments ranked above average in achievement. Private institutions enroll only about one-fifth of the college population.

The Commission of Charities and Correction has general charge of mental hospitals in Norman, Vinita, and Fort Supply, the state penitentiary in McAlester, and the reformatory in Granite. The majority of other similar institutions are under the Department of Institutions, Social and Rehabilitative Services. In spite of a generally conservative attitude toward federal intervention in local social questions, most federal welfare programs are represented. In addition, more than 100 recognized agencies or groups within the state have resulted from minority initiative, by Indians, blacks, and Mexican-Americans.

For financial support of its functions, Oklahoma relies basically on taxes on gasoline, income, and sales. The sales tax is earmarked for welfare, and property taxes are left largely for the support of county, municipal, and school needs. Though in 1970 the state was only 36th in per capita personal income, its appropriations for education, highways, and public welfare comprised the bulk of its expenditures, helping to rank it 21st nationally in per capita general expenditures. It was 29th per capita in outlay for all education, 18th for higher education, 16th for highways, and fourth for welfare. Voter rejection of increased state aid to education in the early 1960s led finally to swift moves by the state legislature to avoid nationwide boycotts of the public schools by the National Education Association. A major check on spending since 1939 has been Oklahoma's "budget balancing" amendment, by which the legislature is forbidden to appropriate more money than in the previous year plus estimated additional revenues.

CULTURAL LIFE AND INSTITUTIONS

Oklahoma is a blend of the old and new. Cowboys and Indians, who undergird a tourist attraction worth nearly $450,000,000 annually, may be seen at the national finals of the Rodeo Association or at the American Indian Exposition. As host of the annual exposition and the site of Indian City U.S.A., the National Indian Hall of Fame, and the Southern Plains Indian Museum, Anadarko is a major tourist attraction. Among the features are full-sized reproductions of the homes of various tribes, pictures and busts of their leaders, and extensive displays of their artifacts. The National Cowboy Hall of Fame is noted for its Western art and its exhibits of cowboy paraphernalia. The Will Rogers Memorial Museum at Claremore stresses exhibits depicting early Oklahoma and Rogers' career as a cowboy and entertainer.

The arts. Oklahoma's best known graphic artists are Indian, and Indian culture as well as works of European masters are represented in many of the museums. Oil as a symbol of Oklahoma has placed derricks on the capitol grounds and made it influential on the international petroleum landscape, but those that it has enriched have contributed much to the artistic scene. The Gilcrease, Philbrook, and Woolaroc museums originally reflected individual tastes, but they have joined other art museums (notably the Oklahoma City Art Center) to offer wide-ranging displays.

The performing arts

Symphony orchestras are supported in Oklahoma City, Tulsa, and Lawton. A public school music program culminates each spring in the Tri-State Music Festival. Several ballerinas of international fame are of Oklahoman Indian descent, the most noted of whom are Rosella Hightower and the sisters Maria and Marjorie Tallchief. Theatres have been sources of entertainment since frontier days. Universities and civic groups continue to provide a wide variety of dramatic experiences and professional training. Several towns feature annual folk plays or pageants, while Tulsa boasts an opera company with a regional reputation. The 7,000-member Tulsa Little Theater has given nearly 50 years of uninterrupted productions, and the Oklahoma City Mummers have constructed a $1,500,000 building and become a resident professional company. The state is unusually active in literature, with numerous writers' clubs, poetry societies, and folklore groups.

Publication and communications. Oklahoma's reputation in publication began with the founding in 1837 of Park Hill Press in the Cherokee Nation. The University of Oklahoma Press publishes both regional and international material. *Books Abroad*, established by the university in 1927, is a unique literary quarterly that also offers a major literary prize. Outstanding research opportunities exist in the holdings of the Oklahoma Historical Society, the Gilcrease Museum Library, and the University of Oklahoma's collections on the history of the West, of business, and of science. Popular sources of information and entertainment are the 52 daily newspapers and the 95 radio and nine television stations.

Prospects. Oklahoma of the 1970s is experiencing rapid change. Improved transportation and communication have eliminated rural isolation, and its population has an increasingly urban outlook. Politically, there is a growing basis for a two-party system, with Republicans making significant gains in recent years. The major economic need is for types of secondary manufacturing that will complement the state's primary mining and agricultural industries. Many persons hope that such additions might provide a broader tax base, making possible better solutions to inadequate financing for education and other social concerns.

Oklahoma's history is unique among American states and far shorter than most in its relation to the modern nation. Infusions of new blood and ideas and the cultural integration and interaction of many citizens in statewide social, economic, educational, and political activities have broken down many historical patterns of distinction and discrimination. In many areas, racial and cultural assimilation is moving more rapidly than in other parts of the nation. It may be soon nearing the time when the citizens of Oklahoma will have little but geography and history with which to distinguish themselves from other Americans.

BIBLIOGRAPHY. The best single volume covering the history, politics, economics, and social conditions is A.M. GIBSON, *Oklahoma: A History of Five Centuries* (1965). For biographical sketches as well as general history, see GASTON LITTON, *History of Oklahoma*, 4 vol. (1957). An updating of an older WPA guide that is strong in local descriptions is KENT RUTH (ed.), *Oklahoma: A Guide to the Sooner State* (1957). The best single source for geography is J.W. MORRIS and E.C. MCREYNOLDS, *Historical Atlas of Oklahoma* (1965). M.H. WRIGHT, *A Guide to the Indian Tribes of Oklahoma* (1951), is essential to understanding the Indian history of the state. IRVIN HURST, *The Forty-Sixth Star: A History of Oklahoma's Constitutional Convention and Early Statehood* (1957), is excellent for the descriptions of Oklahoma's founding fathers. A brief description of Oklahoma's judicial system is found in MARTINDALE-HUBBELL, "Oklahoma Court Calendar," *Law Directory, 1969* (1968). A brief description of the organization and history of higher education is found in the OKLAHOMA STATE REGENTS FOR HIGHER EDUCATION, *The Oklahoma State System of Higher Education* (1971). The best coverage of the oil industry and of ranching is found in C.C. RISTER, *Oil: Titan of the Southwest* (1949); and E.E. DALE, *The Range Cattle Industry* (1960).

(J.S.E.)

Old Age, Social Aspects of

Old age is both a stage in the life of an individual and a segment or stratum of the population of a society. Aging individuals (or cohorts of individuals born at the same time) change over their life course as their personality develops, their experience accumulates, or they make adjustments to new roles. Not all individuals age in the same way, and the aging of each new cohort is affected by the unique situation of its particular era in history as norms, mores, and attitudes change or as wars or economic depressions may occur. Within a society, successive cohorts fit together at any given time to form the age structure of young, middle-aged, and aged segments, and over time the particular individuals composing any particular segment are continually moving on and being replaced. Thus the meaning of old age varies from time to time and from place to place, as individual aging and social change interpenetrate and affect one another.

Even among Western societies today there are no consistent definitions of old age. Biologists are not agreed upon the existence of an inherent biological process, apart from the onslaughts of the social and physical environment, that produces a state of physical deterioration in old age. Epidemiologists describe old age in terms of current conditions of morbidity and mortality in a population. Administrative eligibility for retirement, pensions, or social security may be variously set from ages in the 70s down to ages in the 40s. The public considers as old people those who are anywhere from less than 50 to over 75, and substantial proportions of persons who are in their 60s or older do not look upon themselves as old. Social researchers, who often set the boundary of old age at 65, must reckon with the fact that there are large numbers of people who are themselves old enough to be the parents of other "old people" in the category aged 65 and over. (In the United States, for example, there are some 900,000 persons aged 85 and over.)

THE AGED IN MODERN SOCIETY

The position of older people in modern industrial society must be interpreted in light of the fact that many of them survive beyond the age at which they would perform or be rewarded in major social roles: worker, parent of dependent child, or spouse. By contrast, in most primitive and agrarian societies, old age (at least up to senility) has been associated with some special characteristic regarded as an asset in the respective culture, such as usefulness in the performance of chores, skill in dancing or storytelling, control of property rights, power in the family, seasoned experience, or (especially in preliterate societies) extensive knowledge.

Comparisons of the aged in traditional and industrial societies

The status of the aged in the economically advanced nations today derives from broad changes over the past century in the entire society and its culture. Demographic change has altered the age composition of the overall population, and for the individual, the expectation of life has been increasing as mortality rates decline. Economic development has been accompanied by decreasing proportions of older men in the labour force, posing problems of income maintenance for those no longer economically active. Levels of educational attainment have been rising. Urbanization has transformed the residential setting. The structure of the family has altered. Thus the social context of the aged today differs markedly from the earlier situation when today's older people were reared and almost certainly from the projected future when today's younger people will have grown old.

The demographic base. In Western nations generally, there has been a marked rise in the number of older people, paralleling the expansion of the population as a whole. Over the first half of the 20th century the annual rate of increase in the age category of 65 and over has ranged from about 1 percent in France or Sweden to nearly 3 percent in Canada or the United States. In the United States the number in this age category rose from 3,000,000 persons in 1900 to nearly 17,000,000 by 1960.

Numbers of older people have been increasing even faster than the total population. Thus, in Sweden and France the proportion of the total population aged 65 and over had already reached 8 percent by the turn of the century; in the United States between 1850 and 1960, this proportion rose from 3 percent to 9 percent. This long-term tendency for the older age groups not only to grow larger but to grow faster than the younger segments of the population is characteristic of industrialized societies generally; by contrast, in most of the less developed countries of Asia, Africa, and Latin America, the proportions of persons 65 and over still fall below 5 percent. These changing proportions of old people within the population have resulted primarily from declines in fertility rates, which have reduced the accretions to the younger age categories, and not, as commonly supposed, from the declines in mortality. The improvements in reducing mortality have, in the past, had their effect most directly on infants and small children (though any future benefits in Western countries are likely to accrue at the later ages). Although the increasing proportions of survivors may eventually swell the ranks of old people, they have in the meantime added to the numbers at the young ages, not only initially when they were young themselves but later on when they became parents.

Not only have the number and proportion of older people been changing in the West but among older people the proportion of females has been rising. Except in some underdeveloped areas, male mortality rates are generally higher at all ages than corresponding rates for females (even for some animal species). Hence there is typically an excess of females over males among the aged. In the category of people age 75 and over in the United States, there are only about 73 men for every 100 women, a discrepancy that has been widening.

Increases in females

For individuals within a society, the secular declines in mortality (and the consequent increase in the expectation of life) have greatly improved the chances of surviving into old age. The change is not in the *span* of human life (the ultimate length of life attainable by a member of the human species) but in the proportion of people enduring into the higher ages of this span. Thus the average lifetime in ancient Rome or medieval Europe is estimated at some 20 to 30 years. In several Western countries, a man born in the middle of the last century could look forward to four decades of life; today a man can look forward to living well into his 60s, a woman to living into her 70s. Such demographic changes characterizing Western civilization appear to be unparalleled in human history. They have profound societal implications—for example, for government involvement in programs for older people, for the increasing supply of mature individuals available for the labour force, or for the structure of the family because all the family members, but particularly the females, tend to live longer.

Economic status. The age structure of society is closely linked to the economy. In this realm older people tend, on the average, to be comparatively disadvantaged, although individual variations are great. Their low status is indexed by their comparatively low rates of participation in the labour force and by their low average incomes.

Decline in employment

There has been a steady decline (apart from short-term fluctuations attendant upon mobilization for war) over several decades in the proportion of men 65 and over who are in the labour force in most industrialized countries (in contrast to agricultural countries). In the United States this proportion has dropped to less than one-third from approximately two-thirds in 1900; and estimates suggest a possible reduction to one-fourth by 1975, if present trends continue.

No firm explanations for the long-term decline in the employment of older men have been satisfactorily demonstrated. Extensive research on past trends and on age patterns today point, however, to a variety of associated factors and facilitating mechanisms, such as the possibility that technology may have advanced to the point of a sufficiency of manpower to meet existing demands for goods and services without the participation of older workers; the decline of agriculture and of self-employment, in which the aged are free to continue work or to taper off from it gradually; the steady rise in the educational level of the labour force, which appears to favour

the better educated and the more recently trained younger workers; and the extension of public and private pension plans, which afford alternative sources of income to increasing proportions of older workers and which often specify the age for mandatory retirement.

The striking discrepancy in rates of labour-force participation between older and younger men (most younger men are in the labour force) has differential consequences for the age strata, affecting not only their respective involvement in the goal-directed activities of the society but also their financial status and the time available for their leisure pursuits.

Decline in income

In regard to financial status, people over 60 and even those in their 50s, have markedly lower median incomes than people in their middle years, even after various adjustments are made (for example, for family size). The substantial data available for the United States, for instance, show that, over time, there have been absolute increases in the income and the purchasing power of older people, with a declining share from earnings and a rising share from social security and other retirement benefits (often paid in fixed dollars); however, older people continue to lag behind the young.

To be sure, older people on the average, with their years of asset accumulation mainly behind them, own more and owe less than younger people. Among elderly married couples, substantial proportions (the majority in the United States) own their homes. But assets, which are very unequally distributed, tend to be correlated with income, so that those families with the lowest incomes are also the least likely to have any assets.

What kind of life does the older person's income support? Here again data available for the United States show that he tends to live closely within the limits of his income, spending less than younger age groups for most budgetary items, including leisure items, but spending a larger share for health expenditures when his government has no comprehensive program of medical insurance. His spending is restricted both by age-associated factors (such as reductions in family size after the children leave home) and by his comparatively low income; the income of many older people, when judged by arbitrary standards set by economists, is not adequate to meet their needs.

Education. The age strata of the society are distinguished from one another not only in labour-force participation and in income but also in educational background.

Lower levels of education

Older people have less formal education and less recent education than younger people. Among persons 75 and over in the United States, where the century-long spread of formal schooling accentuates the age differences, more than 70 percent have had only eight years of schooling or less, compared with less than 20 percent of those aged 25 to 29. Moreover, although the average educational attainment of older people is rising, the lag of old behind young persists.

These striking age differences in education have widespread ramifications throughout society. Since comparatively few older people are well educated, few of them possess those characteristics, typically valued in Western culture, that are associated with high education. Thus, the less educated majority of older people are less likely than the few who are well educated to remain active; they are more likely to retire. They are also less likely to belong to voluntary associations or to read or to want to learn more. They have lower incomes. They are often less happy and generally less optimistic about the future, but at the same time, they are less introspective and less ready to doubt their own adequacy as spouses or parents. They are more negative in their view of death and think about it more, although fewer of them are disposed to plan for it.

Age categories in the population differ not only in level but also in recency of education. Thus, the content of knowledge and attitudes acquired vary by cohort. The information imparted to doctors or engineers during their training, for instance, differentiates sharply between the backgrounds of old and young. Yet within each cohort, there is a certain age homogeneity in values and beliefs (about what is good, beautiful, or true) among individuals who were educated at the same point in history.

It is interesting to note that one reason education can make such enormous differences among the age strata in modern society is that formal schooling, rather than being spread over the life course, is almost completely concentrated in the early years of life. A person enters adulthood with a fixed educational background that functions sociologically in almost the same way as such inborn characteristics as skin colour or sex. Adult education (a channel that might close part of the age gap) is developing, but it still reaches only small proportions of people in their middle or later years.

Residential setting. Historic changes in population distribution, in community planning, and in the designing and marketing of houses affect the physical and social environment of the older person; influencing the range of his human contacts, his day-to-day activities, and the community services and facilities available to him. In the modernized nations persons who are now old have, to a great extent, been swept along in the massive movements of the population from agricultural to industrial areas, from farms to cities or suburbs. Many of the elderly, however, have been left behind, as age-mates die off and younger generations move away. Moreover, the houses in which most older people live were built in an earlier era. As their houses have aged, the inhabitants themselves have moved along the life course toward increased leisure in retirement, diminished families, and reduced incomes.

In general, however, most older people do not avail themselves of the freedom from occupational and family commitments to move away from their former homes, a fact that gives some indication of the meaning of the older person's home and the factors mediating his relation to his environment. Rates of moving are generally highest among people in the youngest age groups. Moreover, when older people do move they are more likely to move within the confines of their immediate locale than to change their community setting. It thus appears that, basically, the individual's ties to his place of residence become stronger as he grows older. And older people who own their homes, or those who have strong social connections with the neighbourhood, are the least willing to move.

Family status. The social position of the aged is importantly tied to the structure of the family, which has been gradually changing in Western society over the past century. Although this change is usually described as the isolation of the nuclear family (*i.e.*, parents and their children), perhaps it may be more accurately described as the subdivision into two or even three generations of distinct nuclear families: the young couple with their dependent children, the middle-aged parents, and the aged generation of grandparents.

Independence of the aged

As life expectancy has risen, husbands and wives have become increasingly likely to survive together into old age; and widowhood, even if not curtailed in length, has been postponed. Today, in such countries as the United States, Great Britain, or Denmark, in the 65-and-over age range, most men (about two-thirds), but a minority of women (about one-third), are living with their spouses. Moreover, most of those 65 and over, the married and even the widowed, maintain independent households. In the United States this way of living has been increasing, as decreasing proportions of older people live in the same household with their children. Even when households are shared today, the older person is, more often than not, himself the head of the house or the host, not a guest or subordinate in the home of his child. Living entirely alone is a frequent pattern among older people. This is especially true among women, over one-fourth of whom (mainly widows) live alone. Moreover, the norms appear to favour separate households for older people, except for those who are ill.

Thus a new type of nuclear family, the independent, older family (often a single-person "family") is developing, with little-explored consequences for the older person or for the society as a whole.

THE ORGANIC BASE

The social aspects of old age cannot be understood without reference to the physiological processes that distin-

guish old people from the young. The age structure of the society is fundamentally affected by biology, not only because there is the succession of births and deaths but also because individuals at different stages of life show important differences in both physical structure and function.

With age, there are declines in the number and quality of vital cells and a decreased ability of the organism to adapt to changes in the environment. Associated with such age-related differences is the fact that the health of older people is generally poorer than that of the young. While older people have fewer acute illnesses, they are more subject to chronic conditions: such impairments as failing vision or hearing or such ailments as rheumatism and arthritis, heart disease, and high blood pressure. In the United States, for example, some four out of five persons 65 and over have at least one chronic condition. Thus there are increases by age in physician visits, in the number and duration of hospital stays, and in the number of days spent in restricted activity or in bed. Nevertheless, only a small proportion of old people are so severely handicapped that they cannot carry on their major activity, whether it be a job outside the home or housework.

Behavioral changes

Physical changes associated with age are related not only to health but also to behaviour, though the causal linkages seem far from clear. When compared with younger people, older people are more likely to show deficits in sensory and perceptual skills; in muscular strength; in the ability to react quickly; in complex sensorimotor coordination (such as that required in driving a car); and, most important of all, in certain forms of memory, in learning and in various aspects of intellectual functioning. There are also age-related decrements for both men and women in sexual activity, though sexual capacity persists at least into the 70s or 80s.

Of course, not all the differences between old and young are necessarily reflective of a biological, or even a social, process of aging. It would be fallacious, for example, to attribute entirely to old age the comparatively poor performance of older people on intelligence tests. In cross-section studies (conducted among several age groups at a single time), intelligence appears to reach a peak among people in their late teens or early 20s and then to decline with age in the older strata. But to infer that intelligence falls off so early in the life course of particular individuals because of age-associated changes (as in general health) is open to serious question, as a few longitudinal studies (which trace individuals over the life course) have begun to suggest. In this instance, the interpretative difficulty seems associated with social change, with the long-term upward trend in education. Intelligence test performance is highly correlated with educational level, and education sets older cohorts sharply below younger cohorts.

Unfortunately, there is little information about trends over time in health, physical functioning, or the associated changes in behaviour. For the population as a whole, to be sure, the rising proportions of older people point to the likelihood of increasing prevalence of the chronic conditions associated with senescence; indeed, there have been increases over the past century in death rates resulting from such illnesses of later life as heart disease or cancer. But it is uncertain whether older individuals today are healthier than older individuals in the past. Medical advance may have improved their physical condition, or it may, by interfering with the principle of survival of the fittest, have yielded a larger but less hardy population.

Whatever the trends may be, the inferior physical state of the older segments of society, in contrast to the younger segments, has both societal antecedents and societal consequences. Health is conditioned by such social factors as standard of living, education, and advances in medicine and public health. In turn, the physical state of older people sets limits to their social adjustment and to the contributions that they can make to society.

THE AGED INDIVIDUAL

How, then, does the older individual respond to a life situation in which he tends to be comparatively disadvantaged both in health and physical fitness and as a member of society? Does he in fact conform to the tragic stereotype of the old person as destitute, ill, facing irreparable losses, no longer integrated into society, and no longer subject to society's controls and sanctions? Is it true, as often supposed, that his feelings are fully reflective of the relatively deprived status of the aged within the society and that he is characterized by a loss of self-esteem, a deprecatory view of his low education, a sense of dejection and despair over his losses, and anxiety about his health, finances, and death? A glance at the available clues shows a picture that is, at least in certain respects, at sharp variance with such a stereotype.

Personality. Old people resemble the rest of the population in many of their personality characteristics, despite certain distinctive types of emotional expressions and modes of adaptation. Thus the aged, in comparison with the young, appear to be more rigid, passive, and introverted, more restrained and cautious, and less oriented to achievement. Age differences in personality and the individual's disposition to organize his behaviour in particular ways cannot be separated, however, from his biological condition and the social roles that he plays. And there is little evidence to indicate how many of the unique aspects of the personality of the older individual are a result of accumulated experience, relative lack of education, inexperience with personality tests, weakened sensory and motor skills, or limited opportunities to test out his ideas on other people.

Degree of neurosis, psychosis, and deviance

Disturbances of thought, affect, or behaviour afflict, in varying degrees, somewhat fewer than 10 percent of the people 65 and over in several localities studied in Europe and the United States. The milder mental disorders (neuroses) do not appear to increase by age. But psychoses (and various psychiatric symptoms experienced as bodily illness) are apparently more prevalent among older than among younger people, both because in later life there are certain conditions connected with the degeneration of the central nervous system and because the older the individual is, the greater his chances are of having any chronic or irreversible disease that does not result in early mortality. Although the burden of mentally ill older people upon families and treatment facilities has undoubtedly been rising over time, the rise is due partly, perhaps entirely, to changes in longevity—to the increasing numbers of people who survive into old age.

There are differences by age in various types of deviant behaviour. For example, crime rates, as observed in disparate times and places, show a consistent decline as age advances beyond adolescence. Moreover, life-course analyses demonstrate a curtailment in criminal behaviour with age even among those who have previously engaged in such behaviour. Thus, in the comparatively rare instances when older people commit criminal acts, these are ordinarily relatively harmless offenses rather than aggressive outbursts against society or acts of violence against other persons. In contrast, suicide is most pronounced in the older age categories, a pattern observed in many countries in both the 19th and the 20th centuries.

Social roles. The older the individual is (beyond the middle years), the fewer social roles he plays on the average (as he retires from work, his children leave home, and his age peers die off) and the fewer and less varied are his contacts with other people and with the environment generally. Individuals are not all alike, however, and social activity is greatest among those aged who are in good health or who come from higher rather than lower socioeconomic backgrounds. Moreover, much depends upon the particular exigencies that permit some older individuals, but not others, to retain their major familial and occupational roles.

Work and retirement. The aging individual faces the culmination of his occupational role and a final period of retirement. For older men today, the modal role is retirement, not work. If they grew up in an era when retirement was not widely institutionalized, they now confront a world increasingly populated by age peers who are no longer full-time members of the labour force.

Those men and women who do remain in the labour

Productivity and commitment

force during their later years are not making generally inferior contributions, despite their frequently poorer performance under laboratory conditions. Studies under actual working conditions show older workers performing as well as younger workers, if not better, by most, though not all, measures. (Of course, such age patterns of actual performance are traceable in part to labour-market conditions and to selective processes that allow the retention of the more competent older workers in their jobs.) That there is no necessary decline in creative productivity with aging is evidenced through analyses of the published biographies of contributors to numerous scholarly, scientific, and artistic fields (though widely quoted earlier studies had propounded somewhat misleading contrary findings).

In addition to maintaining their productivity, many older people also take a generally positive view of their occupational role. As compared with the young, older people still in the labour force tend to be more strongly committed to the kind of work they do, to adapt better to the job, and to express greater satisfaction with it. Moreover, despite their inferior education and the lowered aspirations attendant upon age, older workers do not fall very far short of the young in their sense of occupational adequacy or in the emphasis they place on the intrinsic satisfactions of the work itself.

Once older people have retired, they seem on the whole to accept the new role of retired person, although many say that they regret losing their work associates, the feeling of usefulness, the work itself, and the associated earnings. Retirement per se does not appear to have a deleterious effect on health nor does it affect social participation in the other roles of the retired person. To be sure, adjustment to life is generally poorer among retired old people than among age peers who are still working, but this results in part from the selective process whereby the healthier and more advantaged oldsters tend to continue working.

Political roles. Older people, even those with little education, are significantly involved in the political system. They are, for example, at least as likely as younger people to vote. They are better represented among the political elite and in leadership positions in many types of decision making. They are more disposed to keep up with the news and with public affairs through the mass media.

In general, today's generation of older people are more conservative than younger people in their political attitudes. There are important exceptions, however, as on issues affecting their own economic self-interest. In the United States, for example, the old have often tended to support extensions of social security or government health insurance but have frequently voted against school bonds. Yet there is little indication whether future generations of old people, subjected to shifting political climates and to rising levels of education, will perhaps adopt a less conservative stance and a generally greater flexibility of attitude.

Religious roles. Religion assumes greater importance among older than younger segments of the society, although research (which is less complete than research on the polity) does not show whether religious values become intensified with aging or whether the older cohorts reflect a stronger religious emphasis in their early training. The aged are more caught up than the young in such personal observances as prayer and reading scriptures, and elderly members of religious organizations benefit from a variety of special programs, such as home visits and nursing homes. Church attendance, however, declines in the oldest years.

Voluntary associations. In addition to religious and political affiliations, modern pluralistic societies offer a variety of voluntary associations as potential links between older people and the larger community—associations such as clubs, lodges, auxiliaries, and the many other formal organizations to which people belong part-time and without pay. To be sure, membership in such associations is less widespread among older people generally than membership in a church or voting in a national election; and average membership rates show a drop in later life after a rise in the middle years, though individuals vary greatly around this average—depending upon their education and income. The special clubs (such as Golden Age Clubs or Senior Centres) established in many Western countries to provide older people with recreational, educational, health, or welfare services seem to attract comparatively few participants.

Family roles. Old people play a variety of family roles, and they seem no less likely than younger people to feel satisfied and adequate both as spouse and as parent. For those older people who have a living spouse, the marital relationship is of central importance, and older couples share many joint activities; but many of the aged, especially women, must adjust to widowhood.

Ties with children and relatives

Most old people play the enduring role of parent. While their children in turn form families of their own, new linkages develop with children-in-law; and grandchildren, even great-grandchildren, proliferate. Living separately from children does not mean complete isolation from them. Over 80 percent of older people who have living children live less than an hour away from the nearest child, and a similar proportion see one of their children at least every week. Bonds with siblings or other relatives are often maintained, and a few old people continue in the role of offspring to parents who have survived into very old age.

The ties of the elderly to their relatives and to their children in particular are by no means limited to visiting. There are also widespread exchanges of material support of various amounts and kinds, ranging from financial contributions and care in illness to baby-sitting and help with housework and home repairs. Contrary to the often-held theory of a one-way flow of contributions to older people, the flow of support between aged parents and their adult offspring appears to be two-directional, from parent to child or from child to parent as need and opportunity dictate.

Although close parent–child ties seem important to both older people and their offspring, the offspring feel closer to their new family of procreation than to their family of orientation. Thus the norms of intergenerational independence seem to operate to reduce the emotional dependence of the elderly upon their adult children.

Friends and neighbours. Friends and neighbours play an important part in the lives of many older people, often providing help and services as well as informal contact with the world outside the home. They are, however, generally less important to the aged individual than children and other relatives, serving more as a complement than as a substitute for kinship association. Friendships and neighbourly relations, which tend to be maintained well into later life, are most widespread among older people of comparatively high socioeconomic status and among those who have resided for a long time in the same neighbourhood. Older people tend, though by no means exclusively, to have friends who are similar to themselves in status characteristics (notably age) that reflect common experiences or values.

Leisure. Maturity yields new increments of leisure time, as commitments relax to parental and later to occupational roles. Many and varied pursuits fill these hours of leisure. Visiting with friends and relatives is an important activity. Much time is devoted, for purposes of information as well as entertainment, to watching television and reading the newspaper. Only small minorities, however, spend their leisure in crafts, hobbies, or intellectual and artistic activities; and vacations and outings beyond the home, particularly foreign travel, are restricted to the few. Indeed, some older people, especially the very old and the disadvantaged, often have difficulty finding anything at all to do.

When the leisure of old and young is compared, varied age patterns emerge for different sets of activities. In the United States, for example, certain types of activities (gardening or walking) seem characteristic of older people. Other kinds of activities (swimming, museum attendance, or playing a musical instrument) are widespread among the young but largely unfamiliar to today's older generation—a possible portent of social changes to come. Thus a "mass" culture may exist at the younger age levels that will permeate society only as these cohorts (generations) reach maturity.

Attitudes and satisfactions. Age is related not only to psychological characteristics and to performance in roles but also to the individual's estimate of himself, his attitudes toward the environment, and his feelings of gratification or deprivation.

The self. Underlying the several dimensions of his personality, a clear sense of his own identity is experienced by the typical older person. Although conceptions of old age are often assumed to be largely pejorative, the modal older person appears to evaluate quite positively such aspects of the self as his moral virtues or the adequacy of his role performances. At the same time, he minimizes in his self-image many of those aspects that are negative, such as his failing health, his personal appearance, his relative lack of education, or the fact that he is old.

Relatively positive self-evaluations

If many older people, even though objectively disadvantaged, fail to take deprecatory views of themselves, the reasons are not entirely clear. By virtue of their long years of life, these people have acquired experience, possible wisdom, and the opportunity to come to terms with themselves. Perhaps they judge themselves less with reference to younger people or to themselves at an earlier age than with reference to others who are at least as old as they.

Views of life. Older people differ sharply from the young in many of their opinions, feelings, and dispositions toward such central aspects of life as health, personal problems, or death. To a greater degree than younger people, they take aches and pains for granted, put relatively little faith in medical science, and feel they understand their own health best—though the frequency of thinking about health, visiting doctors, and taking medicines rises with age. Certain elements in their conception of physical illness appear also in their outlook on mental illness: the sense of inevitability and incurability and the stricture that the individual should look after himself. They are likely to define their personal problems generally as unchangeable and to stress the individual's responsibility, although they seem ready enough to seek assistance in those instances when assistance is deemed appropriate. Death is typically confronted openly as an inexorable fact. Most older people report that they think about death, are willing to discuss it, and have made preparation for it. Few stress otherworldly aspects, either hope of heaven or fear of a last judgment. Indeed, few show marked fear of any sort, expressing the view rather that death is more tragic for the survivor.

Fatalism combined with stress on individual responsibility

Certain recurrent themes are discernible beneath the myriad specific differences, as old and young define and assess their life situations differently and hence are inclined toward differing courses of action. First, older people tend to have less "sense of mastery" over the conditions of their lives than do younger people, considering the world potentially less changeable. Second, older people tend (paradoxically) to stress the "responsibility of the individual" for his own destiny; whereas younger people are more likely to stress environmental influences. Third, in line with such differences in definitions of the situation, old and young tend to favour different "types of approaches" to life situations. Thus older people, defining many of life's ills as inevitable, may be more disposed to seek palliative rather than corrective or preventive treatment. Or again, if older people attribute to the individual the responsibility for cause and cure of his problems, they may be more disposed to lay exhortation or blame upon him than to attack his social or his somatic conditions. Fourth, when older people do become committed to a particular approach, they may be typically as willing and able as the young to *implement* it. Thus much of the often noted lower activity levels of the aged may perhaps be explained through their differing definitions and evaluations of the situation rather than through any age-related tendency toward generalized passivity.

Such differences in the life attitudes of old and young are probably traceable in part to the differences in educational level and to the impact of the aging process itself. In addition, the differences are undoubtedly reflective of the long-term social and cultural change in which each new cohort (generation) has been socialized to new understandings, norms, and habit patterns.

Life satisfaction. In some general sense, satisfaction with life (happiness, morale, adjustment) seems to diminish with age. This decline, already apparent in early adulthood, is not peculiar to senescence; but it becomes intensified (as both longitudinal and cross-section studies suggest) with age-related deterioration in health, loss of key roles, or reduction of activity. Thus age appears to be associated with a general diminution of the opportunities for happiness.

Nevertheless, when research is focussed, not upon overall satisfaction but upon more specific reactions to particular areas of life, older people appear to differ from younger people in the kinds of gratifications and anxieties experienced and regarded as salient. The typical older person is as likely as the younger person to seem content with his occupational and familial roles and to encounter no greater problems in them. And despite the objective difficulties confronting him, he appears even less likely to worry about his health, for example, or his finances. Thus, the older a person is, the more nearly he seems to have come to terms with many of the specific conditions of his life.

Many studies of life satisfaction have challenged the provocative notion of "disengagement," conceived by some researchers as a mutually satisfying process of withdrawal between the individual and society in preparation for the incapacitating diseases of old age and eventual death. In the main, the evidence lends little support to the theory, although continuing activity is clearly not a *necessary* condition of satisfaction for all older people. For the majority, however, the sense of well-being is associated with high rates of interaction and contact with the environment—just as adjustment is associated also with good health, with favourable occupational and familial circumstances, and with a high sense of self-esteem.

PROGRAMS AND POLICIES

The modern stereotypical view of old age as a situation of utter and inevitable disadvantage has often overemphasized the problems and weaknesses of the elderly without paying proper attention to their strengths and their potential contributions to society. In fact, there are many areas of independence and an absence of serious physical disability among the majority of today's aged. Moreover, the relative deprivations of the older population as a whole are offset by numerous exceptions, since there are segments of this population who enjoy adequate education or income or who exhibit high levels of adjustment and interaction with their fellows. Such exceptions serve both to deny the intractability of the current problems of the elderly and to suggest areas in which ameliorative effort is likely to produce needed solutions.

Nevertheless, the deprivations remain. The central fact is that a great many older people nowadays are not filling socially valued roles in the economically developed countries. This fact can be punishing to the individual, and for the society it can be a costly waste of human resources. Concern for this fact—a concern that has become intensified with the multiplying proportions of old people in the population and with the recently emerging power of the medical profession to prolong the lives of many dying patients—has led to the formulation of various policies and programs.

Types of programs. The problems of the aged, like those of the poor or the disabled, have been handled through a wide variety of institutions and arrangements at different times and in different places. In most agrarian societies, the aged have found security in a familistic social organization in which food is shared, oldsters often marry younger mates, or progenitors hold sway over descendants. The family has been bulwarked by such other structures as the church or the guilds in medieval Europe, by private philanthropy and the giving of alms as broadly established in many cultures, or by the building of institutions to house the aged (which existed in Europe as early as the 3rd and 4th centuries).

Comparisons between agrarian and industrial societies

Modern industrial societies, with their complex structures and their large proportions of older people, have developed many special forms of adaptation.

The family is still of central importance to the older person's security (statutory responsibility of adult children for their needy parents continues to exist in a number of countries); today's aged couple, however, with their increased years of joint survival and their tendency to live apart from their children, are likely to depend primarily upon each other, turning to children or relatives only under conditions of special need.

Following the breakup of medieval institutions in Europe, the aged are variously assisted by state intervention, private and corporate philanthropy, religious and fraternal organizations, or labour unions; and private enterprise is often engaged in building housing and residential communities for the aged or in arranging for the delivery of health care or other services to older people. In the United States, for instance, many older people have secured financial protection above the government-supported minimum through private pensions, voluntary health insurance, or life insurance benefits to the bereaved spouse. Only small fractions of those 65 and over (less than 5 percent in the United States and less than 4 percent in England and Wales) are residents of old age homes and other institutions.

Public programs and governmental acts to meet the needs of the aged, from the bread and circuses of ancient Rome to the modern welfare state, have had varying histories in different countries throughout the world—depending upon particular social philosophies; upon the crises created by war, plague, or famine; or upon the maladjustments and dislocations created by basic changes in economic, religious, military, and other social institutions. In England, for example, a series of 16th-century enactments, culminating in the Poor Law of 1601, recognized a degree of public responsibility for the aged, along with the sick and the poor, although the administration was left to the local parish. In subsequent years, various legislative and administrative modifications were made in the law, including the reversion (under the Poor Law Amendment Act of 1834) to an austere system of workhouses in which a uniform, stringent discipline was imposed upon the aged, the sick, and the able-bodied poor. Not until 1925 did England introduce social insurance for the aged (followed in the 1940s by the legislation underpinning the full welfare state system), in which contributions are regulated by law and benefits disbursed without a means test. Meanwhile in Germany Bismarck instituted as early as the 1880s a full program of social legislation, including old-age pensions. Austria, the Scandinavian countries, the Low Countries, France, and Italy followed the German model; and by the mid-1960s, well over 100 countries had social security programs. The United States, slow to yield its philosophy of individualism, long clung to vestiges of Elizabethan-type poor laws, not enacting an old-age pension program until 1935, when, as a consequence of the Depression, the Social Security Act was signed. The basic plan has been supplemented by other federal, state, and municipal programs, including Medicare, Medicaid, the Older Americans Act, income-tax exemptions, subsidized housing and liberal financing of mortgages, and old-age assistance as the last line of income defense. These and many other aids and services to older people have alleviated some of the difficulties and set a floor to older people's income, but they have not brought this income up to the level of younger people.

Poor laws and social insurance

Development of the professions. The extension of programs to ameliorate the condition of older people has enlarged the demands upon the practicing professions. The long-term intrusion of old age and its concomitants into the established professions, though often unnoticed, has increased the likelihood that the physician or nurse must treat the chronic ailments and disabilities associated with senescence, that the architect must provide housing for older couples or widows living alone, that the social worker or minister must deal with persons no longer absorbed in occupations or parenthood, and that the lawyer must advise on the estates of persons who can expect to live well beyond the age of retirement.

Professional aid to the aged

In addition to increasing in numbers within the clientele of the professions, older people have created demands that often require new kinds of services or shifts in professional emphasis. Thus educators may be expected to provide continued education or retraining to adults, mass communicators to forge links between old people and society, or public-health experts to expand the delivery of health care to this segment of the community. Architects and city planners must arrange for nearby facilities (such as stores, buses, medical services, churches, or recreational activities) on which older people in failing health and vigour are especially dependent; and, since the elderly seem loath to move away from familiar surroundings, homes may be designed to allow expansion and contraction, for example, or ranges of housing suitable for all stages of the life course may be planned within a single neighborhood.

Furthermore, certain developing needs of older individuals, notably their need for help in planning for retirement and for income maintenance, point to gaps in the structures of existing professions. To be sure, advice is proffered by a number of specialists on selected aspects of the older worker's participation in the labour force or on financial management, and thousands of recreation workers offer leisure-time programs for the aged. Yet there is typically no single source of unified advice for the person who is considering retirement, who is seeking to optimize his use of leisure in retirement, or who is attempting to understand the intricacies of social security, pension benefits, insurance annuities, real property, and other assets. This gap becomes increasingly apparent as the mass of today's workers confront the necessity of planning for retirement, suggesting possible future changes in the structure of existing professions to meet developing needs.

The basis in research. The planning of programs and policies and the training of professionals to implement them rests increasingly upon research on old age and on the related processes of individual aging and social change. Early forerunners of research leading to social policy include the monographs of Frédéric Le Play (*Les Ouvriers européens*, 1855), which were used in propaganda for state reform in France entailing patriarchal control in the family; or the detailed house-to-house surveys of Charles Booth (*Labour and Life of the People*, 1889), which led eventually to the Old Age Pensions Act of 1908 in Britain. More general studies of age as an explanatory factor in social behaviour were conducted as early as the mid-19th century, when Adolphe Quételet made quantitative analyses by age categories of crime rates and suicide rates (*Du système social et des lois qui le régissent*, 1848).

Recent social research on old age has benefitted from the advances of the social sciences generally, from the compilation of basic demographic data for many countries, and from the proliferation of empirical studies in the Western world, particularly in the United States. Research has been stimulated by interdisciplinary groups formed to develop gerontology as a special field (in particular, the International Association of Gerontology and its branches in many countries). The focus of such groups, initially centred on biology and clinical medicine, was extended in the 1950s to include psychology and the social sciences as well. The accumulation of knowledge has been fostered in various ways by universities, by foundations, and by governmental and international agencies and has begun to yield not only more rational solutions to practical problems but also a more complete understanding of old age in terms of the dynamic processes of aging and the flow of cohorts through a changing society.

FUTURE STATUS OF THE AGED

Because the state of being old is influenced by the dual processes of aging and cohort succession, the prospects for the people who will become old in the next decades depend upon the social changes, including interventions by policy makers and professional groups, to which each new cohort will respond.

The process of becoming old does not always occur in the same way but is subject to change as when, for example, certain of the current ills of older people are prevented through individual treatment or through manipulation

of the relevant environment. Little is known about the extent of the individual's capacity to adapt or about the conditions for fulfillment of this potential. Yet intervention at earlier stages of the life course might well result eventually in continuing education and development over the lifetime, in anticipatory learning of postretirement skills or of new careers, in lifelong accumulation of savings and estate planning, in preparation for changing housing needs, and even in prevention of certain of the apparent "chronic" ailments of later life.

Prevention of future deprivations of persons not yet arrived at old age is still not only possible because of changes in the way people grow old but also because of differences in the kinds of people who compose the successive cohorts and in the kinds of early training and life attitudes to which they are exposed. It is already clear, for example, that future old people, as compared with the old people of today, will have higher education and higher levels of lifetime earnings and that, since far fewer of them will have been born on a farm, they will be more readily at home in an urbanized society. And if there is a continuation in industrial countries of the long-term trend toward declining labour-force participation and earlier retirement of the older population, this trend can alter drastically the organization of work and leisure, the supporting values, and the total role complexes of individuals at all ages.

Of course, not all the long-term effects of social change will be beneficial. It is by no means clear that future cohorts of old people will have fewer problems or be better off in every respect. For example, the income or the education of tomorrow's aged, *relative* to other age groups, may not improve or may even deteriorate. The social and personal costs may be immeasurably high if the lives of large numbers of helpless and hopeless older people are medically prolonged. And the level of public health, though advancing on many fronts, is subject to possible deleterious effects from such aspects of the urban environment as air or water pollution.

However imprecisely the implications can be foreseen, the forces of social change, whether through deliberate action or as an indirect consequence of existing trends, are not only constantly intervening in the aging process but are also bringing new influences to bear on the situation and on the characteristics of persons who are old.

BIBLIOGRAPHY. Extensive empirical and theoretical work has been collected in the following three handbooks, comprising essays on varied aspects of gerontology: J.E. BIRREN (ed.), *Handbook of Aging and the Individual: Psychological and Biological Aspects* (1959); C. TIBBITTS (ed.), *Handbook of Social Gerontology: Societal Aspects of Aging* (1960); and E.W. BURGESS (ed.), *Aging in Western Societies* (1960). More recent information on several of these topics may be found in a special issue of the *International Social Science Journal*, entitled "Old Age," vol. 15, no. 3 (1963). Findings from some 3,000 social science studies concerning people in their middle and later years have been condensed and organized for ready reference in M.W. RILEY and A. FONER *et al.*, *Aging and Society*, vol. 1, *An Inventory of Research Findings* (1968). The findings from this inventory are interpreted for use in those professions concerned with older people in M.W. RILEY, J.W. RILEY, JR., and M.E. JOHNSON (eds.), *Aging and Society*, vol. 2, *Aging and the Professions* (1969). Systematic research findings for selected times and places appear in the analysis of ethnographic reports in L.W. SIMMONS, *The Role of the Aged in Primitive Society* (1945); H.D. SHELDON, *Older Population of the United States* (1958); and in the parallel surveys reported for Great Britain, Denmark, and the U.S. in E. SHANAS *et al.*, *Old People in Three Industrial Societies* (1968).

(M.W.R./A.F.)

Old Catholic Churches

Churches correctly described as Old Catholic include a number of well-organized local churches that claim to have maintained all the essentials of Christian doctrine and to have kept the historic succession of the episcopate but are not in communion with the see of Rome. The term is used less correctly of certain churches that have come into existence as the result of nationalist movements and in some cases stand outside the main stream of Christianity. It is also used incorrectly of a number of small bodies brought into existence by the enterprise of private individuals, who have in one way or another secured for themselves episcopal consecration and in certain cases have conferred the title Old Catholic on the church bodies that they themselves have brought into being.

NATURE AND SIGNIFICANCE

Different views are held as to the origin of the local episcopate in Christianity. It is clear, however, that from the middle of the 2nd century onward the term *episkopos* ("bishop") was used exclusively of ministers of the church who had been chosen to be chief pastors of cities and of the regions surrounding them, consecrated by their fellow provincial bishops and recognized as being in communion with the entire Christian episcopate throughout the world. But from an early date there were *episcopi vagantes*, or wandering bishops—men who had been deprived of their sees but still claimed the right to minister, bishops expelled from their sees by Muslim occupation (from the 7th century onward) or Protestant occupation (from the 16th century onward), bishops of sees that had ceased to exist, or men who had secured episcopal consecration otherwise than through the regular process of public election and consecration. One of the best known bishops of this type in the Reformation period was Pietro Paolo Vergerio (1497/98–1565), who accepted the principles of the Reformation and served for a time as a pastor in the Grisons (Switzerland) but does not seem to have exercised any further episcopal functions.

Importance of episcopal consecrations

Since the consecration of Jules Ferrette at Homs (Syria) in 1866 by bishop Mar Bedros of the Syrian Jacobite Church, the number of *episcopi vagantes* has enormously increased. It is to be noted that only one such consecration, that of Arnold Harris Mathew (1852–1919)—who was successively an Anglican, Roman Catholic, Old Catholic, and leader of his own series of churches—has been carried out by Old Catholic (Jansenist, or those who followed the mystical and quietist teachings of Bishop Cornelius Otto Jansen, 1585–1638) bishops, and that within three years of the consecration these bishops had withdrawn their recognition of Mathew as a bishop.

Various views have been held as to the validity of such consecrations and of the consecrations and ordinations carried out by bishops of this type. The Eastern Orthodox churches have almost universally condemned them, on the ground that in no case did the consecrators have the *kyriotes* ("authority") from the church to act. Roman Catholic authorities have inclined to the view that such consecrations are valid but irregular. The Anglican Communion, on the ground that a bishop must be consecrated in, by, and for a church, has consistently refused any kind of recognition to such bishops and to the bodies founded by them.

HISTORY

Old Catholic churches. *Origins.* A schism, small in numbers but of great historical importance, was precipitated by the definition and promulgation of the doctrine of the infallibility of the pope at the first Roman Catholic Vatican Council in 1870. That doctrine declared that in matters of faith and morals the pope was infallible. In the 19th century many of the most learned Roman Catholics rejected the doctrine, on the ground that Christ had granted the blessing of infallibility to the church at large and not to any one member of it, however eminent. Others, such as Cardinal J.H. Newman (1801–90, a former Anglican), believed that the declaration of any new dogma would be inopportune. Yet a third group, which Newman described as an "insolent and aggressive faction," but which included the English archbishop, later cardinal, H.E. Manning (1808–92), eagerly desired the declaration. Among the dogma's strongest opponents was the German church historian Ignaz von Döllinger (1799–1890), one of the most learned men of his time. In a long series of works Döllinger had elucidated the history of the church; now his voice was raised in earnest protest against a doctrine that he regard-

Importance of I. von Döllinger

ed as having been unknown in the great ages of the church. After the dogma had been promulgated and the council prorogued, almost all the bishops of the minority group made their submission to Rome. Those who could not repudiate their deeply held convictions turned to Döllinger, who never submitted and died unreconciled in 1890. He never formally joined the Old Catholic movement, as the group of dissidents came to be called, but remained in close touch with the leaders of the movement and was willing to advise. In view of their need for a valid episcopal succession, which they regarded as essential to the being of a church, he recommended them to turn to the Jansenist Church of Holland. Thus, the Old Catholic churches of the continent of Europe came into being.

The Jansenist Church of Holland. On September 8, 1713, Pope Clement XI (1649–1721) in his constitution *Unigenitus* ("Only begotten") condemned what were supposed to be the tenets of Jansenism, which emphasized the doctrine of God's grace. In Holland, the teachings of Cornelius Otto Jansen, bishop of Ypres, were still widely revered, and the condemnation of Jansenist doctrine was not accepted. In 1723 the chapter of the ancient see of Utrecht (founded by St. Willibrord, 658–739) elected as its bishop Cornelius Steenhoven, who secured episcopal consecration at the hands of the missionary bishop of Babylon, Dominique-Marie Varlet, in 1732; this consecration has never been condemned by Rome and is generally accepted as having been valid though irregular. This small but highly cultured and wealthy church has managed to maintain itself in existence over two and a half centuries and can boast of an unbroken episcopal succession. Thus, much sympathy was expressed in Holland for the groups that after 1870 had considered it necessary to separate themselves from Rome in protest against what they believed to be new and unacceptable claims to authority. It was from the Old Catholic bishop of Deventer, Holland (Hermann Heykamp), that Germany in 1873 received its first Old Catholic bishop (J.H. Reinkens), to be followed by Switzerland in 1876 (E. Herzog) and Austria in 1925 (A. Schindelaar)—the Austrian Church had been organized in the 1880s but had not been able to secure the consecration of a bishop.

The Old Catholic churches today. Since 1889 the Old Catholic churches of the world—including the Polish National Catholic churches in Poland and in the United States, but not including the schismatic Mariavite Church in Poland, which in 1924 was disowned by the other Old Catholic churches on the grounds of alleged unorthodoxy in doctrine—have been united in the Union of Utrecht, of which the archbishop of Utrecht is ex officio president. Following the lead given by Döllinger, these churches have always maintained that they have no new doctrine to preach; they exist in order to call the churches back to what they believe to be an older and sounder tradition of Catholicism, and to serve as an instrument in the hand of God for the promotion of the unity of all Christians. With this special sense of vocation to work for Christian unity, the Old Catholic movement has from the start been deeply interested in relationships with other Christian bodies. The Old Catholic conferences of 1874 and 1875, presided over by Döllinger, were attended by Anglicans, Orthodox, Lutherans, and representatives of other Christian traditions. The Anglican Lambeth Conference of 1878 pronounced favourably on the Old Catholic churches, recognizing the group as a genuine reforming movement and not as schismatic.

The Union of Utrecht

National church movements. In the formation of the Old Catholic churches, national feelings played some part—especially in Switzerland—but were not dominant. In a number of other similar movements, national feelings and a striving after national independence played a much stronger role. Such feelings could not but be strengthened by the increasing centralization that the Roman Catholic Church has practiced since the Council of Trent (1545–63).

In Czechoslovakia, at the end of World War I (1918), the Los von Rom ("Away from Rome") movement attracted much interest and support and within a brief period claimed 800,000 members for the newly formed Czech National Church. Within a short time, however, this church adopted Unitarian (*i.e.*, non-Trinitarian) principles, rapidly declined in numbers, and ceased to exercise any powerful influence on church life in Czechoslovakia.

In the United States, Polish Roman Catholics became increasingly dissatisfied with the manner by which the Roman Catholic hierarchy, which was predominantly Irish or German, administered the affairs of the church; and, in the opinion of the Poles, that hierarchy neglected the interests of groups other than those of their Irish or German supporters. Some of their leaders entered into correspondence with Old Catholics in Europe; and in November 1897 Antonius Stanislas Kozlowski was consecrated in Berne (Switzerland) by three Old Catholic bishops. He thus became the first bishop of what has come to be known as the Polish National Catholic Church of America. This has grown to be the largest of all the Old Catholic churches, having (in 1970) a reputed membership of about 350,000.

Philippine Independent Church

In the Philippines, the Spaniards had done little to encourage Filipinos in the ways of independence in either church or state. At the time of the political revolution at the end of the 19th century, an ecclesiastical movement similar to the Polish National Catholic Church came into being, and in a short time (by 1902) the Philippine Independent Church was established under the leadership of Gregorio Aglipay y Labayán and Isabelo de los Reyes y Florentino. Since there was no way of securing regular episcopal consecration, 12 priests followed the custom of the laying on of hands, which is known to have been the practice of the early church of Alexandria, and thus consecrated Aglipay as their bishop. At one time the Philippine Independent Church claimed about 3,000,000 members. It was ill organized, however, and, as it had adopted Unitarian ideas, was not in communion with any other church of the Christian world. A change began when Isabelo de los Reyes, Jr., was consecrated as supreme bishop of the church, a post to which he was elected in 1946. Under his influence the church recovered its orthodox Christian faith and entered into relations with the Protestant Episcopal (Anglican) Church in the United States of America, which in 1948 conferred regular episcopal orders on three bishops of the Philippine Independent Church. Through them the succession was passed on to all other bishops, and in 1961 the Philippine Independent Church was admitted to full communion with the Church of England and with the Old Catholic churches.

Movements headed by episcopi vagantes. Since the second half of the 19th century there has been a considerable number of movements (more than 150) of dissent from the main-line Christian churches, many of them under the lead of virtuous but ill-advised clergymen, who have aspired to become bishops and have believed themselves to have received a special vocation to unite all the now separate sections of the Christian Church. Historically, the first of these new leaders was Jules Ferrette, who had received Roman Catholic orders in 1855 and in 1866 was consecrated with the title Mar Julius I, bishop of Iona, by the Jacobite bishop (who is not recognized by the patriarch of Constantinople) of Homs (Emesa) in Syria. It is affirmed, and also denied, that this consecration had the approval of the Jacobite patriarch. The arrival of Mar Julius in England aroused some interest among Anglicans who hoped for the reunion of Christendom through the Eastern rather than through the Roman Church. Mar Julius seems to have consecrated one successor, Richard Williams Morgan (1874), to whom he gave the title Mar Pelagius I, first patriarch of a restored Ancient British Church, supposedly founded in AD 63 by Joseph of Arimathea. Ferrette later spent some years in the United States but returned to Europe and died in Geneva in 1903. A large number of *episcopi vagantes* claim descent from Ferrette.

The work of Joseph René Vilatte

The second notable figure among the *episcopi vagantes* is Joseph René Vilatte (1854–1929), a French Roman Catholic, who had for a period worked in association with the Protestant Episcopal Church in Wisconsin but

on the recommendation of the bishop of Fond du Lac was ordained deacon and priest by Bishop Herzog of the Swiss Christian Catholic Church in 1885. Having vainly sought episcopal consecration in various directions, in the end Vilatte obtained what he desired from Antonio Alvarez (Mar Julius I), a Goan in South India who was metropolitan of the Independent Catholic Church of Ceylon, Goa, and India, and who had received rather doubtful episcopal orders from prelates in Kerala (India). The consecration took place in Colombo in May 1892. Although Vilatte referred to himself as metropolitan of the Old Catholic Church of America, he was formally repudiated by the Old Catholic bishops of Europe, who did not accept his consecration as valid. Vilatte's high hopes of extensive work among French-speaking Roman Catholics in the United States came to nothing, and toward the end of his life he returned to France. He died on July 8, 1929, having consecrated a number of bishops who in their turn have become the progenitors of a number of espiscopal successions.

Arnold Harris Mathew (1852–1919), after education as an Anglican, joined the Roman Catholic Church in 1875 and was ordained priest in 1877. Having become convinced that there was a great future for an Old Catholic movement in England, he approached the Old Catholic bishops with a view to episcopal consecration. In spite of the fact that Mathew was married—whereas the Dutch Church had not abolished the rule of celibacy for the clergy—he was consecrated at Utrecht in April 1908. Shortly after his return to England, Mathew took the title of "Old Catholic archbishop of London." But before long he gravely offended the Old Catholic bishops by raising to the episcopate two dissident Roman Catholic priests, without consultation with any other bishop. From this time on these bishops made it clear that they had disowned Mathew and that they would have nothing more to do with him. Mathew made various approaches to the Church of England, but these were not enthusiastically welcomed. He had been hopelessly mistaken in his estimation that there were possibilities of creating a large Old Catholic movement in England; his following was never other than very small, and no such movement ever came into being.

Succession of the *episcopi vagantes*

Almost all of the *episcopi vagantes* trace their succession to one or other of the above three bishops. The majority of these bishops were and are good and simple men, in part inspired by personal ambition but in many cases driven on by a genuine desire to be of service to the cause of Christian unity. Few of them have had any deep theological training. A notable exception is F. Heiler (1892–1967). Professor of the University of Marburg and author of many distinguished theological works (*Das Gebet*, 1918), he secured consecration in the Vilatte succession in 1932 with the title of bishop of the Communio Evangelica Eucharistica (Evangelical Eucharistic Communion).

Almost all of these bishops have an intense and one-sided belief in the power of the laying on of hands by anyone who stands in any line of allegedly valid episcopal succession. One of the strangest features of their belief is the desire to collect in one person as many lines of episcopal succession as possible, as a kind of extra guarantee for the validity of their claims. Thus, in April 1944 Hugh George de Willmott Newman (1905–) was raised to the episcopate by Mar Basilius Abdullah III (W.B. Crow) of the Ferrette line and given the name Mar Georgius and the title "Archbishop and Metropolitan of the Holy Metropolis of Glastonbury, the Occidental Jerusalem, and Catholicos of the West." Mar Georgius is understood to have undergone in the course of the next ten years a number of reconsecrations, as a result of which no fewer than 23 lines of succession have come to be united in his person. At the recent consecration of a Lutheran pastor in Germany, it was affirmed that 15 distinct lines of succession were represented.

In spite of the multiplication of successions, not one of these churches has received recognition from any one of the historic Old Catholic churches or from any one of the major churches of Christendom. Thus, their contribution to the cause of Christian unity has been minimal. The communities over which such bishops rule are very small and have exercised no major influence on the developments that have taken place in the Christian world in the last century.

DOCTRINE, PRACTICE, AND ORGANIZATION

Claim to the true tradition of the Western Church

Doctrine and practice. The Old Catholic churches of the continent of Europe claim to maintain the true tradition of the Western Church, without the new Roman Catholic doctrines of the Immaculate Conception (*i.e.*, that Mary, the Mother of Jesus, was conceived without sin; 1854), the corporal Assumption of Mary (*i.e.*, that Mary was physically taken up into heaven; 1950), and the infallibility of the pope (1870). True to the commission laid upon them by Döllinger, these churches have avoided doctrinal innovation. Yet all seem to have been influenced to some extent by the doctrinal ferment around them and to have given some attention to Protestant theological thought.

The churches founded by the *episcopi vagantes* have made little contribution to theological thought. Most of them may be classed as orthodox in theology, and this is an indispensable condition for the fulfillment of their self-imposed mission of working for the unity of all the great Christian traditions. The one notable exception is the Liberal Catholic Church, which from the beginning held theosophical (Oriental-Western mystical) views, and, since it denies the uniqueness of Jesus Christ, stands only on the very fringe of the Christian tradition.

Liturgical practice. In worship, the one radical change made by the Old Catholic churches was the introduction of the vernaculars in church services. In Holland, however, Latin was used until 1909 (a new liturgy was promulgated in 1960). The Swiss liturgy was issued both in French and German in 1880, and that for Germany in 1888. The giving of communion to the laity in both kinds (bread and wine) seems to have been practiced for the first time at Berne in 1879. All these liturgies are conservative, following closely the traditions of the Western Church but omitting or modifying everything that is considered to be inconsistent with true Catholic theology. The unrecognized churches have brought forth a multiplicity of liturgies. These are almost without exception eclectic in character, taking characteristic features from all the known liturgies of Christendom but themselves without clear character as the expression of any particular form of Christian theology.

Ecumenical relationships. True to the charge laid upon them by Döllinger, the Old Catholics have always been in the forefront of the modern ecumenical movement. They have been present at all the great ecumenical conferences and at the present time are represented on the Central Committee of the World Council of Churches. Perhaps their greatest contribution has been in the field of ecumenical journalism. The periodical known today as the *Internationale kirchliche Zeitschrift* ("International Church Journal"), founded in 1893 as the *Internationale theologische Zeitschrift* and from 1895 to 1911 as the *Revue internationale de théologie*, and edited for many years by the Swiss Bishop A. Küry, is indispensable to the student of church history for that period; in particular, the quarterly survey of events in the Eastern churches is unique in the variety and reliability of the information provided.

Relationships with other communions

The Old Catholic churches have always been especially interested in the Orthodox churches and in closer relations with them. Efforts continue to be made in this direction, but so far no formal step has been taken. They likewise have not remained unaffected by the new currents in the Roman Catholic Church. Relations, which naturally had been strained for a century, are probably better today than they have been at any time since 1870. Since 1931 the Old Catholic churches have had a steadily increasing measure of communion with the churches of the Anglican Communion. According to the Declaration of Bonn (Germany) of 1931, accepted first by the Church of England but subsequently by most of the other Anglican churches, each of the negotiating groups of

churches recognized the catholicity and independence of the other, though without being committed to being identical in regard to all doctrinal opinions or liturgical practices. Since that date interconsecration of bishops has been frequently practiced; it is probable that today 80 percent of Anglican bishops and clergy have the Old Catholic as well as the Anglican succession of the episcopate and the ministry. In 1946 the Polish National Catholic Church of America entered into similar relations of communion with the Protestant Episcopal Church in the U.S.A. This has gradually been extended to most of the other churches of the Anglican Communion. Similarly, the Philippine Independent Church, which first entered into relationship with the Protestant Episcopal Church, now has relations with a wide range of churches, including the Old Catholic churches.

Being small in numbers and preoccupied with maintaining their own existence, the Old Catholic churches have never had missionary work of their own. With ecumenical contacts, a new interest in missions in the non-Christian world has grown up, and some support has been given to Anglican missions, especially in South Africa.

Organization. The Old Catholic churches hold firmly to episcopacy as an essential element in the life of the church. Each bishop has considerable liberty in his own diocese, and national churches are self-determining also in the fields of liturgy and of the organization of parochial life. The centre of unity for the scattered churches is the Union of Utrecht, to which all the Old Catholic churches in Europe have adhered since 1889, and in which the conference of Old Catholic bishops is accepted as the highest authority in the church. As such, the conference determined the terms of intercommunion with the Church of England in 1931, affirmed the adherence of the Old Catholic churches to the World Council of Churches in 1948, and defined the attitude of these churches to the dogma of the Assumption of the Blessed Virgin in 1950.

The churches brought into existence by the *episcopi vagantes* depend very much on the charismatic qualities of the founder and can hardly be said to have any recognized organization.

Official statistics for the Old Catholic churches were given in 1970 as follows: Holland 12,000; Switzerland, 30,000; Germany 30,000; Austria 38,000; Czechoslovakia 4,000; Polish National Catholic Church of America 350,000, in Poland 90,000; smaller groups 25,000. The Philippine Independent Church claims 3,500,000 adherents under 56 bishops and 640 priests; it seems likely, however, that this is a considerable overestimate and that there are not more than 1,000,000 members closely in touch with the church.

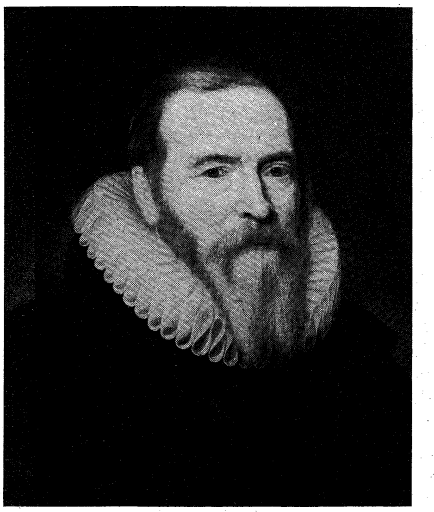
BIBLIOGRAPHY. For the main-line Old Catholic movement the standard work in English is C.B. MOSS, *The Old Catholic Movement: Its Origin and History*, 2nd ed. (1964). In German, U. KURY, *Die Altkatholische Kirche: ihre Geschichte, ihre Lehre, ihr Anliegen* (1966), is valuable as a comprehensive survey. The Philippine Independent Church, still a highly controversial subject, is treated at great length in P.S. DE ACHUTEGUI and M.A. BERNARD, *Religious Revolution in the Philippines: The Life and Church of Gregorio Aglipay, 1860–1960*, 2 vol. (1960–66). Another point of view is presented in L.B. WHITTEMORE, *Struggle for Freedom: History of the Philippine Independent Church* (1961). The most recent and very full study of the *episcopi vagantes* is P. ANSON, *Bishops at Large* (1964), including a splendid bibliography; however, Anson does not entirely supersede the pioneer work on the subject: H.R.T. BRANDRETH, *Episcopi Vagantes and the Anglican Church*, 2nd rev. ed. (1961). The affair of Bishop A.H. Mathew has been dealt with from the Anglican side by G.K.A. BELL in *Randall Davidson: Archbishop of Canterbury*, 3rd. ed., pp. 1016–1023 (1952).

(S.C.N.)

Oldenbarnevelt, Johan van

Johan van Oldenbarnevelt was a Dutch statesman and (after William I the Silent, prince of Orange) the second founding father of Dutch independence.

Born at Amersfoort (province of Utrecht) on September 14, 1547, Oldenbarnevelt studied law at Louvain, Bourges, and Heidelberg (where his ultimate conversion to Protestantism first germinated) and, probably, Padua. After his return to the Netherlands he settled down as an *advocaat* ("counsel") at the Hof van Holland, roughly speaking, the court of appeal for the province of that name, established at The Hague.



Oldenbarnevelt, painting by M.J. van Mierevelt (1567–1641). In the Rijksmuseum, Amsterdam.
By courtesy of the Rijksmuseum, Amsterdam

Role in Dutch independence

When, in 1572, two of the Netherlands provinces, Holland and Zeeland, succeeded in shaking off the Spanish rule from Brussels, Oldenbarnevelt did not follow the Court of Appeal, which fled to Utrecht, but decided to throw in his lot with the movement of national liberation. He even took part in an attempt to relieve the besieged towns of Haarlem and Leiden. In 1576 he was appointed "pensionary" (strictly, legal adviser, but as a full-time job) of Rotterdam, an office that automatically implied membership of the provincial states (assemblies), and, when the national revolt had spread to the other provinces, frequent attendance at the States-General in Brussels or Antwerp. In 1578, when a total reconquest by the Spanish armies under the leadership of Alessandro Farnese, duke of Parma, threatened, Oldenbarnevelt was one of the negotiators of the so-called closer Union of Utrecht (concluded January 1579), which was to serve as a kind of makeshift constitution for the United Provinces until 1795. During the negotiations, it became apparent that Oldenbarnevelt was aiming at securing for Holland the politically unassailable position to which the strategically all but unassailable province considered itself entitled after having borne the brunt of the revolt alone with Zeeland for nearly seven years. These activities also brought him into fairly close contact with William the Silent.

In 1586, two years after the Prince's assassination, Oldenbarnevelt accepted the appointment by the States of Holland as the province's *landsadvocaat—i.e.*, the office that, after his downfall, was renamed *raadpensionaris* ("great pensionary"); in this office he became the real trustee of William the Silent's political inheritance. While the latter's son, Maurice of Nassau, a brilliant military commander, was in charge of the actual warfare in the field, it was Oldenbarnevelt who, at first in close collaboration with him, mobilized and coordinated the country's available energy and resources, thus making warfare possible. As one of these activities, he took an active part in the founding of the Dutch East India Company.

Although theoretically a servant of only one out of seven sovereign provinces, Oldenbarnevelt, who himself undertook several diplomatic missions to France and England, was virtually the union's foreign secretary. In this capacity, too, it was he who continued the work of William the Silent and succeeded in integrating the somewhat suspicious, politically unorthodox new commonwealth of rebels against their lawful overlord. In this respect, his greatest triumph was the conclusion of a full-fledged triple alliance with France and England in 1596.

Oldenbarnevelt's main achievement in the field of foreign policy was, however, the so-called Twelve Years' Truce, concluded in 1609 after long-protracted negotiations, by which the original national program of ousting the Spaniards from the whole of the Netherlands was virtually abandoned and the northern commonwealth of the seven provinces established as such.

On the other hand, it was also in this achievement that the all-pervading flaw in Oldenbarnevelt's position as the union's leading statesman became apparent—*i.e.*, the very circumstance that he was pledged by oath to only one out of those seven provinces of the ancient Netherlands that, as the saying was, "remained with the union," the province of Holland. Filling a vacuum created by the fact that the formerly predominant provinces, Brabant and Flanders, had been reconquered by the Spanish arms, Holland, a hitherto peripheral county, had become preeminent; it contributed close to 59 percent of the federal budget. Under those conditions, a large degree of "Hollandocentrism" or even of "Hollandism" appeared more or less inevitable, and the landsadvocaat, or "great pensionary," was its chief exponent and advocate within the union setup.

In the same way, the princes of Orange found themselves almost automatically cast in the role of prime exponents of the union conception, and it is in the light of this that the ensuing conflict between Oldenbarnevelt and Maurice becomes fundamentally understandable, even without a detailed examination of its multiple origins. In this context it is utterly irrelevant that, after his father's death, Maurice had been made stadholder of Holland by Oldenbarnevelt and his political friends, with the express purpose of safeguarding the province's individuality and special position in the days (1586) when Robert Dudley, earl of Leicester, as governor general, tried to impose his conception of centralized government on the various provinces.

Religious conflicts during the Twelve Years' Truce

During the Twelve Years' Truce, the latent conflict crystallized around its religious facets. The astonishing success of the Netherlands' independence movement was indissolubly connected with the fact that, sooner or later, all the provinces "remaining with the union" came to be ruled by Calvinist minorities for whom, short of renouncing their faith, there existed no possibility of reconciliation with the abjured ruler, Philip II of Spain. As for Oldenbarnevelt, like William the Silent, he too had accepted membership of the Reformed Church, but he and his fellow "regents" in Holland cherished the ideal of a church that was, though based on the Reformation in its Calvinist shape, sufficiently latitudinarian in its dogma to attract and satisfy all those who were willing to relinquish the Roman obedience. According to these rulers of Holland, the nation had rebelled against the centralizing and tyrannical tendencies of its Hispanicized overlord for the sake of freedom, including the freedom from inquisitorial practices, of which they were equally unwilling to accept the Calvinist Genevan as well as the Roman or Spanish brand. As seen by many theologians and preachers, on the other hand, the revolt had taken place for the sake of the Reformed religion in its most uncompromisingly strict dogmatic variety. When, on the perilous issue of predestination, the antithesis became polarized in a conflict between two professors of theology at Leiden—the strict Gomarus and the more moderate Arminius—Oldenbarnevelt and the majority of the voting towns in Holland, though not Amsterdam, favoured the Arminians against the bulk of the Calvinized masses, who were staunchly Gomarist or, as they were commonly called, Counter-Remonstrant. Paradoxically, Oldenbarnevelt and his adherents even had to safeguard the principle of tolerance by somewhat intolerant measures; those preachers who, in spite of various decrees to the purpose, remained stubbornly unwilling to refrain from preaching controversial sermons were dismissed and occasionally sent into exile.

The religious controversy was, moreover, inextricably interwoven with the antithesis of province versus union, the Counter-Remonstrant preachers consistently referring to their religion as the God-given "cement" that kept the union together. Translated into terms of actual politics, this meant that they wanted to convoke a "national"—*i.e.*, an interprovincial—synod trusting (as it turned out, wrongly) that it would establish a church triumphant on the Genevan model, completely independent from all civil authorities whose worthiness and consequently whose right to govern would be judged by the churchmen. For obvious reasons, the States of Holland, led by Oldenbarnevelt, considered a synod of this kind far too risky and withheld their consent to its meeting.

Disputes with Prince Maurice

Seen in retrospect, the climax announced itself when in July 1617 Prince Maurice sided openly, not to say defiantly, with the Counter-Remonstrants. This veiled declaration of war on Oldenbarnevelt and the Holland regents' party was answered by a so-called sharp resolution voted by the States of Holland on August 4, 1617, which, among other things, encouraged the various towns in the province to recruit armed units of their own, not integrated in the federal army and not even subject to Maurice's command as the province's captain general. The states remained within their rights in taking such measures. It is understandable that a man like Maurice considered such measures as an intolerable violation of the union statute. Slow-moving tactician that he was, the Prince spent no less than a whole year in reinforcing his position throughout the union, until suddenly, on August 29, 1618, he took Oldenbarnevelt prisoner, together with some of his closest collaborators, chief among whom was his informal "crown prince," the famous Hugo Grotius, then pensionary of Rotterdam.

Never in the course of Dutch history was the problem of union versus province more crudely manifest than when it materialized in the vexing question of how and by whom Oldenbarnevelt was to be tried for his life. His own thesis, unassailable at least in theory, was that, having exclusively acted as a civil servant of the sovereign province of Holland, he was responsible only to the judiciary of the province of Holland; his enemies, on the other hand, wanted to have him tried for felony against the union. As, however, no federal judiciary existed, the only possible expedient was to summon an extraordinary tribunal ad hoc; it consisted of 24 judges, by no means all of whom were qualified lawyers, and not a few of whom, besides being political opponents, were also personally antagonistic to Oldenbarnevelt. Even so, after more than half a year's imprisonment and interrogation, he was condemned to death not for high treason, for which public opinion had been carefully propagandized, but for the "subversion" of the country's religion and policy. On May 13, 1619, he was beheaded on the Binnenhof, at The Hague. More than any other event in the country's history, his execution has continued to haunt Dutch historiography and even Dutch politics, almost until the present day.

BIBLIOGRAPHY. In spite of its enduring literary merits, J.L. MOTLEY, *Life and Death of John of Barneveld, Advocate of Holland*, 2 vol. (1874), is utterly untrustworthy as far as the subject matter is concerned, although, to a lesser extent, its extreme partiality justifies the same verdict on the book written against Motley by the Dutch historian and politician, GROEN VAN PRINSTERER, *Maurice et Barnevelt* (1875). The authoritative biography is the Dutch work by JAN DEN TEX, *Oldenbarnevelt*, 3 vol. (1960–71).

(J.J.P.)

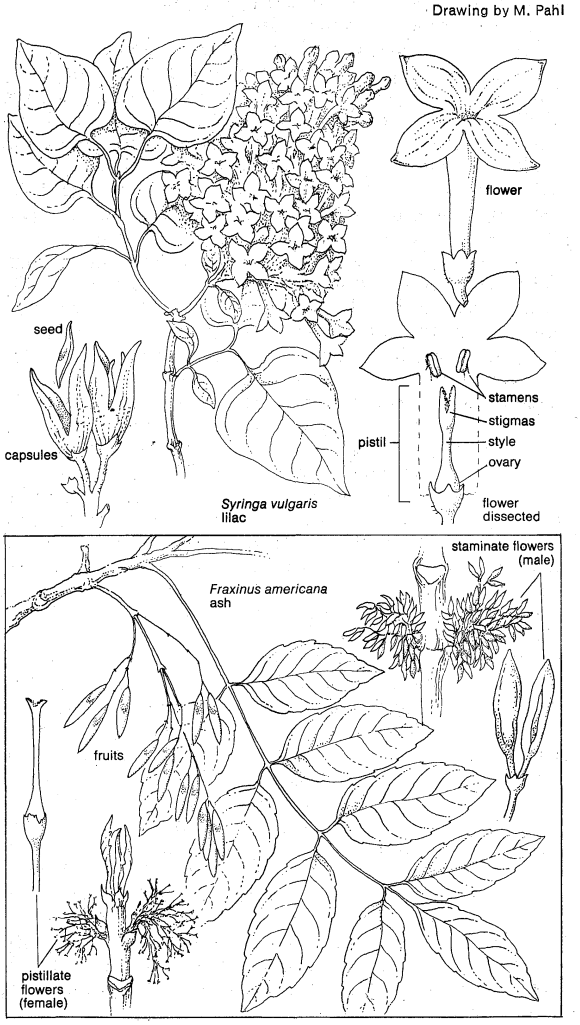
Oleales

The olive order (Oleales) is a small order of flowering plants containing a single family, the Oleaceae. The woody plants comprising this group are distributed throughout the world, except in the Arctic. Although they are found in many different habitats, only rarely are they the most dominant plants.

General features. *Diversity and distribution.* A number of plants in this order are of economic or aesthetic importance—for example, the olive (*Olea europaea*) is the source of olives and olive oil; the ashes (genus *Fraxinus*) are noted for their hardwood timber; and many genera are famous for their horticultural merit—*e.g.*, *Syringa* (lilacs), *Jasminum* (jasmines), *Ligustrum* (priv-

ets), *Forsythia*, and *Osmanthus*. Most members are trees or shrubs, but a number, such as most jasmines, are woody climbers. The tropical and warm-temperate species are evergreen; those from the colder North Temperate Zone are deciduous. Certain genera are found in both Eurasia and North America (*e.g.*, *Fraxinus*), others are scattered throughout the tropics (*e.g.*, *Linociera*). The largest genus, *Jasminum*, is confined to the Old World; any species growing wild in the Americas today have escaped from cultivation. The genus *Menodora* has an unusual disjunct distribution (found in widely separated regions without apparent connections) and is found in North America, South America, and South Africa. This distribution is also remarkable in that one species, *M. heterophylla*, is native to both South America and South Africa.

Drawing by M. Pahl



Representative plants of two genera in the order Oleales.

Origin of the cultivated olive

Economic importance. The most important economic product is the olive, the fruit of *Olea europaea*. This tree, particularly characteristic of the Mediterranean region and of very ancient origin, appears to have no truly wild progenitors, and, although the cultivated olive can seed itself, the offspring fail to produce the same large fruits. The olive probably arose at the dawn of agriculture, perhaps in the area of the Fertile Crescent in the Middle East. Very closely related plants with small, generally inedible fruits are found over a large part of the Old World—in southwest Asia; in eastern Africa as far south as South Africa, including the highest mountains of the Sahara; on Madeira Island in the North Atlantic; and in the drier areas of the Sino-Himalayan region. Today the olive is widely grown as a commercial crop in every country with a Mediterranean-like climate.

Although several members of the order Oleales are well-known in horticulture, the most famous is the lilac, with its colourful, sweetly scented blossoms. Over 400 cultivars of the European lilac, *Syringa vulgaris*, with colours ranging from lilac, purple, red, and pink to white and with single or double flowers, have been developed. Lilacs have been grown in gardens for centuries. The Persian lilac, *S. persica*, was one of the first plants introduced to Europe from the East; other Asiatic species have been introduced more recently.

Natural history. Members of this order are generally characteristic of woodlands, woodland margins, and forests of various types. Their ecological amplitude is not particularly great, and no members are especially adapted to desert conditions, salty soils, or the Arctic climate. Some species of *Fraxinus*, the ashes; of *Olea*, the olives; and of *Menodora* grow in dry climates, but not in true desert.

Pollination and diversity of flowers. Pollination is usually carried out by insects; many types are involved, and no particular specialization exists between the flowers and their pollinators. Nectar is usually produced within and at the base of the petals, and where these form a tube, it is more or less concealed. This ensures that in obtaining nectar the insect pushes down beyond the stamens (male pollen-producing structures) and stigma (the female pollen-receiving structure), touching both. A wide variety of insects visit flowers within this order, including bees, beetles of various kinds, and several different flies.

The most common flower colour in the order is white, although lilacs occur in a range of colours, and in *Forsythia*, *Menodora*, and a few species of *Jasminum* the flowers are yellow. In many genera, where the individual flowers are not very large, their conspicuousness to pollinating insects is increased by aggregation into fairly compact inflorescences (flower clusters) and by the production of a strong scent. In fact, the family Oleaceae, especially lilacs, jasmines, and *Osmanthus*, is noted for its sweetly fragrant flowers. The name *Osmanthus* means "fragrance flower," and this genus is particularly prized in China and Japan, where the dried flowers are used to scent certain Oriental teas, as are those of a few species of *Jasminum*. *J. sambac*, whose native provenance is uncertain but is probably India or Southeast Asia, is cultivated throughout the tropics and has been for many centuries. Its flowers are often used in making necklaces, or leis; and in Hawaii, where it is not native, it is especially grown for this purpose. *J. grandiflorum*, commonly called Spanish or Catalonian jasmine, has long been prized for its scent and cultivated in Spain and in southern France, where it is grown commercially for the extraction of oil of jasmine.

Scent and colour

Other genera in the order are also noted for their fragrance, and even the privets (species of *Ligustrum*) have a strong, sweet odour. It is probable that some of the heavily scented, white-flowered members are pollinated at night. The pollinating insects of *Jasminum*, the fused petals of which form a longish tube, probably are night-flying moths with a long proboscis, a tubular sucking organ with the mouth at the tip.

In a few species of *Fraxinus*, such as the manna or flowering ash, *F. ornus*, which is native to the Mediterranean region, the flowers have small, white, strap-shaped petals and produce nectar. (The name manna ash refers to an edible sugary material exuded from the cut stem of the tree and at one time used in medicine.) Most species of *Fraxinus*, however, like the majority of forest trees in temperate regions, are wind-pollinated. The flowers of the white ash (*F. americana*), many other American species, and the common European ash (*F. excelsior*) possess no petals; quantities of pollen are blown by the wind, and the sexes of flowers on individual trees are often separate.

The chance of cross-pollination is increased by the development of unisexual flowers in several genera. The American genus *Forestiera*, like many ashes, has unisexual flowers without petals. In other genera, such as *Osmanthus*, the flowers of individual trees are often functionally unisexual because either the small ovary or the stamens are nonfunctional. In *Forsythia*, *Abeliophyllum*, and *Jasminum* the flowers of individual plants have either

short styles (the narrow upper end of the ovary, with the stigma at its top), with the stigma below the level of the stamens, or long ones, with the stamens borne below the stigma. Seed normally develops only if pollen is transferred from a long-styled to a short-styled flower or vice versa. The plants in fact are self-sterile, and cross-pollination is essential.

Seed dispersal. Seed dispersal in a number of genera is aided by wind. The whole fruit is winged in *Fontanesia*, *Abeliophyllum*, and the ashes, in which the fruit is traditionally known as a "key." Individual seeds are somewhat flattened in *Forsythia*, *Syringa*, and the tropical genus *Schrebera* so that, after being shed from the capsule, each has a greater chance of being blown away from the immediate vicinity of the parent plant. In most members of the family (*e.g.*, olive, most privets, and jasmines), however, the fruit is fleshy, generally dark purple or black, and has a hard, stony seed. In the natural state these fruits are consumed by birds, which helps to disperse the plant species.

Form and function. *Vegetative features.* A vegetative feature common to almost every member of the order is that of opposite leaves; a few species of *Jasminum* are the only exceptions. In many genera the leaves are pinnately compound—*i.e.*, with numerous small leaflets arranged on both sides of a central axis, the whole constituting a single leaf. Examples include most species of *Fraxinus*, many jasmines, two lilacs (*S. laciniata* and *S. pinnatifolia*), and some species of *Schrebera*. In *Forsythia suspensa* the leaves are trifoliolate (three leaflets) on some shoots, especially those growing rapidly, or single-bladed, as in the other forsythias. The groups of genera closely related to the olive have simple leaves, often thick and leathery, and in some species of *Osmanthus* the leaves are holly-like and prickly.

Characteristic wood of the Oleales

Anatomically the family is fairly uniform in its characteristics. Typically the vessels (*i.e.*, the water-conducting cells) of the wood are small, and accordingly, the wood of the olive, with its fine and close grain, is prized for carving and turnery. The timbers of most species of *Olea* and their close relatives are very durable—even when used out-of-doors—and their hardness is often reflected in their native names; *e.g.*, black ironwood, for *Olea capensis* of South Africa, and devilwood, for *Osmanthus americanus* of southeastern United States. Ash timber, from species of *Fraxinus*, has a worldwide reputation for toughness and resilience and is used for purposes where these characteristics are required—*e.g.*, in tool handles, hockey sticks, and frames for tennis rackets.

Floral features. Generally speaking, flower structure throughout the family is very uniform with four sepals, four petals, two stamens, and two fused carpels forming a single superior (*i.e.*, positioned above the other flower parts) ovary. The four petals (in *Jasminum* there may be seven, eight, or nine) are generally joined at the base to a greater or lesser extent to form a tube. Most species of *Fraxinus*, however, have no petals; nor do the four native olives of New Zealand (*Nestegis* species), the American genus *Forestiera*, and one of two Asiatic species of *Olea*. In *Chionanthus*, the fringe-tree, the long, strap-shaped petals form a very short tube at the base. Superficially similar flowers are found in the relatives of *Fraxinus ornus*; in *Linociera* the petals, split to the base, are held together in pairs by the base of stalks of the two stamens.

Although two is the characteristic stamen number, there are a few exceptions. *Nestegis* species, four of which are from New Zealand and one, *N. sandwicensis*, from Hawaii, often bear four stamens (occasionally up to six), as does the Malesian *Osmanthus scortechinii*. The ovary is much more constant in its characters, producing only one, or at most two, mature seeds in a fleshy olive-like fruit, in a winged "key" as in *Fraxinus*, or in a dehiscing (splitting-open) capsule as in the lilacs. In *Forsythia* the capsule may contain several seeds.

The pollen produced by members of the order is more or less similar in structure and of a rather nonspecialized type. The grains are ellipsoid with three (or four) longitudinal furrows and a finely reticulate surface. The minor differences in pollen structure among the various genera help confirm that the members form a natural group. Pollen grains somewhat similar in appearance are found, however, in several other families, including, for example, certain members of the family Celastraceae (order Celastrales).

Evolution. *Fossil record.* The order Oleales is relatively poorly known from fossils. The oldest one recognized as belonging to the order is a leaf impression identified as *Fraxinus* from the early Cretaceous beds of Greenland (of about 80,000,000 years ago). Similarly, a leaf identified as *Fraxinus* is recorded from the Upper Cretaceous in the United States (about 65,000,000 years ago). Numerous fruits and leaf impressions, many bearing a very close resemblance to present-day species, have been found in later beds throughout North Temperate regions, through to the Pliocene (7,000,000 years ago), and in even more recent deposits.

Oldest known members of the order

Other fossil members of the Oleales are first known from a fruit named *Olea headonensis* from the early Paleocene to Oligocene deposits in Britain (65,000,000 to 38,000,000 years ago) and one called *Oleaecarpium germanicum* from the upper Oligocene of Europe (about 30,000,000 years ago). Good leaf impressions of *Olea* dating from 20,000,000 to 12,000,000 years ago have been found in south and central Europe and North America. Leaves said to represent *Notelaea*, a present-day Australian genus, have also been found in the same regions, but in this case the generic identification is not so certain. A fruit identified with the tropical genus *Nathusia* (now known as *Schrebera*)—several living species occur in Africa, one in South America, and one in Asia—and called *N. rugosa* has been found in European deposits from the middle to upper Oligocene (30,000,000 to 26,000,000 years ago). Another tropical genus, *Jasminum*, has been recorded from fruits found in the Netherlands dating from 7,000,000 years ago in the lower Pliocene.

Phylogenetic-evolutionary relationships. The order Oleales contains the single, somewhat isolated family Oleaceae. A relationship has been suggested to certain families in the order Celastrales, such as the holly family, Aquifoliaceae, the Salvadoraceae, and the Icacinaceae, and it has also been suggested that these groups probably share a common ancestor in the order Saxifragales. There is little unanimity among botanists, however, as to how the family Oleaceae should be classified. It has been variously given ordinal status, sometimes alone or with the small family Salvadoraceae; classed within the order Gentianales; and sometimes placed in a restricted part of the Gentianales that is separated and called the order Loganiales. Most recently it has been placed within a modified order Scrophulariales. Because in most genera the petals are joined into a tube and the ovary is always two-parted, the family Oleaceae has usually been included in orders in which there is always a corolla (the collection of petals) of united petals and often in that group of families that have in the past been loosely called the "Bicarpellatae." It seems most probable, however, that the order Celastrales or its progenitors were the group from which the family Oleaceae arose.

Various classification schemes

The family Oleaceae seems to consist of two main groups of genera: those placed in the subfamily Jasminoideae and those in the subfamily Oleoideae. A study of the chromosome numbers of different genera confirms the affinity of various genera otherwise grouped on morphological characters. *Forsythia* and *Abeliophyllum* (which has been called white forsythia) are unique in that both possess 28 chromosomes in their vegetative cells; in a recent classification the two genera have been placed in the subfamily Jasminoideae along with *Jasminum* and *Menodora*, which have 26 and 22 chromosomes respectively. They are perhaps best treated separately, however, as the two pairs of genera bear little resemblance to each other.

Within the subfamily Oleoideae the genus *Fraxinus*, which, because the best known species lack petals, has traditionally been classified alone or with *Fontanesia* (because of the latter's winged fruits), is probably quite closely related to the many genera that produce olive-like fruits. The genus *Syringa* (lilacs) and the genus *Ligus-*

trum (privets) also belong in this subfamily and are closely related, despite their dissimilar fruits.

The small genus *Nyctanthes* (best known from *N. arbor-tristis*, a tree sacred to the Hindus), native to India, together with the related monotypic (having a single species) *Dimetra* of Thailand, has been variously classified in the family Oleaceae or in the family Verbenaceae (order Lamiales). Traditionally it belongs to the former and is possibly, if distantly, related to *Jasminum* or *Schrebera*; various anatomical and morphological characteristics suggest, however, that they should be treated as a subfamily within the family Verbenaceae.

Classification. *Distinguishing taxonomic features.* The family Oleaceae is most easily recognized by a combination of characters: it has a woody habit with opposite leaves; flowers consisting of four sepals and petals (usually united), two stamens, and a united ovary of two chambers. There are exceptions, however, to most of these individual features.

The characters used to classify the genera within the family are the presence or absence of the petals and whether they are united to form a tube, however short, or free, and whether the petals overlap in the bud (imbricate) or are arranged with their edges touching (valvate). The type of fruit has also been important, whether it is a fleshy olive-like drupe (a stony-seeded fruit), a capsule that splits open to release dry seeds, or a winged samara (or key) as in the ashes. The fruit type, however, is no sure guide to relationships, nor is the absence of petals; *Nestegis*, for example, generally has no petals but is related to the petaloid genera *Olea, Osmanthus*, and *Notelaea*. The wild olive of Hawaii (*Nestegis sandwicensis*), however, has petals in its flowers. Similarly, although most species of *Fraxinus* have no petals, those related to *F. ornus* have well-developed ones. Most ashes, such as the common European ash, *F. excelsior*, have flowers reduced to complete simplicity, for there are no sepals either, and because the sexes are often separated in different inflorescences or on different individual trees, the flowers consist only of a single, naked ovary or a pair of stamens. Many of these, however, are crowded together in branching inflorescences. There seems no doubt that this simplicity results from reduction and does not reflect primitiveness.

Structural variation among close relatives

Annotated classification.

ORDER OLEALES

Woody plants, trees, shrubs, or climbers. Leaves opposite, rarely alternate, simple or pinnately compound, and lacking stipules (small, leaflike appendages at the base of the leafstalk). Inflorescence basically cymose (determinate; *i.e.*, the flowers mature in sequence from the tip, or central flower, downward) but paniculate, decussate, fasciculate (various modes of branching), or single-flowered. Flowers regular (radially symmetrical); bisexual, or rarely unisexual, sometimes by abortion. Calyx (sepals) of 4 lobes (rarely more, or lacking). Corolla (petals) of 4 (rarely to 12) united lobes, rarely almost free or lacking, aestivation (condition of the petals in the bud) imbricate or valvate (rarely contorted). Disc absent. Stamens 2 (rarely 3, 4, or 6), borne on the corolla and alternate with the lobes; anthers usually with a slight extension to the connective, dehiscing (opening) longitudinally. Ovary superior, of 2 fused carpels, each locule with two (rarely 1 to 4) axile, pendulous or ascending, ovules, style 1 or lacking, usually with a 2-lobed stigma. Fruit usually 1-seeded, a drupe, berry, capsule, or samara, seed with cellular endosperm (nutrient tissue for the developing embryo) or none. About 28 genera and between 500 and 600 species in temperate and tropical regions throughout the world, but especially the Old World, eastern Asia, and the tropics.

Family Oleaceae

The only family in the order, it has the characters of the order. It is divided into 2 subfamilies, Jasminoideae and Oleoideae.

Critical appraisal. The Oleaceae comprise a natural family, despite the number of individual exceptions to characteristic features, such as the lack of petals in some members and the occasional occurrence of alternately arranged leaves. Moreover, its classification in an isolated position is probably correct, for relationships with other families and orders are not clear. New information from the traditional study of morphology or anatomy that might help to elucidate these relationships is unlikely, but the rapidly developing fields of palynology (the study of pollen grains) and chemotaxonomy (the study of biochemical characters in relation to taxonomy) may well provide valuable new evidence.

Within the family the relationships of certain genera are uncertain. The subfamily Jasminoideae, as at present constituted, appears to be a rather heterogeneous assemblage containing the genera *Jasminum, Forsythia, Fontanesia, Schrebera*, and *Myxopyrum*. It is doubtful whether these genera are really more closely related to one another than to others now placed in the second subfamily, the Oleoideae. Here, however, the correct delineation of the genera themselves is often the problem. Should *Olea* be split into two or three separate genera (*e.g.*, into *Steganthus* and *Tetrapilus*) as has been proposed, and within the genus *Osmanthus*, should the sections *Leiolea* and *Siphosmanthus* be raised to generic rank? As with problems at interordinal and interfamilial levels, new and significant morphological data are unlikely to be forthcoming. Chromosomal information seems of no assistance in this subfamily, but modern developments in biochemistry may eventually be of value.

Taxonomic problems

Perhaps the outstanding problem at the species level concerns the relationships and origin of the cultivated olive. *Olea europaea* is not known in the wild, and, although closely related plants exist in many tropical and warm temperate parts of the Old World, it is uncertain whether the domesticated olive was developed from them and, if so, how it was done. Chemotaxonomic studies together with breeding and cultural experiments—growing species and plants from different provenances side by side and crossing them—might provide decisive information, but because the olive is a relatively slow growing tree such investigation could be undertaken only on a long-term basis.

BIBLIOGRAPHY. A. CRONQUIST, *The Evolution and Classification of Flowering Plants* (1968), a recent classification with discussion; P.S. GREEN, "The Olive Family in Cultivation," *Arnoldia*, 25:13–27 (1965), a review of cultivated members of the olive family; L.A.S. JOHNSON, "A Review of the Family Oleaceae," *Contr. N.S.W. Natn. Herb.*, 2:395–418 (1957), a review of classification within the family; A. LINGELSHEIM, "Oleaceae-Fraxineae et Syringeae," in A. ENGLER (ed.), *Das Pflanzenreich IV* (1920), a monograph of a few genera; S.D. MCKELVEY, *The Lilac: a Monograph* (1928), standard monograph on *Syringa*; H. TAYLOR, "Cyto-Taxonomy and Phylogeny of the Oleaceae," *Brittonia*, 5:337–367 (1945), a review of intrafamilial classification; K.A. WILSON and C.E. WOOD, "The Genera of Oleaceae in the Southeastern United States," *J. Arnold Arbor.*, 40:369–384 (1959), a discussion, with bibliography, of eastern North American genera.

(P.S.G.)

Olivares, Conde-Duque de

Prime minister and court favourite of King Philip IV, the Conde-Duque de Olivares governed Spain for over 20 years during a period that saw the climax of an epoch of literary and artistic splendour as well as the most chronic political decadence.

Olivares was born Don Gaspar de Guzmán y Pimental, in Rome on January 16, 1587. His father, Don Enrique de Guzmán, was the Spanish ambassador to that city. His mother, Doña María Pimental Fonseca, was a member of the Castilian nobility.

As second-born son of an aristocratic family, Don Gaspar studied for the priesthood, obtaining a degree from the University of Salamanca in law, theology, and the arts (1601–04). With the death of his older brother, however, he renounced his position as canon in Seville (to which he had been appointed by Pope Clement VIII) and joined his father in Valladolid, then the location of the Spanish royal court. In 1607, orphaned and heir not only to a noble title but also to one of the largest fortunes of the kingdom, Don Gaspar married his cousin and niece (they were related through both sides of the family), Doña Inés de Zúñiga y Velasco, lady-in-waiting to Queen Margaret. Their only child was a girl, María, who died shortly after her marriage. Don Gaspar also had an ille-

Olivares, oil painting by Velázquez (1599–1660). In the Prado, Madrid.
Archivo Mas, Barcelona

gitimate son, Julianillo Valcárcel, whom he later recognized and named Enrique Felipez de Guzmán, marquis of Mairena.

In 1615 Olivares became one of Prince Philip's six personal attendants. When Philip was crowned king in April of 1621, he had just reached 16 years of age, and Olivares was approaching the age of 34. By this time Olivares, a man of unpleasing appearance and changing moods, had become the young king's irreplaceable companion. As Philip's favourite he was given the rank of grandee, the title most coveted by Castilian nobility. Reluctant to drop any part of his title, he styled himself "conde-duque" (count-duke).

Domestic policies

From 1623 until January 24, 1643, Olivares served as prime minister of Spain. He was unswervingly loyal to the King and was vehemently patriotic. He was also avid for power—both for himself and for Spain. The main objective of his domestic policy was to engender national unity among the separate kingdoms of the peninsula, kingdoms that he described as "anachronistic as crossbows." He attempted many economic reforms aimed at relieving the difficult situation that had arisen as a result of long reliance on the influx of precious metals from the New World. Among these programs were restrictions on granting favours (except for honorary titles); recoining of the old copper alloy moneys; introduction of paper money; promotion, with the aid of the Castilian Cortes (council), of various royal decrees to stop the industrial and commercial decline of the kingdom; and a project whereby the shipping companies would be able to compete more advantageously with the Dutch, English, and French commercial fleets. But his attempts to promote trade and industry met with failure, due largely to the fact that aristocratic Castilians, slaves to the idea of a rigid class structure, looked down upon all mercantile professions. His moves toward centralizing power in the hands of the king and his ministers were partly responsible for the revolts of the Catalans and the Portuguese, which began in 1640, and for an abortive conspiracy to form a separate Andalusian kingdom (1641).

In foreign policy Olivares was guided by the dream of *austracismo,* a joint European hegemony of the Austrian and Spanish Habsburg kingdoms. This policy meant continued Spanish involvement in the Thirty Years' War and ended with the eclipse of Spanish power by France. Yet in the period of the Counter-Reformation, it is difficult to conceive of Spain following a different course: in this sense it was almost inevitable, and Olivares can hardly be judged in terms of its ultimate failure.

Olivares' character

One biographer has painted a picture of Olivares as the typical *picnico,* a man who was strong and slightly heavy, a man subject to wide and sudden ranges of mood—from a maniacal, irrational exultation and optimism to complete despair and depression. He thus presented a sharp contrast to his main rival, Cardinal Richelieu of France, a cold, irritable, reserved man, known for using few words and taking swift action. In 1632 Olivares envisioned the inevitable destruction of the Spanish monarchy; in 1635 he was convinced that, if all his war plans were executed, there would be no enemy capable of opposing or defeating Spain. Scarcely two months later, in another mercurial change of mind, he again gave up any hope of Spain's retention of power. From that time onward, his moods tended more toward deep depression, with occasional euphoric interludes, such as that caused by the success of Spanish armed forces threatening Paris in December 1636.

As a result of a court intrigue headed by the Queen (Elizabeth of France), Philip removed his ailing favourite from office in January 1643. Although the King would undoubtedly have liked to recall him later, other grandees, long jealous of his power, continued to discredit him. Eventually Olivares was exiled, along with his wife, to the city of Toro. In December 1644 the Inquisition began to investigate his conduct. He died at Toro on July 22, 1645.

BIBLIOGRAPHY. G. MARANON, *El conde-duque de Olivares (la pasión de mandar)*, 5th ed. (1965), an excellent psychological study, although to some degree vindictive; E. ZUDAIRE HUARTE, *El conde-duque y Cataluña* (1964), a valuable study of the fight between Olivares' modern concept of the State and Catalonia's separatist policy; J.H. ELLIOTT, *The Revolt of the Catalans: A Study in the Decline of Spain, 1598–1640* (1963), an analysis of the political relations between Castille and Catalonia during the reign of Olivares, and *Imperial Spain, 1469–1716* (1963), an excellent history of Spain dealing with the tensions that resulted from Castille's unitarian policies and Catalonia's federalist policies.

(E.Z.H.)

Olivines

The olivines comprise a group of common rock-forming silicate minerals that are closely related on chemical and structural grounds. The name olivine alludes to the quasi-olive green colour of the most abundant varieties, forsterite and fayalite, which are silicates of magnesium and ferrous iron, respectively. Other members of the group may contain manganese or calcium in addition to, or substituting for, iron or magnesium; among these, the more important named varieties are tephroite (containing manganese), monticellite (calcium and magnesium), kirschsteinite (calcium and iron), and glaucochroite (calcium and manganese). These varieties and some of their physical properties are listed in the accompanying Table.

Some Physical Properties of the Olivines

end members	composition	specific gravity	hardness on Mohs scale	unit cell dimensions*		
				a	b	c
Forsterite	Mg_2SiO_4	3.22	7	4.756	10.195	5.918
Fayalite	Fe_2SiO_4	4.39	6½	4.817	10.477	6.105
Tephroite	Mn_2SiO_4	3.78	6	4.90	10.60	6.25
Monticellite	$CaMgSiO_4$	3.08	5½	4.815	11.08	6.37
Kirschsteinite	$CaFeSiO_4$	—	—	—	—	—
Glaucochroite	$CaMnSiO_4$	3.41	—	4.92	11.19	6.51

*The unit cell is the smallest volume of a mineral containing a complete sample of the atomic or molecular groups that comprise it. Unit cell dimensions along the three crystallographic axes (*a, b, c*) are given in angstrom units (one angstrom equals 10^{-8} centimetres).

The olivine minerals that contain magnesium and iron are generally believed to be among the most important minerals in the Earth's upper mantle (the region directly beneath the Earth's crust). They are particularly characteristic of basic and ultrabasic igneous rocks; that is, rocks that are relatively poor in silica and rich in iron and magnesium that crystallize from molten material within the Earth (Figure 1). They also occur in metamorphic

Figure 1: Phenocryst of olivine (centre) in a fine-grained groundmass; olivine basalt from Ookala, Mauna Kea, Hawaii (magnified 28 ×).
By courtesy of C.E. Tilley

(altered) rocks and in some meteorites. When clear and of good colour, magnesium-rich olivine is used as a gemstone and has the names peridot (for dark-coloured varieties) and chrysolite (light-coloured). Because of its high melting point and resistance to chemical reagents, magnesium olivine is an important refractory material—*i.e.*, it can be used in furnace linings and in kilns when other materials are subjected to heat and chemical processes.

This article treats the crystal structure and chemical composition of the olivines and their physical properties, occurrence in nature, and synthesis. For further information on crystal structures see CRYSTALLOGRAPHY, and for the general relations of olivines to other mineral groups, see MINERALS and SILICATE MINERALS. Ornamental varieties are covered in the article GEMSTONES.

FORSTERITE–FAYALITE SERIES

Chemical composition. The most abundant olivines are intimate mixtures of forsterite, pure magnesium silicate (Mg_2SiO_4), and fayalite, pure ferrous silicate (Fe_2SiO_4); most of the naturally occurring specimens are intermediate in composition to these two pure compounds, called end-members, and have the general chemical formula $(Mg,Fe)_2SiO_4$ (Figure 2). The members of this mixture, the forsterite–fayalite solid-solution series (a single crystalline phase that varies in composition between finite limits—*i.e.*, the end-members), are given the general mineral name olivine. The name forsterite is restricted to those species with no more than 10 percent iron substituting for magnesium; fayalite (from Fayal Island in the Azores, where it was believed to occur in a local volcanic rock but probably was obtained from slag brought to the island as ship's ballast) is restricted to species with no more than 10 percent magnesium substituting for iron. Species intermediate in composition to forsterite and fayalite are also named; from most magnesium-rich to most iron-rich they are: chrysolite (90–70 percent Mg), hyalosiderite (70–50 percent Mg), hortonolite (50–70 percent Fe), and ferrohortonite (70–90 percent Fe).

Solid-solution series

The continuity in the forsterite–fayalite series has been verified experimentally. At the magnesium-rich end of the solid solution series, natural crystals may contain very small amounts of calcium, nickel, and chromium; the iron-rich members near the other end of the series may incorporate small amounts of manganese and calcium. Apart from ferrous iron, the crystalline structure of the olivines is also capable of accommodating relatively small amounts of ferric iron; dendrites (small branching crystals) of magnetite or chromite found oriented with respect to some crystallographic direction within such olivines may be attributed to exsolution (that is, to precipitation while cooling in the crystalline state). But the presence of relatively large amounts of ferric oxide in the analyses of olivines clearly indicates either an advanced state of oxidation or the mechanical inclusion of co-precipitating magnetite upon crystallization from the magma.

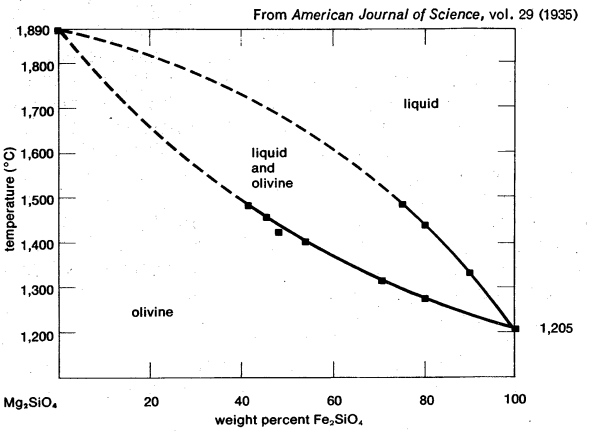


Figure 2: Equilibrium diagram of the system Mg_2SiO_4-Fe_2SiO_4.

X-ray diffraction data for powders of synthetic forsterite and fayalite permit estimation of the composition of any solid solution between these two end-members to an accuracy of better than 4 percent.

In addition to the forsterite–fayalite series, other complete solid-solution series exist among the various olivine minerals. Fayalite is soluble in all proportions with ashgray tephroite (from Greek *tephros*, "ashen"), pure manganese silicate (Mn_2SiO_4); the intermediate in the series is knebelite ($FeMnSiO_4$). Tephroite and knebelite are known only from manganese and iron-ore deposits, from metamorphosed manganese-rich sedimentary rocks, and from slags.

Crystal structure. The olivines are classified as nesosilicates because their crystalline structure (Figure 3) consists of silicate tetrahedra (structures in which four oxygen atoms surround and are bonded to a central silicon atom) that are completely separated from each other by the various metal cations (positively charged ions)—*i.e.*, Mg^{+2}; Fe^{+2}. These positive cations are in octahedral coordination; that is, they are immediately surrounded by six oxygen atoms at the corners of an octahedron, with each oxygen atom belonging to a different silicate tetrahedron. The symmetry of the structure is orthorhombic (referable to three mutually perpendicular crys-

By courtesy of the Crystallography Laboratory, University of Amsterdam

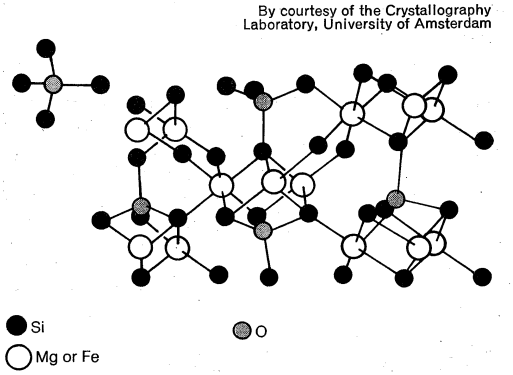


Figure 3: Model of the structure of olivine.

tallographic axes of unequal length), and the unit cell (the smallest volume containing a complete sample of the atomic or molecular groups present) contains four molecules. The oxygen atoms form parallel sheets so stacked that their arrangement approximates the intimate configuration known as hexagonal close packing, wherein the position of every second sheet approximates the position of the initial sheet.

Monticellite, calcium and magnesium silicate ($CaMgSiO_4$), and kirschsteinite, calcium and iron silicate ($CaFeSiO_4$), have essentially the same structure as the forsterite–fayalite minerals, but the calcium atoms are so placed that the magnesium and iron atoms are restricted to certain positions.

PROPERTIES OF OLIVINES

Banded structures

Physical properties. The specific gravity and hardness of the olivines are listed in the Table. There are at least two cleavages—*i.e.*, the tendency to split along preferred crystallographic directions (perpendicular to the *a* and *b* axes in this case)—both of which are better developed in the iron-rich varieties. Forsterite contained in certain ultrabasic rocks may show a banded structure when observed in thin sections with a polarizing microscope; in some dunites (a rock consisting nearly entirely of olivine), for example, olivine is preferentially oriented so that the cleavage plane perpendicular to the *b* axis is parallel to the microscopic laminated structure of the rock. Individual grains of olivine within such rocks typically appear as oriented bands with angles of up to 10 degrees between them. Such banding, which is undoubtedly the product of incipient mechanical deformation, also can be observed within the olivine nodules of some basalts.

To the unaided eye pure forsterite appears colourless, but as the content of ferrous oxide increases, specimens show yellow-green, dark-green, and, eventually, brown to black tints. In thin sections under the microscope, however, even pure fayalite appears pale yellow. Pure tephroite is gray, and monticellite also appears gray or colourless.

Some variations of optical properties observed in natural olivine crystals probably result from small but varying replacements of magnesium and iron by calcium and manganese and of silicon by titanium, chromium, and ferric iron.

Crystal habit and form. The magnesium-iron olivines occur most commonly as compact or granular masses. Except for the well-shaped phenocrysts (single crystals) of such olivines found embedded in the fine-grained matrices (groundmass) of basalts (see Figure 1), distinctly developed crystals are relatively rare. The phenocrysts in basalts are characterized by prominent pinacoid (*i.e.*, having a pair of parallel faces) and prism faces, so that their cross sections often appear six- or eight-sided. With fayalite the morphology is often simple. Monticellite and tephroite commonly show prominent pyramidal faces. Twinning, the intergrowth of two or more grains of the same mineral according to some crystallographic pattern, is not very frequently observed. When twinning does occur, trillings (the intergrowth of three grains) may be produced and, in monticellite, six-pointed star shapes, as reported from the Highwood Mountains in Montana.

OCCURRENCE IN NATURE

In igneous and metamorphic rocks. The magnesium-iron olivines are essential minerals in basic and ultrabasic rocks. Their composition in these rocks ranges from Fo_{92} (92 percent forsterite by molecular weight) in dunites and about Fo_{88} in peridotites through the interval Fo_{88}–Fo_{50} in basalts, gabbros, and dolerites. The minerals associated with olivine in basalts and in gabbros are indicated in the expanded basalt tetrahedron (see Figure 4). This tetrahedron represents the chemical relationships between the minerals larnite (La), nepheline (Ne), forsterite (Fo), and the various phases of silica (Qz). These minerals are shown at the corners of the tetrahedron, which represent the pure compounds. Moving along the edges and faces of the tetrahedron from one corner increases the content of the other mineral phases. Thus, along the edges and faces of this tetrahedron are located the intermediate phases enstatite (En), albite (Ab), wollastonite (Wo), monticellite (Mo), diopside (Di), akermanite (Ak), and sodamelilite (Sm). On the tetrahedron, the four points representing the minerals nepheline, forsterite, quartz, and diopside form the corners of another tetrahedron, the simple basalt tetrahedron: Ne - Fo - Qz - Di. The three points Di, Ne, and Fo define a critical plane (shaded) of undersaturation with respect to silica. This plane separates the basic mineral assemblages of the simple basalt tetrahedron, which contain plagioclase (Ab) and olivine (represented below the plane) from the olivine-bearing ultrabasic mineral assemblages that are devoid of plagioclase but contain monticellite (Mo) and melilite (Ak) (represented above the plane).

The basalt tetrahedron

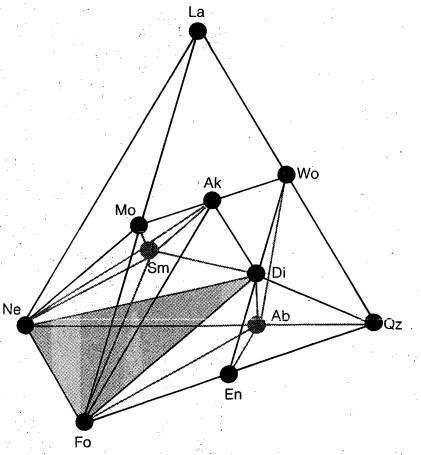


Figure 4: Expanded basalt tetrahedron showing mineral associates of magnesium-iron olivine (see text for explanation of symbols).

From J.F. Schairer and H.S. Yoder, Jr., "Crystal and Liquid Trends in Simplified Alkali Basalts," *Carnegie Institution of Washington Year Book 63*, p. 65, 1964

Olivines richer in iron than Fa_{50} are less common; they do occur in the iron-enriched layers of some intrusive rocks, however. Fayalite itself occurs in small amounts in some silicic volcanic rocks, both as a primary mineral and in the lithophysae and vugs (bubble-like hollows) of rhyolites and obsidians (volcanic glass). It also occurs in acidic plutonic rocks such as granites, in association with iron-enriched amphiboles and pyroxenes (*qq.v.*).

Olivines also occur in metamorphic environments. Both forsterite and monticellite typically develop in the zones in which igneous intrusions make contact with dolomites. Forsterite tends to develop at lower temperatures than monticellite as the process of decarbonation in the contact zone progresses. Fayalitic olivines develop within metamorphosed iron-rich sediments. In the quaternary (*i.e.*, four-component) system Fe_2O_3 - FeO - SiO_2 - H_2O, fayalite is associated with the minerals greenalite (iron-serpentine), minnesotaite (iron-talc), and grunerite (iron-amphibole) in various metamorphic stages. In chemically more complex environments, which, in addition to the above components, also involve lime (CaO) and alumina (Al_2O_3), fayalite may be associated with hedenbergite, eulite (iron-rich orthopyroxene), grunerite, and almandine (iron-garnet).

In meteorites and in the Earth's mantle. In meteorites, the olivine is usually a forsteritic variety containing only Fa_{15} to Fa_{30}. In the Nakhla (Egypt) meteorite, an achondrite, the olivine is more ferrous, however, containing as much as Fa_{65}. In the chondrites (stony meteorites), the olivine commonly is incorporated in the distinctive spheroidal bodies referred to as chondrules, which range up to one millimetre in diameter.

Because the rocks of the upper mantle directly below the Mohorovičić discontinuity (the zone separating the crust from the mantle, in which the velocity of transmission of seismic waves changes) are believed to consist of peridotite and garnetiferous peridotite that contain oliv-

ines as their most abundant minerals, it is of crucial importance to establish experimentally the behaviour of these minerals when subjected to high pressures (see also EARTH, STRUCTURE AND COMPOSITION OF). Study of the olivine-like compound magnesium germanate, Mg_2GeO_4, showed that it has polymorphs (same chemical composition but different crystal structures) that have both the olivine structure and spinel structure. In the spinel structure, the oxygen atoms are arranged in cubic close packing (in which the position of every third layer repeats that of the initial layer) instead of hexagonal close packing (in which the position of every second layer repeats that of the initial layer) of the olivine structure. The spinel form of Mg_2GeO_4 was found to have a density exceeding that of the olivine form by 9 percent. In 1936 it was suggested that at high pressures Mg_2SiO_4 might also transform to a spinel structure; this suggestion was adopted in 1937 as a basis for explaining the so-called 20°-discontinuity, an observed seismic discontinuity in the mantle at a depth of about 400 kilometres (250 miles).

Relation to the spinel structure

In 1966 it was shown that each of the three synthetic olivines, Fe_2SiO_4, Ni_2SiO_4, and Co_2SiO_4, could be transformed directly to a spinel structure at a temperature of 700° C (1,300° F) and at pressures below 70 kilobars (the bar is a unit of pressure equal to atmospheric pressure about 100 metres above sea level). These spinel structures were denser by approximately 10 percent than the corresponding olivine structures. In 1968 a series of synthetic magnesium and iron olivines was subjected to a range of pressures between 50 kilobars and 200 kilobars (700,000 pounds per square inch and 3,000,000 pounds per square inch) at a temperature of 1,000° C (1,800° F). In the composition range Fe_2SiO_4 to $(Mg_{0.8}Fe_{0.2})_2SiO_4$, these olivines were transformed completely to their spinel polymorphs, which are isometric crystals (referable to three mutually perpendicular crystallographic axes of equal length), with an accompanying increase in density of 10 percent. In the composition range $(Mg_{0.8}Fe_{0.2})_2SiO_4$ to Mg_2SiO_4, however, the olivines were transformed to another orthorhombic structure (called β-orthosilicate) at a pressure of about 130 kilobars (1,900,000 pounds per square inch) and a temperature of 1,000° C (1,800° F). This β-phase polymorph, with a density only 8 percent greater than that of the corresponding olivine structure, is believed to be the stable phase in the field of its synthesis. The change in the crystalline structure of olivine to its spinel polymorph, accompanied by a change in the structure of magnesium-iron pyroxenes to a new garnet-like structure at depths of 350 kilometres (220 miles) to 450 kilometres (280 miles) in the mantle, is believed to be responsible for the observed abrupt change in the velocity of seismic waves at these depths (see also EARTHQUAKES).

The spinel polymorph of olivine has been recorded in the Tenham (Queensland) chondrite as pseudomorphs after olivine. Portions of some large grains of olivine immediately adjacent to black, shock-generated veins are recognized as transforms to the spinel phase; the associated plagioclase feldspar was converted to maskelynite. The composition of the spinel phase in the meteorite has been analyzed by means of an electron probe and found to be $(Mg_{0.75}Fe_{0.25})_2SiO_4$; in thin sections it appears blue-gray to violet-blue. It has been named ringwoodite after A.E. Ringwood, an Australian earth scientist who synthesized spinel phases with compositions and properties close to those of the mineral found in the meteorite. More recently, ringwoodite also has been found in the Coorara (Western Australia) meteorite in association with a garnet phase. The β-phase polymorph has not yet been observed in shocked meteorites—*i.e.*, those that have undergone impact shock—but it is highly probable that it, too, exists in relative abundance within the Earth's mantle.

Other occurrences. Knebelite olivines are restricted to iron-manganese ore deposits, to their associated skarn (lime-bearing silicate rocks) zones, and to metamorphosed manganiferous sediments. At Franklin, New Jersey, tephroite and glaucochroite occur in the same deposit as roepperite, a knebelite containing 10.7 weight percent zinc oxide (ZnO).

Monticellite occurs in some alkali peridotites and within limestones near their contact with peridotites. Pure kirschsteinite is known only from slags and has not yet been observed as part of a natural mineral assemblage. The most plausible natural environments for kirschsteinite should be altered limestones, and it is possible that the mineral has remained unrecognized in such rocks because its optical properties (the chief means of identification) are similar to those of the much more common magnesium-iron olivines. A kirschsteinite containing 31 percent by weight of other olivines, particularly monticellite, has been reported from a nepheline-melilite in north Kivu Province, Zaire.

Sources of monticellite

Glaucochroite, pure calcium and manganese silicate ($CaMnSiO_4$), is rare; it has only been reported from an ore deposit in Franklin, New Jersey, where it occurs with tephroite. It is thought that the limited availability of manganese in parent magmas accounts for the rarity of minerals intermediate in the solid solution series between the calcium-rich olivines monticellite, glaucochroite, and kirschsteinite.

Alteration products and weathering. Olivines gelatinize in even weak acids and offer little resistance to attack by weathering agents and hot mineralizing (hydrothermal) solutions. The forsteritic olivines are altered principally through leaching, which results in the removal of magnesium and the addition of water and some iron. The chemical reactions are usually complex and involve hydration, oxidation, and carbonation. The fayalitic olivines are altered principally through oxidation and the removal of silica. The usual products of alteration are the minerals serpentine, iddingsite, and bowlingite, all of which may occur as pseudomorphs (forms with the outward appearance of the original mineral but which have been completely replaced by another mineral). Serpentine, which is the most common alteration product of olivine in ultrabasic rocks, often is accompanied by magnesite. Iddingsite and bowlingite are variable in composition and, even though they appear optically homogeneous, each actually consists of an intimate mixture of several distinct minerals. X-ray diffraction analyses of iddingsite show that the mineral goethite ($FeO{\cdot}OH$) is a frequent component. Also included with the mixture may be hematite (Fe_2O_3), as well as a silicate phase whose structure, even though very irregular and disordered, appears to be related to that of the clay minerals vermiculite or montmorillonite. Iddingsite develops almost exclusively from the olivines of extrusive and hypabyssal (minor intrusive, such as sill and dike) rocks and practically never from the olivines of plutonic and metamorphic rocks. The process of "iddingsitation" of olivine may set in even before the complete consolidation of the lava or the hypabyssal magma. This is evident from the observation of rock textures in which shells of iddingsite surrounding a core of olivine are, in turn, surrounded by more unaltered olivine.

The mechanical weathering of olivine-rich rocks leads to the release of olivine particles that, in the absence of much chemical weathering, may accumulate to produce green or greenish-black sands. Conspicuous examples of such sands occur on the beaches of the islands of Oahu and Hawaii, particularly at Diamond Head (Oahu) and South Point (Hawaii). Alluvial sands rich in olivine are also known from the Navajo County of Arizona and from New Mexico; these sands provide clear olivine used in jewelry.

Olivine in beach sands

EXPERIMENTAL STUDIES

The general stability fields of the olivines are shown in Figure 5, on which the olivines, pyroxenes, and silica phases present are plotted. With regard to olivine end-members, forsterite is readily prepared by heating its component oxides (2MgO and SiO_2) to temperatures of at least 500° C (900° F) at water-vapour pressures of from 140 to 2,800 atmospheres (2,000 to 4,000 pounds per square inch; one atmosphere is equal to the atmospheric pressure at sea level). Although forsterite remains stable

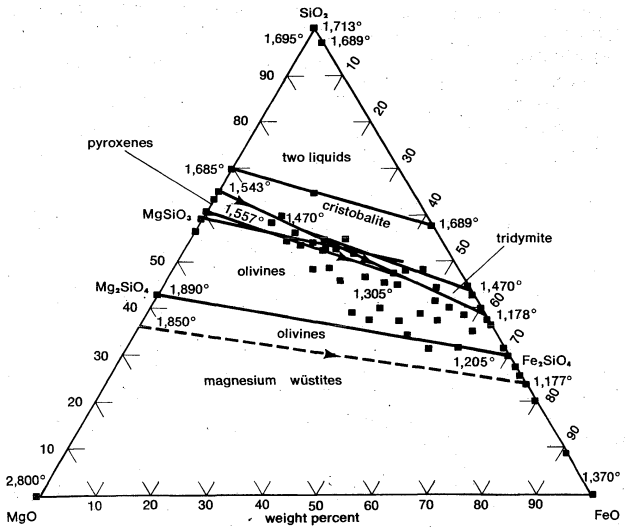


Figure 5: Equilibrium diagram of the system MgO-FeO-SiO_2. Temperatures are given in degrees Celsius.
From *American Journal of Science*, vol. 29 (1935)

in contact with water vapour at all temperatures above 400° C (750° F), below this temperature it is attacked by water, yielding serpentine and brucite:

$$\underset{\text{forsterite}}{2Mg_2SiO_4} + \underset{\text{water}}{3H_2O} \rightleftharpoons \underset{\text{serpentine}}{H_4Mg_3Si_2O_9} + \underset{\text{brucite}}{Mg(OH)_2}.$$

Fayalite also can be prepared hydrothermally from iron oxide and silica ($2FeO + SiO_2 \rightleftharpoons Fe_2SiO_4$) at low partial pressures of oxygen and at temperatures above 250° C. Below this temperature, fayalite is unstable in contact with water, and hydrous phases such as greenalite, accompanied by magnetite, appear instead.

Many of the experimental studies are of great importance for the interpretation of the role of olivine in rocks. In synthetic systems fayalite is found to be quite stable even in the presence of free silica, but forsterite is unstable unless the system is undersaturated with respect to silica. In the binary system MgO - SiO_2 the stability field of the pyroxene-like metasilicate called protoenstatite separates the stability field of forsterite from that of cristobalite. Upon heating to nearly 1,560° C (2,840° F), the pyroxene melts according to the reaction: $2MgSiO_3 \rightleftharpoons MgSiO_4 + SiO_2$. This melting relationship, called incongruent melting because the solid and the melt are not identical, is carried over into the ternary system CaO - MgO - SiO_2, involving diopside, forsterite, and silica, and into the quaternary system Al_2O_3 - CaO - MgO - SiO_2, involving anorthite, diopside, forsterite, and silica. The reverse reaction occurs when forsterite, in contact with the liquid, cools to the incongruent melting temperature. It is recorded in basic and ultrabasic rocks in which partially resorbed crystals of olivine are observed to be encased in a shell of pyroxene, the product of the incongruent melting reaction.

From *American Journal of Science*, vol. 29 (1935)

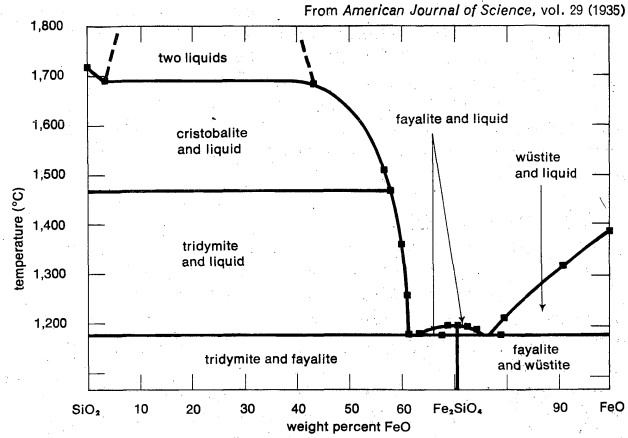


Figure 6: Equilibrium diagram of the system FeO-SiO_2.

Unlike the MgO - SiO_2 system, the binary system FeO - SiO_2 (Figure 6) has no metasilicate phase ($FeSiO_3$) intervening at liquidus temperatures, and there is no obstacle to the coexistence of fayalite (Fe_2SiO_4) with tridymite (SiO_2). But at 1,205° C (2,201° F), fayalite undergoes incongruent melting, with the separation of metallic iron and the simultaneous enrichment of the liquid in Fe_2O_3. This incongruent melting is characteristic of all phase assemblages in the system FeO - SiO_2.

BIBLIOGRAPHY. For a systematic treatment of olivines, see W.H. DEER, R.A. HOWIE, and J. ZUSSMAN, *Rock-Forming Minerals*, vol. 1 (1962); and for synthesis and studies of phase equilibrium of olivines in artificial systems, see N.L. BOWEN, J.F. SCHAIRER, and E. POSNJAK, "The System CaO-FeO-SiO_2" *Am. J. Sci.*, 26:193–284 (1933). Mineral associates of olivine in basaltic rocks are covered in J.F. SCHAIRER and H.S. YODER, JR., "Crystal and Liquid Trends in Simplified Alkali Basalts," *Yb. Carnegie Instn. Wash.*, *63*, pp. 65–74 (1964). A further summary is provided by H.S. YODER, JR. and C.E. TILLEY, "Origin of Basalt Magmas: An Experimental Study of Natural and Synthetic Rock Systems," *J. Petrology*, 3:342–532 (1962). For phase transformations of olivine in meteorites and in the Earth's mantle, see R.A. BINNS, "$(Mg,Fe)_2SiO_4$ Spinel in a Meteorite," in *Physics of the Earth and Planetary Interiors*, vol. 3, pp. 156–160 (1970); H. JEFFREYS, "On the Materials and Density of the Earth's Crust," *Mon. Not. R. Astr. Soc. Geophys. Suppl.*, 4:50–61 (1937); A.E. RINGWOOD, "A Major Synthesis of Mg_2SiO_4-Fe_2SiO_4 Spinel Solid Solutions," *Earth Planet. Sci. Letters*, 1:241–245 (1966), and with ALAN MAJOR, "Phase Transformations in the Mantle," *ibid.*, 5:401–412 (1969).

(C.E.T.)

Olympia

Olympia, a place in Greece in the western Peloponnese, was an ancient religious sanctuary and the scene of the Olympic Games. It lies on the northern bank of the Alpheius (Alfios) River about 10 miles (16 kilometres) from its mouth. A tributary stream, the Cladeus (Kladios), joins the Alpheius just below Olympia, to the south. The country is rich and well watered, consisting of low, wooded hills alternating with farmland.

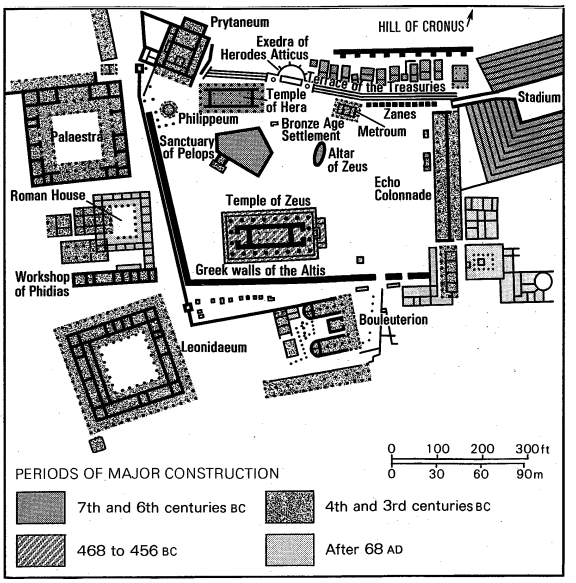
History and excavations. The earliest remains date from 2000–1600 BC, the sanctuary itself from around 1000. First controlled by the town of Pisa, after 570 Olympia came under the jurisdiction of Elis and Sparta. The religious festival, of which the Games were a part, was held there every four years from the 8th century BC until the end of the 4th century AD.

The first excavations were conducted on the site of the Temple of Zeus in 1829 by the French Expédition Scientifique de Morée (A. Blouet). The temple was sufficiently cleared to reveal its general plan, and fragments of three sculptured metopes (panels) were found, which were later placed in the Louvre, in Paris. The great German excavations of 1875–81 cleared the whole of the sacred precinct and some buildings that lay outside it; the position of the stadium was located by exploratory trenches. Thus the plan of a great Greek sanctuary was revealed for the first time. In the early 20th century some small-scale exploratory digging was done in the deeper layers in the sanctuary. Large-scale work was resumed by the Germans in 1936, one of the chief aims being the excavation and restoration of the stadium. Interrupted by World War II in 1942, work was resumed in 1952, and in 1960 the excavation of the stadium was completed, with its restoration in 1961. Other structures were explored in this period, the most important of which was the workshop of the sculptor Phidias.

The remains. The sacred precinct, the Altis, or "sacred grove of Zeus," was an irregular quadrangle over 200 yards (182.9 metres) on a side, bounded on the north by the hill of Cronus and enclosed by a wall on the other three sides. In it were the temples of Zeus and Hera, the principal altars and votive offerings, the treasuries, and administration buildings. Outside were the athletic installations and the hostels, baths, and other accommodations for visitors.

The Temple of Zeus was the largest and most important building at Olympia and one of the largest Doric temples in Greece. Built about 460 BC by the architect Libon of

The Temple of Zeus



Plan of the sanctuary at Olympia.
Adapted from *Westermann Grosser Atlas zur Weltgeschichte;* Georg Westermann Verlag, Braunschweig, West Germany

Elis, the temple was made of a coarse local shell conglomerate, the exposed surfaces being covered with a coat of fine white stucco. The temple had six columns across the front and 13 on the sides. There was a *pronaos* (porch) and an *opisthodomos* (rear porch), and the *naos* (cella; part enclosed by the walls) was divided into three aisles by two rows of slender columns arranged in two stories. The roof tiles were of marble.

The temple was richly decorated with sculpture, much of which has survived and is to be seen in the Olympia Museum. In the front gable the chariot race between Pelops and Oenomaus was represented, and both parties were shown preparing for the race. In the back gable was the battle of the Lapiths and Centaurs at the wedding of Perithous. These sculptures are masterpieces of the early classical style, but the name of the artist is not known. Pausanias' attribution of them to Paeonius and Alcamenes is generally rejected because these sculptors are known to have worked in the later 5th century. The frieze that ran above the front and back porches had sculptured metopes with the 12 labours of Heracles, six at each end. At the peak of the gable was a gilded figure of Victory, and at each corner a gilded caldron, but these have not survived.

Within the temple was the great gold and ivory statue of Zeus, the work of the Athenian sculptor Phidias, the most famous of all ancient statues and counted one of the seven wonders of the world. It made a profound impression on all who saw it, and people generally agreed that Phidias had succeeded in creating the image of Homer's Zeus. The god was represented seated on an elaborately wrought throne. He held a figure of the goddess of victory (Nike) in his right hand and a sceptre in his left.

"The workshop of Phidias"

This statue was made piece by piece by Phidias and his collaborators in a building just outside the Altis to the west of the temple. Subsequently converted into a church, the building was still known in the time of Pausanias as the "workshop of Phidias." The excavations of 1954–58 brought dramatic confirmation of the identification. In the deep layers in and around the building, particularly toward the south, a great mass of material, evidently waste from an artist's atelier, was found. This material included tools, slivers and worked fragments of ivory and bone, glass ornaments, and molds. The clay molds, of a very heavy fabric, like roof tiles, with the larger ones sometimes reinforced with iron rods, are of an unusual open form and were evidently used for hammering into shape the thin plates of gold that formed the statue's drapery. Pottery found with this debris indicates that the workshop was active in the years around 430 BC, an important fact because it settles an old controversy as to whether Phidias made the Zeus before or after his other great chryselephantine statue, the Athena Parthenos, which was completed in 438 BC. The new evidence is decisively in favour of the later date. One of the pieces of pottery, a ribbed mug, has inscribed on its bottom in neat clear letters the words "I am [the property] of Phidias."

The great altar of Olympian Zeus was not in front of the temple, as might have been expected, but to one side and nearer the Temple of Hera. It was elliptical in shape and consisted of an elevated base approached by steps. From the base rose a large mound made of the ashes of the thighs of victims sacrificed to Zeus. The whole height of the altar was 22 feet (6.7 metres).

Temple of Hera

The oldest temple at Olympia and one of the most venerable in all Greece was that of Hera, originally a joint temple of Hera and Zeus until a separate temple was built for him. It has sometimes been thought that the Temple of Hera was built in the 10th or 11th century BC, but this view is now rejected. The existing temple was probably built about 600 BC, and an earlier phase, without peristyle (colonnade), may go back to the 8th century. The temple is long and narrow, having six columns across the ends and 16 along the sides. The columns are Doric, showing a great variety of styles because they were originally of wood and were gradually replaced in stone. In the 2nd century AD there was still one wooden column in the *opisthodomos*. The entablature was of wood, and the upper parts of the walls were of mud brick. The cella had two interior rows of columns, alternate columns being attached by spurs to the cella walls and thus forming bays. Pausanias says that in the temple was an image of Hera seated on a throne with an image of Zeus standing beside her. An archaic limestone head thought to be that of the Hera has been found. Pausanias also reports a stone statue of Hermes carrying the young Dionysus, a work of Praxiteles that was found in the cella of the temple in 1877 and is one of the most prized possessions of the Olympia Museum.

A row of 12 treasuries overlooked the Altis from the lowest slopes of the hill of Cronus. These small structures in the form of Doric temples, date from the 6th century BC. All were erected by Dorian states ranging from Byzantium to Gela in Sicily and Cyrene in north Africa. In the case of only three, Sicyon, Megara, and Gela, is enough material available to allow a reconstruction on paper. These treasuries were erected by the several states either as thank offerings for Olympic victories gained by its citizens or as a general mark of homage to Olympian Zeus and to contain the dedicated gifts in which the wealth of the sanctuary consisted.

Between the temples of Zeus and Hera, the Elean hero Pelops had a sanctuary in the Altis that was open to the sky and surrounded by a wall, with trees and statues within.

The Metroum, or Temple of the Great Mother of the Gods, was a small Doric temple of the 4th century BC just below the treasuries. Because the cult no longer existed in Roman times, the excavated temple contained statues of Roman emperors.

A round building of the Ionic order, with Corinthian half columns on the inside, was erected by Philip of Macedon to commemorate his victory over the Greeks at Chaeronea in 338 BC. The building contained gold and ivory statues of Philip, Alexander, and other members of the family.

The Prytaneum, in the northwest corner of the Altis, was a building that contained a hearth on which burned a perpetual fire and a banquet room in which the Olympic victors were feasted. A large, lavishly decorated fountain, on an apsidal (semicircular) plan, was erected by Herodes Atticus in the name of his wife Regilla. On it were displayed some 20 statues of Herodes and his family and of the Roman emperors Hadrian and Antoninus Pius.

This building was officially called the Stoa Poikile, or "painted colonnade," from the paintings that used to be on its walls, but it was popularly called the Echo Colonnade because an echo repeated a word seven times or more. The colonnade closed the east side of the Altis and was separated from the east Altis wall, which supported

the stadium embankment, by a narrow passage. The colonnade was built soon after the middle of the 4th century BC. Deep down beneath its floor, the starting line of the early classical stadium has been found.

Zanes were bronze statues of Zeus erected with money from fines imposed on those who wantonly violated the rules of the Games. The bases of 16 of these have been found just outside the covered entrance to the stadium, the entrance by which the athletes entered.

The Bouleuterion, or council house, lies just outside the Altis to the south. It comprised two Doric buildings of different date but of identical oblong form with apsidal ends toward the west. In the space between was a rectangular court at the centre of which stood the statue of Zeus Horkios ("Zeus Who Presides over Oaths"). Beside this statue the athletes took the oath not to indulge in foul play during the contests.

Outside the Altis to the southwest stood the Leonidaeum, a large hostel for the reception of distinguished visitors, which was built in the 4th century BC and remodelled in Roman times. To the northwest were the Palaestra, where wrestlers and boxers trained, and the gymnasium, which included an elaborate entrance gateway and a covered running track.

The stadium

The stadium lay to the east of the Altis. In early classical times it was not cut off from the sanctuary, and one end of the track was in the area directly in front of the temple and the great ash altar of Zeus (beneath the later Echo Colonnade). About the middle of the 4th century BC the stadium was shifted about 90 yards (about 82 metres) eastward and a little to the north. The track was surrounded by massive sloping embankments of earth for the accommodation of the spectators, except to the north where the natural slope of the hill sufficed. The western embankment, parallel to which the Echo Colonnade was built, effectively cut the stadium off from the Altis. Connection between the two was maintained by what was called the Krypte, or "covered entrance," which pierced the embankment and, in Roman times, was covered with a stone vault. This entrance was used by the athletes and the umpires. There were no stone seats in the stadium except for a box on the south side about one-third of the way from the starting line nearest the Altis; here the *hellanodikai*, or chief judges of the Games, sat. Directly opposite the box was the altar of Demeter Chamyne, from which the priestess of that cult was privileged to watch the Games (married women were excluded from the Olympic festival, but unmarried girls were permitted). The track was about 230 yards (210 metres) long and 35 yards (32 metres) wide and separated from the sloping embankments by a low stone parapet beside which ran an open stone water channel with basins at intervals. The actual course was marked by stone starting lines at either end. These were about 210 yards apart (192.28 metres = 600 Olympic feet). There was space for 20 runners at a time. The classic race was the stade—*i.e.*, one length of the course. There was also a *diaulos*, two lengths, and a *dolichos*, or long-distance race, the length of which varied and might be as much as 24 stades or nearly three miles. Other athletic contests were also held in the stadium. This 4th-century stadium has been fully excavated and its track and embankments restored so that it may be seen as it was in late classical times.

When the stadium embankments were excavated many votive offerings were discovered. Some of these were works of art of various kinds, including bronze statuettes and reliefs and several terra-cotta statues, of which the most noteworthy was a group of Zeus and Ganymede, about half-life-size and dating from around 470 BC. Others were arms or armour that had been dedicated in the sanctuary; the Olympia Museum houses the largest collection of ancient Greek weapons in the world, some of which have identifying inscriptions on them that are interesting historical documents, such as a Persian helmet with the inscription "The Athenians [dedicated the helmet] to Zeus, having taken it from the Medes."

The hippodrome where the horse races were held lay south of the stadium in the open valley of the Alpheius. No trace of this has been found. Pausanias gives a long description of the hippodrome and of the elaborate starting machinery.

BIBLIOGRAPHY. J.G. FRAZER, *Pausanias's Description of Greece*, trans. with commentary (1913), is the story of the Greek traveller Pausanias, who visited Olympia in the 2nd century AD and wrote a long description of the place in his books v and vi. Frazer's commentary is thorough and makes full use of modern knowledge of Olympia gained from the original German excavations. E. NORMAN GARDINER, *Olympia: Its History and Remains* (1925), is a good general account; and LUDWIG DREES, *Olympia: Götter, Künstler und Athleten* (1967; Eng. trans., *Olympia: Gods, Artists and Athletes*, 1968), an up-to-date account, richly illustrated, with sections on the religious festival, the Games, and the buildings. See also BERNARD ASHMOLE and NICHOLAS YALOURIS, *Olympia: The Sculptures of the Temple of Zeus*, with new photographs by ALISON FRANTZ (1967).

(E.V.)

Oman

Oman—until 1970, Muscat and Oman—is an independent sultanate that occupies the southeastern coast of the Arabian Peninsula. It is bounded on the southwest by the People's Democratic Republic of Yemen, on the south and east by the Arabian Sea, on the north by the Gulf of Oman, on the northwest by the United Arab Emirates, and on the west by Saudi Arabia. In addition, Ru'ūs al-Jibāl (the Mountaintops), a small territory occupying the northern tip of the Musandam Peninsula between the Gulf of Oman and the Persian Gulf, forms a part of Oman; it is separated from the rest of the country by the United Arab Emirates. The United Kingdom assigned three of the nine villages in the oasis of al-Buraymī (Buraimi), south of the United Arab Emirates, to Oman in 1955, but the oasis is still claimed by Saudi Arabia, which also claims the eastern part of the sand desert of the Rub'al-Khali (the Empty Quarter). Because the boundaries with Saudi Arabia are not fixed, the area of Oman cannot be determined; the maximum estimate is about 82,000 square miles (212,400 square kilometres). Since 1967 the Kuria Muria Islands, 25 miles off the south coast of Oman, have also been under Omani sovereignty. The population of the country is about 750,000. Muscat (Masqaṭ), a port on the Gulf of Oman, is the capital.

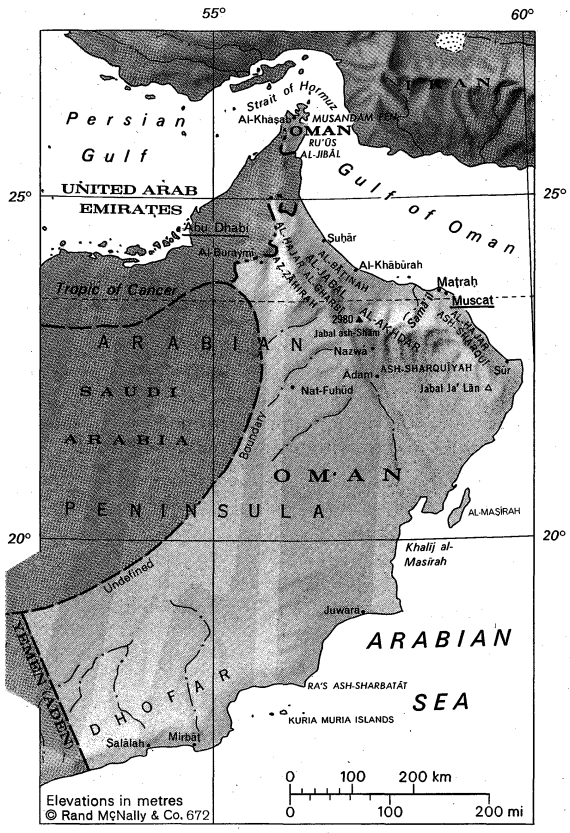
Sultan Sa'īd ibn Taymūr, who acceded to the then British-protected sultanate in 1932, kept Oman isolated from most of the world—maintaining diplomatic relations only with the United Kingdom, the United States, and India—and did little to develop the country. Sultan Qābūs ibn Sa'īd, since overthrowing his father, Sa'īd, in 1970, has taken steps to liberalize the government. Trying to reverse his father's record in external and internal affairs, Qābūs has been handicapped by the fact that many Arabs still regard Oman as too closely linked with the United Kingdom. (For an associated physical feature, see ARABIAN DESERT.)

The landscape. *Relief features and drainage.* The outstanding relief feature in northern Oman is the mountain chain of al-Ḥajar (the Stone), which parallels the coast of the Gulf of Oman from Ru'ūs al-Jibāl to Ra's al-Ḥadd, the eastern extremity of the Arabian Peninsula. These bleak mountains reach a height of 10,194 feet (3,107 metres); their steeper slopes are on the seaward side. Rainfall from the mountains is brought to the oases by ancient underground conduits of a type common in Iran. Mountains also line the Arabian Sea coast of Dhofar (Ẓufār), Oman's southernmost province, reaching their greatest elevation in the hinterland of Ṣalālah, the capital of Dhofar province.

The al-Ḥajar mountains

The Kuria Muria Islands are a group of five; from east to west they are al-Ḥāskīyah, as-Sawdā', al-Ḥallānīyah, Qarzawīt, and al-Qiblīyah. Only al-Ḥallānīyah—the largest, with a population of less than 100 people—is inhabited.

Climate, vegetation, and animal life. Oman is a hot and arid country with high humidity along the coasts. Annual rainfall is between three and four inches. Vast plantations of date palms cover the plain of al-Bāṭinah, and oases dot the mountains and the interior plains. Fruits and some



OMAN

grains are grown, but the staple food, rice, must be imported. The camels of Oman are recognized as the best breeds in the Arabian Peninsula, the mountain donkeys compare with mules in strength and endurance, and a variety of game is found at the higher elevations.

Human settlement. Al-Bāṭinah, the Gulf of Oman coastal plain lying east of al-Ḥajar and reaching south almost to Muscat, is sometimes 20 miles broad and is intersected by many *wādī*s (seasonal watercourses). Ṣuḥār, in the north, is the principal town of al-Bāṭinah. East of Muscat the mountains crowd close to the sea, leaving little room for farming or travel. Al-Ḥajar al-Gharbī ends west of Muscat in al-Jabal al-Akhḍar (the Green Mountain), distinguished from the rest of the range by its larger population and comparatively rich cultivation.

The more gently inclining landward side of western al-Ḥajar is called aẓ-Ẓāhirah (the Upland). The region just below al-Jabal al-Akhḍar is Oman Proper (ʿUmān al-Wusṭā), with Nazwā (Nizwa) as its most important town. Landward of al-Ḥajar ash-Sharqī lies ash-Sharqīyah (the Eastern Region) and Jaʿlan, the latter region extending to the coast of the Arabian Sea, off which lies the island of Maṣīrah, used by the United Kingdom as an air base. Farther south is the region of Dhofar, with the port of Ṣalālah, from which Sultan Saʿīd governed the country.

The harbour city of Muscat, with a population of about 10,000, is ringed by steep, gaunt hills. Access to the interior is easier from its twin port of Maṭraḥ (population 16,000) through the pass of Wādī Samāʾil, which divides al-Ḥajar into its western and eastern halves.

Omani towns are usually dominated by old forts, often now crumbling. Houses are commonly built of stone or palm thatch. Many settlements on the landward side of al-Ḥajar consist of a cluster of villages on an open plain or along a *wādī* bed.

People and population. Of Oman's total population of about 750,000, about 200,000 are settled, the rest being nomadic. Only 5 percent of the population are urban, with 95 percent rural. Arabs constitute the great majority of the people. There are considerable non-Arab elements in the ports—Persians, Baluchis (people of Baluchistan, Pakistan), Pakistanis, Indians, and blacks from East Africa. In aẓ-Ẓāhirah, a thriving community of Baluchi villagers has become thoroughly Arabized.

Most of the Arabs are affiliated with tribes that long kept alive the ancient rivalry between Southern Arabs (descendants of Qaḥṭān) and Northern Arabs (descendants of Nizār). In Oman the tribesmen claiming descent from Qaḥṭān are called Hināwīs, and those claiming descent from Nizār are called Ghāfirīs. This factionalism, once one of the most important features of the society of Oman, appeared to be on the wane in the early 1970s.

Rivalry between the Southern and Northern Arabs

Oman harbours the largest group of Ibāḍīs in the Islāmic world; the Ibāḍīs are a branch of the Khārijites, the first sect to break away from the main body of Islām. Since the 8th century the Ibāḍīs have elected *imāms* in Oman, the last of whom—because of his opposition to Saʿīd—was driven into exile by the British in 1958. Sultan Qābūs, though an Ibāḍī, is not recognized as the religious head of the sect. The Ibāḍīs are concentrated on the landward side of al-Ḥajar, with Nazwā as the old capital of the *imāms*. Sunnites (members of one of the two main religious divisions of Islām, who regard the first four caliphs as legitimate successors of Muḥammad), however, are also fairly numerous in northern aẓ-Ẓāhirah and Jaʿlan, where the influence of Wahhābism (a fundamentalist, puritanical doctrine that opposes all practices not sanctioned by the Qurʾān and the *sunnah*) spread in the 19th century. A sprinkling of Shīʿites (members of the second of the two main religious divisions of Islām, who regard ʿAlī, the son-in-law of Muḥammad, as his legitimate successor) and Ismāʿīlī Muslims (a Shīʿite sect with an esoteric philosophy) and followers of other religions is found among the foreign elements.

The oil industry and development of the country have brought a number of Westerners into Oman, with the government showing a strong preference for hiring British experts. Oman is also attracting Arabs from other countries.

Oman, Area and Population

	area*		population†
	sq mi	sq km	1970 estimate
Provinces (*liwāʾs*)*			
al-Bāṭinah	...	...	...
al-Ḥajar al-Gharbī	...	...	...
al-Ḥajar ash-Sharqī	...	...	...
ash-Sharqīyah	...	...	...
aẓ-Ẓāhirah	...	...	...
Jaʿlan	...	...	...
Muscat	...	...	...
Ruʾūs al-Jibāl	...	...	...
ʿUmān al-Wusṭā	...	...	...
Dhofar	...	...	...
Total Oman	82,000	212,380	750,000

*No boundaries or administrative seats have been established; hence, no breakdown is available. †No complete census has ever been taken in Oman.
Source: Official government figures.

The national economy. *Economic activity.* The discovery of oil in commercial quantities was announced in 1964. A pipeline was constructed from the Nati-Fuhūd area on the edge of the Rubʿ al-Khali down to the Gulf of Oman, and export of oil began in 1967. Production, however, is modest, averaging 300,000 barrels a day in 1971, and reserves are regarded as relatively small. Petroleum is the country's only commercial mineral resource, accounting for more than 90 percent of the total revenue, the remainder of which came from customs duties and a religious tax.

Oil production

Trade is the chief industry of Muscat and Maṭraḥ. Elsewhere the economy is agricultural, the principal crops being dates, limes, coconuts, papayas, bananas, wheat, sugarcane, oranges, and grapes. Agriculture is limited by the shortage of water for irrigation. Camel breeding is practical in the interior and fishing on the coast. Dried fish, dates, and limes are shipped abroad, but, apart from oil, imports greatly exceed exports. The government, which dominates the economy, is heavily burdened with military expenditures.

Transport. In 1971 Oman had two main roads—the paved road from Muscat to the military base of Bayt al-Falaj, next to Maṭraḥ (three miles), and the dirt road from the oil-shipping terminal west of Maṭraḥ across al-Ḥajar to the oil fields. A new road is being constructed from Muscat to Ṣuḥār. Country tracks are often difficult or impassable in al-Bāṭinah and al-Ḥajar but are usually easier on the landward side of the mountains. Improved and unimproved roads total about 600 miles. In 1971 a modern port was being built at Maṭraḥ and an international airport at Muscat. Small steamers call at some of the principal ports, all of which are much frequented by ancient types of sailing craft. Dates are exported from Ṣuḥar, and dried fish and fish meal from Ṣūr and Mirbāt. There is also some shipping at Ṣalālah. Even with the improvements being introduced, development of the country continues to be greatly hampered by inadequate communications.

Administration and social conditions. Sultan Saʿīd had kept personal control of even trivial affairs of state. In the early 1970s, Sultan Qābūs began building up a more elaborate apparatus of government, with ministries and departments, but foreign advisers remained essential as few citizens were qualified for administrative and technical positions. The relatively well-equipped armed forces are trained and led by British officers.

Health and education

Oman faces serious health problems, with many diseases being widespread and the infant mortality rate very high. The government is spending large sums for new hospitals, clinics, doctors, and nurses.

Under Sultan Saʿīd, Oman had three elementary schools, with 900 boys enrolled. In 1970 the number of schools rose to 15, including one for girls; the number of students increased to 7,000. Secondary and higher education still had to be obtained abroad.

Cultural life and institutions. The cultural life of Oman has been confined by the strict form of Islāmic belief prevailing there, particularly among the Ibāḍīs. While representational art has been forbidden, Oman has developed a picturesque style of architecture, seen at its best in the old forts. The cultivation of literature has been impeded by the lack of a printing press, but manuscripts on religious subjects and history have been produced. Until the departure of Sultan Saʿīd, music, the cinema, fraternal associations, and other diversions and activities were forbidden. In the early 1970s cultural life was beginning to expand under the new regime. Oman's first weekly newspaper is printed in Beirut.

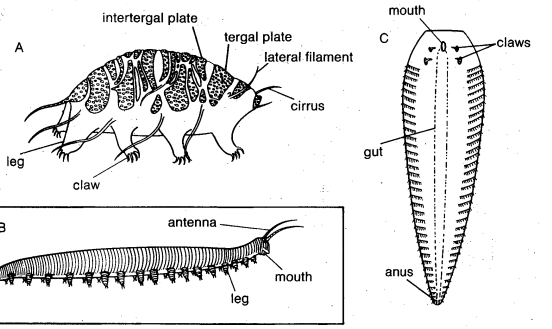
Prospects for the future. The viability of the sultanate depends primarily on the success of its development program and of its campaign against revolutionary Arab leftists and militant tribesmen in Dhofar and the interior. Domestic progress is calculated to weaken popular support for the dissidents, and military successes, to strengthen the government's control throughout the country. Acceptance of Oman in the Arab world requires a demonstration of decreasing reliance on Great Britain.

BIBLIOGRAPHY. P.S. ALLFREE, *Warlords of Oman* (1967), Memoirs of a British officer who served in al-Buraymī and Oman in the 1950s and early 1960s; JAMES T. and MABEL BENT, *Southern Arabia* (1900), a travel account by a British couple who visited Muscat and Dhofar in the late 19th century; PAUL HARRISON, *Doctor in Arabia* (1940), experiences of an American medical missionary in Oman; ROBERT LANDEN, *Oman Since 1856* (1967), primarily a history, but much information of a general nature is included; J.G. LORIMER, *Gazetteer of the Persian Gulf, 'Oman, and Central Arabia*, 2 vol. (1908–15), a mine of historical and geographical data drawn from British official records, some of which is inaccurate or out of date; JAMES MORRIS, *Sultan in Oman* (1957), report by a British journalist who accompanied Sultan Saʿid on a trip from Dhofar to al-Buraymī; WENDELL PHILLIPS, *Unknown Oman* (1966) and *Oman: A History* (1967), works by an American involved in archaeology and oil concessions who, as a close associate of Sultan Saʿid, travelled extensively in Oman—also contains geographical material; GEORGE RENTZ (ed.), *Oman and the Southern Shore of the Persian Gulf* (1952), a study that exploits a number of Arabic sources, written and oral; WILFRED THESIGER, *Arabian Sands* (1959), journeys by one of the great explorers of Arabia, who travelled through Dhofar and the western regions of Oman; BERTRAM THOMAS, *Alarms and Excursions in Arabia* (1931), travels and experiences of an Englishman who was financial adviser to the Sultan of Muscat in the 1920s; J. RAYMOND WELLSTED, *Travels in Arabia* (1838), an account by one of the first Westerners to penetrate the interior of Oman; SAMUEL M. ZWEMER, *Arabia: The Cradle of Islam* (1900), by an American missionary who worked in Muscat and travelled in the interior.

(G.Re.)

Oncopod

The term oncopod (sometimes Pararthropoda) is the collective name for three groups of animals usually considered classes: Onychophora and Tardigrada, which are free-living forms, and Pentastomida, which are parasitic (*i.e.*, deriving nourishment from the body of another living animal). The groups are of interest chiefly because they bear strong resemblances to primitive annelidan forms as well as to the phylum Arthropoda (*e.g.*, insects, crustaceans)—the most highly evolved invertebrates.



Body plans of (A) tardigrade, *Echiniscus scrofa;* (B) peripatus, *Peripatopsis capensis;* (C) pentastomid, *Linguatula taenoides.*

The Onychophora (meaning "claw bearers") were once thought to be intermediate between the annelid worms and the arthropods. They comprise fewer than 100 species and exhibit little diversity of form. Onychophores, which vary in length from about 14 to 150 millimetres (about ½ to six inches), are elongated, terrestrial animals covered by a soft, furrowed skin. They occur widely in tropical and subtropical regions of the world and, because they are unable to control water loss, must live in a moist habitat. *Peripatus,* a typical genus, is the best known of the group.

The Tardigrada, the so-called bear animalcules, or water bears, comprise about 350 species, which are distributed worldwide, mostly in damp or aquatic habitats. Their length ranges from about 0.05 to 1.2 millimetres (0.002 to 0.05 inch).

Pentastomids, comprising some 70 species, are wormlike parasites up to nine centimetres (about four inches) in length that live in the respiratory systems of vertebrate hosts. Although they are mostly tropical or subtropical, those with homoiothermic, or warm-blooded, hosts may also be found in cold regions. A few species are of medical interest because man occasionally serves as the host.

Natural history. *Reproduction and life cycle.* The sexes are separate; *i.e.*, male and female sex organs do not occur in one animal. Except in many tardigrades, in which the sexes are similar in appearance, males are smaller than females. The sex organs (gonads) are always on the back side and may be paired or single. The gonad cavity is a remnant of the coelom, the general body cavity. The genital pore, from which the eggs or sperm are discharged, is located either ventrally (*i.e.*, on the lower side) in front of the anus (onychophores, some tardigrades, females of a few pentastomids) or in the anterior, or front, part of the trunk (pentastomids); in some tardigrades the genital ducts open into the rectum.

Transfer of sperm

Copulation, with internal fertilization of the egg, is the general rule, though modifications occur. In many onychophores the male deposits a spermatophore, or sperm case, on the skin of the female; the skin then dissolves,

allowing the sperm to reach the ovaries via the hemocoel (*i.e.*, a network of spaces between the internal organs). In some tardigrades fertilization is external. Oncopods are oviparous (*i.e.*, the young hatch from eggs that have been laid), with the exception of most onychophores. In most viviparous (*i.e.*, giving birth to live young) onychophores, the embryos receive nourishment from the mother. Development is direct—*i.e.*, there is no radical change in form except in size—in tardigrades and onychophores, but larval stages occur in pentastomids. Many pentastomids must parasitize an intermediate host—usually a vertebrate prey of the adult pentastomid's host—in order to complete their life cycles; others are thought to have a single host. Molting, or shedding of the skin, occurs throughout the lives of all oncopods.

Ecology and behaviour. Because onychophores, which are terrestrial, cannot control water loss, they cannot tolerate dry habitats. They live among litter in forests, under stones or fallen logs, inside galleries of fallen logs, inside nests of termites, or in crevices and galleries in the soil, sometimes to depths of more than one metre (about three feet). Onychophores have a remarkable ability to squeeze themselves through narrow passages, an adaptation that permits them to find shelters of satisfactory humidity and safety. Onychophores are carnivorous (*i.e.*, flesh eaters); they use their jaws to open captured prey and suck out the juices. A quick-hardening slime squirted from the oral papillae, fingerlike projections near the mouth, is used for defense or to subdue prey. Onychophores avoid light and are usually well concealed from casual observation. Touch is important, and onychophores are provided with tactile spines that are sensitive to air currents.

A few tardigrades are marine; others are found in freshwaters, and many occur in terrestrial habitats in which droplets or a film of water are available. Thus, terrestrial species occur in mosses, lichens, or other cushiony plants; they are also regular inhabitants of the soil and may be locally abundant in a given area. Most are herbivorous (*i.e.*, plant eaters), piercing individual cells with their stylets (spearlike structures near the mouth) and sucking out the cell contents. A few tardigrades are predacious carnivores. Species normally living in habitats subject to sudden drying or freezing are able to enter into a resistant state. Some species may encyst within the molted skin. In such states tardigrades are easily distributed by wind, water, and other means.

Modes of locomotion

Onychophores and tardigrades move by using their legs, the body remaining straight. In walking, the onychophore foot usually remains lifted, the weight of the body resting on the creeping pads; the claws are used only on slippery surfaces. Many tardigrades use only the three anterior pairs of legs in walking; the fourth trails behind.

Form and function. *General features.* A common feature of oncopods is the presence, in some stage of the life history, of oncopodia—*i.e.*, locomotory lobopodia, which are nonjointed, usually hollow, paired appendages with claws at the tips. The onychophoran oncopodium has a conical basal portion with a furrowed, papillose (*i.e.*, with many short protuberances) skin; a specialized distal, or end, portion comprises a broad spiny pad and a foot bearing the pair of terminal claws. Ventrally, at the base of the leg, is an excretory pore. Onychophores have two pairs of modified oncopodia: the oral papillae, at the tip of which open the ducts of large slime glands, and the jaws.

Tardigrade oncopodia are simpler and conical or cylindrical in shape; they are often divided into a thinner basal portion that may telescope into a broader basal portion, ending in two, four, or more claws. In pentastomids, typical oncopodia bearing two claws and similar to those of tardigrades occur only in embryos and some larvae.

Organ systems. Of the three oncopod groups, only the Onychophora have well-developed organ systems. Both the Tardigrada and Pentastomida have undergone reductions in, or even loss of, some organ systems, a consequence of small size in the former and of parasitism in the latter. Thus, neither group has respiratory or circulatory organs; excretory organs, in addition, are found neither in pentastomids nor in many tardigrades.

A cuticle, or skin, secreted by epidermis, or surface cells, and molted periodically, is present in all oncopods. It is generally soft, sometimes with local thickenings, and is heavily sclerotized, or hardened, only in structures such as the claws of oncopodia. Chitin, a tough, resistant material, is a component of the cuticle in onychophores and pentastomids; in tardigrades the cuticle consists of albumin, a protein. In onychophores the epidermis forms numerous papillae over the body, many of which are tipped by a sensory spine; under the epidermis is a strong fibrous layer.

The muscle fibres of onychophores and tardigrades are smooth (*i.e.*, nonstriated), while those of pentastomids are striated. Beneath the skin of onychophores are circular, diagonal, and longitudinal muscle layers as occur in annelids (segmented worms); in addition, transverse muscles also occur. Pentastomids have a similar muscle organization but do not have a diagonal layer of muscle. Tardigrades have discrete muscle fibres segmentally arranged.

Structure of the nervous system

The nervous system consists of a brain dorsal to the pharynx and connected by a pair of nerves encircling the pharynx to the ventral nerve cords. These usually form a ventral chain of ganglia, or nerve centres. Tardigrades have a brain, often containing a pair of eyespots, and a ventral chain of five ganglia. In pentastomids there may be a short ventral chain, or a single ventral ganglion; the brain halves shift ventrally, leaving a band of connecting nerve fibres. Onychophores have a well-developed brain and two widely separated, nonganglionated ventral nerve cords that unite caudally (*i.e.*, toward the tail) above the rectum, a very primitive condition. The eyes of Onychophora are similar in structure to those of some annelids.

In oncopods the gut is usually a straight tube. The foregut and hindgut are lined with cuticle. The mouth is usually ventral, but at the anterior tip in most tardigrades. In onychophores the mouth is bordered by a row of oral lobes (lips). The salivary glands are modified segmental organs of the oral papillae segment. In the mouth cavity are two pairs of jaw blades, which are the modified claws of a pair of reduced oncopodia. In tardigrades the mouth is small and round; a pair of stylets may protrude through it. A sucking pharynx and salivary glands are also present in tardigrades. The anus is terminal.

The body cavity is a hemocoel. In onychophores the dorsal portion, separated as a pericardium, contains the long tubular heart provided with segmental pairs of ostia, or openings.

The respiratory system of onychophores consists of numerous, irregularly distributed pits, from which many thin tracheae, or breathing tubes, originate. The tracheae may penetrate deep into the body and may branch.

Onychophores have a pair of excretory segmental organs on each leg-bearing segment. A ciliated (*i.e.*, with hairlike structures) funnel leads to the outside at the excretory pore. Each segmental organ eliminates water and other substances, but nitrogenous excretion, in the form of uric acid, takes place into the midgut. Many tardigrades have three excretory rectal glands.

Early fossil forms

Evolution. *Paleontology.* Two fossils, probably both marine, are directly relevant to the oncopods. The first, *Xenusion auerswaldae,* possibly from the Precambrian (more than 570,000,000 years ago), may well represent a lobopod stage (*i.e.*, with fingerlike body extensions) in the evolution of oncopods and arthropods. The clearly segmented body is provided with relatively large, annulate (ringed) lobopodia; the presence of claws could not be ascertained, and the cephalic (head) end was not preserved. The second fossil, *Aysheaia pedunculata,* from the Middle Cambrian (about 535,000,000 years ago), is generally considered to be an onychophore. The body has ten or 11 postcephalic segments, each with four rings bearing skin papillae, and a pair of ringed oncopodia provided with six apical (*i.e.*, at the apex, or tip) claws. The head may have had antennae and, probably, oral papillae; jaws were not detectable.

Historical development. The evolution of oncopods, as of other very old groups with few or no fossils, is a speculative subject, on which no consensus has yet been reached. Oncopods, as well as arthropods, originated in Precambrian times from marine, bottom-dwelling forms that may also have been ancestral to modern annelids. The adaptation that resulted in the evolution of oncopods and arthropods was the acquisition of lobopodia, locomotory appendages that could work independently of the waves of contraction of the body. The acquisition of lobopodia led to the dissolution of separate coelomic compartments and to the formation of a hemocoel; it also permitted the development of a firmer cuticle, which made molting necessary. The development of the cuticle also led to the loss of external cilia.

Tardigrades evolved toward small size, with a consequent simplification of body organization. Some of their primitive characteristics—*e.g.*, the cuticle of protein, brain, and absence of specialized legs on the head—suggest that their ancestors diverged very early from the oncopod–arthropod stem. Tardigrades show interesting similarities to some aschelminths (roundworms).

The evolution of pentastomids has been greatly influenced by their endoparasitic (internal parasitic) mode of life; it also entailed simplifications. Their chitinous cuticle, striated muscles, and brain form indicate that they had ancestors more active and advanced than those of tardigrades; it is possible that they shared a common ancestor with the arthropods. Pentastomids could have evolved from free-living forms, first as parasites of fishes, then becoming parasites of terrestrial vertebrates as they appeared in the Devonian Period (345,000,000 to 395,000,000 years ago).

The formation of a strong "skeleton" under a soft epidermis is a characteristic of onychophore evolution and permits the body to be deformed. Among oncopods, onychophores have the most extensive combinations of primitive features—*e.g.*, structure of nerve cords, persistence of cilia, smooth muscles—with oncopodan characters or basic arthropodan characters (*e.g.*, chitin and molting, hemocoel, gonads, embryonic development) or both. They also developed specializations of their own (*e.g.*, dermis, tracheae, viviparity). Both onychophores and arthropods could have arisen in Precambrian seas from a *Xenusion*-like ancestor, arthropods acquiring sclerites (*i.e.*, hard body plates) and joints and onychophores developing a strong dermis while keeping a thin cuticle. According to some authorities, the myriapod–insect line originated from onychophore ancestors, independently from the other arthropods.

Classification. *Distinguishing taxonomic features.* The various groups considered together as oncopods have several common features: similar paired locomotory appendages called oncopodia at some stage in the life cycle; a body cavity (hemocoel); a cuticle (skin) secreted by surface cells and shed periodically (molting); a gut that is usually a straight tube; and separate sexes and gonads. The groups also differ: only the onychophores have well-developed organ systems; those of pentastomids and tardigrades are reduced or lacking. The pentastomids are parasites; the other two groups are free-living.

Annotated classification.

ONCOPOD

General term commonly used to refer to three arthropod-like animal groups; tardigrades, pentastomids, and onychophores.

Class Onychophora

Length about 14–150 mm; predatory on other small organisms; cuticle containing chitin, a tough, complex nitrogen-containing carbohydrate; genera include *Peripatus, Peripatopsis;* fewer than 100 species in tropics and subtropics.

Class Pentastomida (or Linguatulida)

Parasites of vertebrates; length up to 9 cm; cuticle containing chitin; about 70 species; mostly tropical or subtropical.

Order Cephalobaenida

Genera include *Cephalobaena, Raillietiella, Reighardia.*

Order Porocephalida

Genera include *Porocephalus, Linguatula.*

Class Tardigrada (water bears)

Size 1 mm or less; found in marine waters and freshwaters and in various terrestrial habitats (*e.g.*, soil); mostly plant eaters; cuticle composed of protein; worldwide distribution; about 350 species.

Order Heterotardigrada

Genera include *Echiniscus, Batillipes.*

Order Eutardigrada

Genera include *Macrobiotus, Hypsibius, Milnesium.*

Critical appraisal. The name oncopod is used for convenience to refer to these three groups of invertebrates of uncertain relationship to the phylum Arthropoda. Some authorities have elevated the Onychophora to the rank of phylum and regard the Tardigrada and Pentastomida either as Proarthropoda or as orders of the arthropod class Arachnida (*e.g.*, spiders). Other authorities regard the three groups as distinct phyla.

BIBLIOGRAPHY. L. CUENOT, "Onychophores, Tardigrades, Pentastomides," in P.P. GRASSE (ed.), *Traité de zoologie,* vol. 6, pp. 1–75 (1949), comprehensive, in French; R. HEYMONS, "Pentastomida," in *Bronn's Klassen und Ordnungen des Tierreichs,* vol. 5, bk. 1–4 (1935), classic, in German, still the best monograph of the group; H.R. HILL, "Annotated Bibliography of the Linguatulida," *Bull. Sth. Calif. Acad. Sci.,* 47:56–73 (1948); A. KAESTNER, *Invertebrate Zoology,* vol. 2, trans. and adapted from the German by H.W. and L.R. LEVI (1968), an excellent graduate-level textbook; E. MARCUS, "Tardigrada," in *Bronn's Klassèn und Ordnungen des Tierreichs* (1929), a classic, comprehensive monograph in German; G. RAMAZZOTTI, "Il Phylum Tardigrada," *Memorie Ist. Ital. Idrobiol.,* 14:1–595 (1962) and suppl. no. 1, *ibid.,* 19:101–212 (1965), a modern monograph; O.W. TIEGS and S.M. MANTON, "The Evolution of the Arthropoda," *Biol. Rev.,* 33:255–337 (1958), a scholarly discussion with extensive bibliography that defends the view that myriapods and insects originated from onychophores; F. ZACHER, "Onychophora," in K. KUEKENTHAL and T. KRUMBACH (eds.), *Handbuch der Zoologie,* vol. 3, pp. 79–138 (1933), a classic general presentation, in German.

(C.G.F.)

Onega, Lake

Lake Onega (Onezhskoye Ozero in Russian; spelled Onezhskoje Ozero in the transliteration system of the Akademija Nauk) lies in the northwest part of the European portion of the Soviet Union. It is situated mainly in the Karelian Autonomous Soviet Socialist Republic (Karelskaya Avtonomnaya Sovetskaya Sotsialisticheskaya Respublika), but its southern shores are in the Leningrad and Vologda *oblasti* (regions) of the Russian Soviet Federated Socialist Republic. It covers an area of 3,753 square miles (9,720 square kilometres), being the second largest lake in Europe after Lake Ladoga.

Geological structure

The hollow of the lake was formed by movements of the Earth's crust, but Quaternary glaciers (10,000 to 2,500,000 years ago) elongated it from northwest to southeast. It is 154 miles long; its greatest width is 50 miles; and its greatest depth is about 380 feet, the average being about 100 feet. The shores to the north and northwest are high and rocky, built of layered granite and covered with forest. There are deep bays at Petrozavodsk, Kondopoga, and Povenots. The southern shores are narrow, sandy, and often marshy or flooded. The floor of the lake in the north and northwestern part is broken and has deep depressions, the deepest one being more than 300 feet in places and extending out 380 feet to the west of Bolshoy Lelikov Island. The centre of the lake is deepest toward the west, while the floor of the southern part is flat, reaching depths of 150 feet only in its western part. The depth in the northeast is similar, though Povents Bay reaches 300 feet in places. The floor is rocky and sandy at the shores and, in the open part, covered with silt. Onega has about 1,650 islands covering a total of about 100 square miles, mostly in the northern and northwestern bays.

Fifty rivers enter Onega, the largest being the Shuya and Suna in the northwest and the Vodla in the east, together accounting for 60 percent of the water inflow. In the southeast and east are the Andoma, Vytegra, and Megra rivers. Lake Onega itself empties into the River Svir. The

rivers supply 75 percent of the lake's water, and the rest comes from the atmosphere. It loses 84 percent of its water output through evaporation. The water level reaches its highest point in the summer, and its lowest in March–April, varying about 24 inches annually. Water levels at the northern end of the lake may differ by 20 to 26 inches from those at the southern end. The water circulates in a twisting pattern within the lake because of differences of temperature between the coastal and the open regions. The rate of flow at the surface is eight to ten inches per second. During autumn gales, waves sometimes reach 14 or 15 feet.

The region has a cold climate. Air temperatures in February average 10° F (−12° C) in the north and 14° F (−10° C) in the south, while July temperatures average 61° F (16° C). The absolute minimum air temperature is −54° F (−48° C), and the absolute maximum is 93° F (34° C). The highest temperature of the water at the surface in the open part of the lake is 64° or 68° F (18° or 20° C) and 75° or 81° F (24° or 27° C) in the bays. The bottom layers of the water are much colder, varying from about 36° F (2° C) in the winter to 39° or 43° F (4° or 6° C) in the summer. The coastal parts and the small bays begin to freeze at the end of November, and the deeper central parts in the middle of January, although in some years the central parts do not freeze. Thawing begins at the end of April.

The colour of Onega's water is dark yellowish brown in the open part and grayish brown along the shores. Its average transparency in the open lake and deep bays is 16 to 28 feet, and three to 12 feet near the shore. The lake contains 47 species of fish, including ryapushka (a small member of the salmon family), smelt, burbot (a freshwater variant of the cod), bream, pike, perch, roach, and salmon. The fish are of considerable economic value.

Onega as a transport route

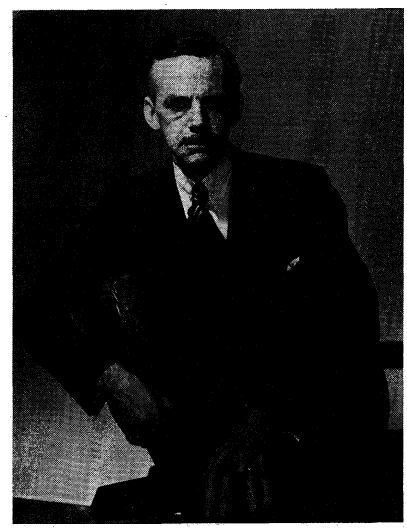
Lake Onega is connected with the Baltic and White seas by the White Sea–Baltic Canal and with the basin of the Volga River by the Volga–Baltic Waterway, which enable it to play an important part in both internal and international transportation. Goods are shipped over this route from Finland, Sweden, Denmark, and East Germany to points in the east and north. For protection against storms, a bypass canal has been dug along the south and southeastern shores from the mouth of the River Vytegra to the source of the Svir. The cities of Petrozavodsk, Kondopoga, and Medvezhyegorsk are on Lake Onega. The island of Kizhi houses the internationally known architectural organization of that name.

(B.B.Bo.)

O'Neill, Eugene

Eugene O'Neill was the first American dramatist to regard the stage as a literary medium and the only American playwright ever to receive the Nobel Prize for Literature. Through his efforts, the American theatre grew up during the 1920s, developing into a cultural medium that could take its place with the best in American fiction, painting, and music. Until his *Beyond the Horizon* was produced, in 1920, Broadway theatrical fare, apart from musicals and an occasional European import of quality, had consisted largely of contrived melodrama and farce. O'Neill saw the theatre as a valid forum for the presentation of serious ideas. Imbued with the tragic sense of life, he aimed for a contemporary drama that had its roots in the most powerful of ancient Greek tragedy—a drama that could rise to the emotional heights of Shakespeare. For more than 20 years, both with such masterpieces as *Desire Under the Elms, Mourning Becomes Electra,* and *The Iceman Cometh* and by his inspiration to other serious dramatists, O'Neill set the pace for the blossoming of the Broadway theatre.

Early life. O'Neill was born into the theatre. His father, James O'Neill, was a successful touring actor in the last quarter of the 19th century whose most famous role was that of the Count of Monte Cristo in a stage adaptation of the Alexandre Dumas *père* novel. His mother, Ella, accompanied her husband back and forth across the country, settling down only briefly for the birth of her first son, James, Jr., and of Eugene.



O'Neill.
By courtesy of the Collection of American Literature, Yale University Library

Shattering experiences of family life

Eugene, who was born in a hotel in New York City, on October 16, 1888, spent his early childhood in hotel rooms, on trains, and backstage. Although he later deplored the nightmare insecurity of his early years and blamed his father for the difficult, rough-and-tumble life the family led—a life that resulted in his mother's drug addiction—Eugene had the theatre in his blood. He was also, as a child, steeped in the peasant Irish Catholicism of his father and the more genteel, mystical piety of his mother, two influences, often in dramatic conflict, which account for the high sense of drama and the struggle with God and religion that distinguish O'Neill's plays.

O'Neill was educated at boarding schools—Mt. St. Vincent in the Bronx and Betts Academy in Stamford, Connecticut. His summers were spent at the family's only permanent home, a modest house overlooking the Thames River in New London, Connecticut. He attended Princeton University for one year (1906–07), after which he left school to begin what he later regarded as his real education in "life experience." The next six years very nearly ended his life. He shipped to sea; lived a derelict's existence on the waterfronts of Buenos Aires, Liverpool, and New York; submerged himself in alcohol; and attempted suicide. Recovering briefly at the age of 24, he held a job for a few months as a reporter and contributor to the poetry column of the *New London Telegraph* but soon came down with tuberculosis. Confined to the Gaylord Farm Sanitarium in Wallingford, Connecticut, for six months (1912–13), he confronted himself soberly and nakedly for the first time and seized the chance for what he later called his "rebirth." He began to write plays.

Entry into theatre. O'Neill's first efforts were awkward melodramas, but they were about people and subjects that had, up to that time, been in the province of serious novels and not considered fit subjects—prostitutes, derelicts, lonely sailors, God's injustice to man—for presentation on the American stage. A theatre critic persuaded his father to send him to Harvard to study with George Pierce Baker in his famous playwriting course. Although what O'Neill produced during that year (1914–15) owed little to Baker's academic instruction, the chance to work steadily at writing set him firmly on his chosen path.

O'Neill's first appearance as a playwright came in the summer of 1916, in the quiet fishing village of Provincetown, Massachusetts, where a group of young writers and painters had launched an experimental theatre. In their tiny, ramshackle playhouse on a wharf, they produced his one-act sea play *Bound East for Cardiff.* The talent inherent in the play was immediately evident to the group, which that fall formed the Playwrights' Theater in Greenwich Village. Their first bill, on November 3, 1916, included *Bound East for Cardiff*—O'Neill's New York debut. Although he was only one of several writers whose

plays were produced by the Playwrights' Theater, his contribution within the next few years made the group's reputation. Between 1916 and 1920, the group produced all of O'Neill's one-act sea plays, along with a number of his lesser efforts. By the time his first full-length play, *Beyond the Horizon*, was produced on Broadway, February 2, 1920, at the Morosco Theater, the young playwright already had a small reputation.

Beyond the Horizon impressed the critics with its tragic realism, won for O'Neill the first of four Pulitzer prizes in drama—others were for *Anna Christie* (1922), *Strange Interlude* (1928), and *Long Day's Journey into Night* (posthumously, 1957)—and brought him to the attention of a wider theatre public. For the next 20 years his reputation grew steadily, both in the United States and abroad; after Shakespeare and Shaw, O'Neill became the most widely translated and produced dramatist.

Period of the major works. O'Neill's capacity for and commitment to work were staggering. Between 1920 and 1943, he completed 20 long plays—several of them double and triple length—and a number of shorter ones. He wrote and rewrote many of his manuscripts half a dozen times before he was satisfied, and he filled shelves of notebooks with research notes, outlines, play ideas, and other memoranda. His most distinguished short plays include the four early sea plays, *Bound East for Cardiff*, *In the Zone*, *The Long Voyage Home*, and *The Moon of the Caribbees*, which were written between 1913 and 1917 and produced in 1924 under the overall title *S.S. Glencairn; The Emperor Jones* (about the disintegration of a Negro Pullman porter turned tropical-island dictator); and *The Hairy Ape* (about the disintegration of a displaced steamship coal stoker).

Evolving themes of tragedy and conflict

O'Neill's plays were written from an intensely personal point of view, deriving directly from the scarring effects of his family's tragic relationships—his mother and father, who loved and tormented each other; his older brother, who loved and corrupted him and died of alcoholism in middle age; and O'Neill himself, caught and torn between love for, and rage at, all three.

Among his most celebrated long plays is *Anna Christie*, perhaps the classic American example of the ancient "harlot with a heart of gold" theme; it became an instant popular success. O'Neill's serious, almost solemn treatment of the struggle of a poor Swedish-American girl to live down her early, enforced life of prostitution and to find happiness with a likable but unimaginative young sailor is his least complicated tragedy. He himself disliked it from the moment he finished it, for, in his words, it had been "too easy."

The first full-length play in which O'Neill successfully evoked the starkness and inevitability of Greek tragedy that he felt in his own life was *Desire Under the Elms*. Drawing on Greek themes of incest, infanticide, and fateful retribution, he framed his story in the context of his own family's conflicts. This story of a lustful father, a weak son, and an adulterous wife who murders her infant son was told with a fine disregard for the conventions of the contemporary Broadway theatre. Because of the sparseness of its style, its avoidance of melodrama, and its total honesty of emotion, the play was acclaimed immediately as a powerful tragedy and has continued to rank among the great American plays of the 20th century.

Innovative playwriting and staging techniques

In *The Great God Brown* O'Neill dealt with a major theme that he expressed more effectively in later plays—the conflict between idealism and materialism. Although the play was too metaphysically intricate to be staged successfully in 1926, it was significant for its symbolic use of masks and for the experimentation with expressionistic dialogue and action—devices that since have become commonly accepted both on the stage and in motion pictures. In spite of its confusing structure, the play is rich in symbolism and poetry, as well as in daring technique, and it became a forerunner of avant-garde movements in American theatre.

O'Neill's innovative writing continued with *Strange Interlude*. It was revolutionary in style and length: when first produced, it opened in late afternoon, broke for a dinner intermission, and ended at the conventional hour. Techniques new to the modern theatre included spoken asides, or soliloquies, to express the characters' hidden thoughts. The play is the saga of Everywoman, who ritualistically acts out her roles as daughter, wife, mistress, mother, and platonic friend. Although it was innovative and startling in 1928, its obvious Freudian overtones have rapidly dated it.

One of O'Neill's enduring masterpieces, *Mourning Becomes Electra*, represents the playwright's most complete use of Greek forms, themes, and characters. Based on the *Oresteia* trilogy by Aeschylus, it was itself three plays in one. To give the story contemporary credibility, O'Neill set the play in the New England of the Civil War period, yet he retained the forms and the conflicts of the Greek characters: the heroic leader returning from war; his adulterous wife, who murders him; his jealous, repressed daughter who avenges him through the murder of her mother; and his weak, incestuous son who is goaded by his sister first to matricide and then to suicide.

His one comic vision

Following a long succession of tragic visions, O'Neill's only comedy, *Ah, Wilderness!* appeared on Broadway in 1933. Written in a lighthearted, nostalgic mood, the work was inspired in part by the playwright's mischievous desire to demonstrate that he could portray the comic as well as the tragic side of life. Significantly, the play is set in the same place and period, a small New England town in the early 1900s, as his later tragic masterpiece, *Long Day's Journey into Night*. Dealing with the growing pains of a sensitive, adolescent boy, *Ah, Wilderness!* was characterized by O'Neill as "the other side of the coin," meaning that it represented his fantasy of what his own youth might have been, rather than what he believed it to have been (as dramatized later in *Long Day's Journey into Night*).

The Iceman Cometh, the most complex and perhaps the finest of the O'Neill tragedies, followed in 1939, although it did not appear on Broadway until 1946. Laced with subtle religious symbolism, the play is a study of man's need to cling to his hope for a better life, even if he must delude himself to do so.

Even in his last writings, O'Neill's youth continued to absorb his attention. The posthumous production of *Long Day's Journey Into Night* brought to light an agonizingly autobiographical play, one of O'Neill's greatest. It is straightforward in style but shattering in its depiction of the agonized relations between father, mother, and two sons. Spanning one day in the life of a family, the play strips away layer after layer from each of the four central figures, revealing the mother as a defeated drug addict, the father as a man frustrated in his career and failed as a husband and father, the older son as a bitter alcoholic, and the younger son as a tubercular, disillusioned youth with only the slenderest chance for physical and spiritual survival.

O'Neill's tragic view of life was perpetuated in his relationship with the three women he married—two of whom he divorced—and with his three children. His elder son, Eugene O'Neill, Jr. (by his first wife, Kathleen Jenkins), committed suicide at 40, while his younger son, Shane (by his second wife, Agnes Boulton), drifted into a life of emotional instability. His daughter, Oona (also by Agnes Boulton), was cut out of his life when, at 18, she infuriated him by marrying Charlie Chaplin, who was O'Neill's age.

Until some years after his death in 1953, O'Neill, although respected in the U.S., was more highly regarded abroad. Sweden, in particular, always held him in high esteem, partly because of his publicly acknowledged debt to the influence of the Swedish playwright August Strindberg, whose tragic themes often echo in O'Neill's plays. In 1936 the Swedish Academy gave O'Neill the Nobel Prize for Literature, the first time the award had been conferred on a U.S. playwright.

Unfinished plays and unwritten cycles

O'Neill's most ambitious project for the theatre was one that he never completed. In the late 1930s he conceived of a cycle of 11 plays, to be performed on 11 consecutive nights, tracing the lives of an American family from the early 1800s to modern times. He wrote scenarios and

outlines for several of the plays and drafts of others but completed only one play in the cycle—*A Touch of the Poet*—before a crippling illness ended his ability to hold a pencil. An unfinished rough draft of another of the cycle plays, *More Stately Mansions*, was published in 1964 and produced three years later on Broadway, in spite of written instructions left by O'Neill that the incomplete manuscript be destroyed after his death.

O'Neill's final years were spent in grim frustration. Unable to work, he longed for his death and sat waiting for it in a Boston hotel, seeing no one except his doctor, a nurse, and his third wife, Carlotta Monterey. As broken and tragic a figure as any he had created for the stage, O'Neill died on November 27, 1953.

MAJOR WORKS

ONE ACT PLAYS OF THE SEA: *Bound East for Cardiff* (performed 1916); *The Long Voyage Home* (performed 1917), later used as the title of a film version of O'Neill's plays of the sea; *Ile* (performed 1917); *In the Zone* (performed 1917); *The Moon of the Caribbees* (performed 1918), published in a collection, *The Moon of the Caribees, and Six Other Plays of the Sea* (1919), which included the first book publication of the above plays plus *The Rope* (performed 1918), and *Where the Cross Is Made* (performed 1918). This same collection was published in 1940 as *The Long Voyage Home: Seven Plays of the Sea.*

LONGER PLAYS: *Beyond the Horizon* (performed and published 1920); *The Emperor Jones* (performed 1920, published 1921); *Anna Christie* (performed 1921, published 1922); *The Hairy Ape* (1922); *All God's Chillun Got Wings* (1924); *Desire Under the Elms* (performed 1924, published 1925); *The Great God Brown* (1926); *Marco Millions* (performed 1928, published 1927); *Strange Interlude* (1928), a two-part play in 9 acts; *Mourning Becomes Electra* (1931), a trilogy comprising *Homecoming, The Hunted,* and *The Haunted; Ah, Wilderness!* (1933), O'Neill's only comedy; *The Iceman Cometh* (written 1939, performed and published 1946); *A Touch of the Poet* (written 1935–42; performed and published posthumously, 1957), third of a projected cycle of 11 plays to be collectively entitled *A Tale of the Possessors, Self-Dispossessed; Long Day's Journey Into Night* (written 1939–41; performed and published posthumously, 1956); *A Moon for the Misbegotten* (written 1943; performed 1957, published 1952); *Hughie* (written 1941; first performed 1964; published 1959, one of a projected cycle of one-act plays, to have been collectively entitled *By Way of Orbit*); *More Stately Mansions* (written 1935–41; performed 1962; published 1964).

The handiest source for all the plays listed from the beginning of this Major Works section through *Mourning Becomes Electra* is *The Plays of Eugene O'Neill,* 3 vol. (1951). Some of the other plays have been published as separate volumes.

BIBLIOGRAPHY. ARTHUR and BARBARA GELB, *O'Neill* (1962), is the definitive critical biography; see also DORIS ALEXANDER, *The Tempering of Eugene O'Neill* (1962); AGNES BOULTON, *Part of a Long Story* (1958), a memoir of their married life, by O'Neill's second wife; HELEN DEUTSCH and STELLA HANAU, *The Provincetown* (1931), a history of the Provincetown Players, who produced O'Neill's early works in New York; BARRETT H. CLARK, *Eugene O'Neill: The Man and His Plays,* rev. ed. (1936, reprinted 1967), a brief biography and critical analysis of the plays by a contemporary of O'Neill; and RALPH SANBORN and BARRETT H. CLARK (comps.), *A Bibliography of the Works of Eugene O'Neill* (1931, reprinted 1965).

(B.Ge./A.Ge.)

Ontario

Ontario, the second largest province of Canada, occupies the strip of the Canadian mainland lying between Hudson and James bays, on the north, and the St. Lawrence Rivers an area of 412,582 square miles (1,068,582 square by Quebec Province and to the west by Manitoba, it covers an area of 412,582 square miles (1,068,582 square kilometres). It is the most populous Canadian province, with more than 7,700,000 inhabitants representing more than one-third of the country's total by the early 1970s.

Ontario is also the nation's wealthiest province, having a substantial share of the country's natural resources and its most mature and diversified industrial economy. It is at once Canada's economic pacemaker and a major force in national politics. Some Ontarians call it the "Empire Province," while to those Canadians living outside its boundaries, its pre-eminent position and the influence of Toronto, the provincial capital, have constituted a not infrequent source of regional resentment. (For associated physical features see GREAT LAKES; SAINT LAWRENCE RIVER; HUDSON BAY; for related city articles see TORONTO; OTTAWA; and DETROIT; and for a detailed discussion of history see CANADA, HISTORY OF.)

HISTORY

The earliest known Indian inhabitants of the Ontario region included the agricultural Huron, Tobacco, and Erie tribes of the south and the hunting groups of the Algonquin, Ojibwa, and Cree of the north. The French explorer Étienne Brûlé was the first known white man to travel among them, doing so on an expedition to the Ottawa River in 1610–11. He was shortly followed by Samuel de Champlain and other French explorers, fur traders, and missionaries. In 1648–49 the southern tribes were dispersed when the Iroquois destroyed the Jesuit mission at Ft. Ste.-Marie, and France established Ft. Frontenac (present-day Kingston) in 1673 to begin the military protection of its westward-spreading fur empire. When Canada was ceded to Great Britain in 1763, however, no French colonization had taken place, except for a small farming settlement in the Detroit area.

Early European settlement

The Quebec Act of 1774 established Ontario as part of an extended colony ruled from Quebec. During the American Revolution, the region was a base for Loyalist and Indian attacks upon the American frontier, and in 1784 it was settled by approximately 10,000 Loyalists and those of the Iroquois Indians who had fought for the British. The Constitutional Act of 1791 divided Quebec colony into Lower Canada, with a French majority, and the new Loyalist province of Upper Canada. Upper Canada received representative government; provision was made for the support of the colonial administration and an established church by substantial land endowments called the Crown and Clergy Reserves.

John Graves Simcoe, the vigorous first lieutenant governor of Upper Canada, supervised the introduction of English legal and local government practices, laid out the land-granting pattern, supported the construction of trunk roads, and fixed the capital at York (now Toronto). His policy of welcoming massive American immigration was a source of tension between the newcomers and the established anti-American Loyalists, a rift that deepened during the War of 1812.

From 1815 to 1841 the province was dominated by a conservative coalition that was known as the "Family Compact" because it was alleged to be a self-perpetuating elite, tied together by family relationships and intermarriage. It favoured the Anglican Church and the Crown and Clergy Reserve system. Reformers demanded responsible government, and in 1837 a radical minority led by William Lyon Mackenzie attempted an unsuccessful revolt. In 1841 the provinces of Upper and Lower Canada were united, and Upper Canada became known as Canada West. Responsible cabinet government was achieved with the formation of the Robert Baldwin–Louis Hippolyte Lafontaine ministry in 1848, and the present system of municipal government and the province's educational system were created. The 1850s brought the railways, the beginnings of industrialization, and the emergence of Toronto as a commercial rival for Montreal.

Achievement of provincial status

Political deadlock and the impetus of the new Ontario economy contributed to the movement for Canadian political union. Canadian federation—achieved in 1867—was brought about in large part by such Ontario politicians as the Conservative John A. Macdonald and the Liberal George Brown. Canada West became the province of Ontario, and the capital was located at Toronto. For a generation Ontario's government was headed by Oliver Mowat, the Liberal premier who won a boundary dispute with Manitoba that doubled the size of Ontario and confirmed the supremacy of provincial governments within their constitutionally-assigned powers.

In the 20th century, the chief concern of Ontario's governments has been the fostering of economic growth. The

province has been transformed into a multicultural society engaged in a highly industrialized economy. With the harnessing of Niagara Falls in 1882, Ontario entered into an energy revolution that also encompasses the power potential of the north, the St. Lawrence River, and uranium-generated thermal power. Since 1914, there have been major discoveries of gold, silver, iron, nickel, copper, and other minerals, as well as the expansion of the forest products industry, particularly in the development of pulp and paper. Tied to the swift emergence of the northern natural resources, the rise of manufacturing in southern Ontario was spurred by the two world wars to the production of a surplus for export. The provincial government has also been concerned with the growth of services, especially education, for its burgeoning population and ever-expanding cities.

THE LANDSCAPE

Northern Ontario. Ontario is composed of two regions of widely different character. Northern Ontario, as usually defined, lies north of a line drawn from the confluence of the Mattawa and Ottawa rivers southwest to the mouth of the French River on Georgian Bay. Most of the region's 350,000 square miles is part of the ancient Canadian Shield, characteristically marked with a profusion of lakes and rivers, muskeg (bogs), and densely forested, rocky, and rugged terrain. A low plateau, it is generally no more than 1,500 feet (450 metres) above sea level, although it contains the highest point in the province, Ogidaki Mountain, which rises to 2,183 feet (665 metres) on the eastern shore of Lake Superior. Although the region has always been an obstacle to transportation and communication, its rich mineral deposits, huge forest reserves, and the hydroelectric power potential of its swift rivers have made it a major source of the province's contemporary wealth.

The forests of the north

The vegetation is that of the boreal (northern, mountain) forest, and includes the black and white spruce, jack pine, tamarack, poplar, white birch, and balsam. The soils of the region, apart from peat, consist largely of brown podzolic (mineral-covered, leached) soils unsuitable for agriculture, except for two clay belts in the Timiskaming and Cochrane farming areas. The region contains parts of two major drainage basins—that of the Great Lakes to the south and of Hudson Bay to the north—separated by a band of higher land running from Lake of the Woods to Kirkland Lake. Major rivers of the northern system are the Severn, the Winisk, and the Albany, while the major rivers in the southern system are the Ottawa and French. Both Hudson Bay and James Bay are bordered by sparsely forested lowlands less than 500 feet above sea level, and, at the northern limit of the province, there is a band of tundra along Hudson Bay.

The climate varies from that of the districts close to the Great Lakes, which are frost free on more than 100 days a year, to the harsh climate of the Hudson Bay area, where the period free from frost may be as low as 40 days. At Thunder Bay on Lake Superior the mean temperature in January is 7° F (−14° C), and in July 64° F (18° C), the annual precipitation is 29 inches (737 millimetres), and the annual snowfall 85 inches (216 centimetres).

Southern Ontario. Though covering less than 10 percent of the area of the province, southern Ontario contains 90 percent of its population and is, in addition, the major urban-industrial region of Canada. It is a land of gentle relief; its lowest area on the Ottawa River is only 150 feet above sea level, and its highest point—on Blue Mountain south of Georgian Bay—is just over 1,700 feet. The east is divided from the rest of the region by an extension of the Canadian Shield, known as the Frontenac Axis. This crosses the St. Lawrence River east of Kingston and forms the Thousand Islands region. Along the southern edge of the shield lies a series of beautiful lake districts—including the Muskoka Lakes, the lakes of the Haliburton Highlands, and the Rideau chain—which are the province's best known resort areas. The most dramatic feature of the landscape is the Niagara Escarpment, running from Niagara Falls to the Bruce Peninsula; roads and rail lines pass through its notched valleys, and a nature trail runs along its length. The landforms of southern Ontario were shaped by glacial action, and most of the region consists of gently rolling plains. Both the Ottawa–St. Lawrence lowlands of eastern Ontario and the lands at the western tip of the Ontario Peninsula are, however, quite flat. The retreating glaciers left over most of the region a thick overburden of gray-brown podzolic soil, fertile in character, although sand plains are also found north of Lake Erie and along the eastern Lake Ontario shore. Rivers of the region are short, draining into the Great Lakes from the Western Ontario Upland and from the Oak Ridges Moraine north of Lake Ontario. Eastern Ontario, however, is drained chiefly by tributaries of the Ottawa River.

The Niagara Escarpment

The original natural vegetation of the area consisted of hardwood forests with great stands of white and red pines on the lighter soils, but during the 19th century land clearing and lumbering removed most of the original forest cover. Since the 1880s, the provincial government has engaged in farm woodlot and reforestation programs; it has also encouraged the formation of watershed-conservation authorities concerned with problems of soil erosion, drainage, forestry, and pollution.

The climate of the whole region is generally favourable to agriculture, although considerable local variation exists. The eastern section, away from the moderating influence of the lakes, tends to be cooler and more humid than the southern and southwestern zones. Ottawa receives 34 inches of rain and 86 inches of snow yearly, as compared to Toronto's 31 inches of rain and 54 inches of snow. The mean temperature in January for most of the southern region is about 25° F (−4° C), and for July, about 72° F (22° C). The Ottawa Valley, however, has means of 13° F (−11° C) and 69° F (21° C).

Pollution. In both regions of the province, industrialization and urbanization have created problems of pollution, the most acute of which are the death of Lake Erie and the polluted air of the Toronto urban complex. Air and water pollution associated with the mining and pulp and paper industries of the north has also emerged and increasing concern has been expressed about the presence of mercury in some northern lakes and rivers. The resort region is also endangered because of the high concentration of cottagers.

Settlement patterns. In northern Ontario, settlement has little agricultural base and is largely connected with major industries and transportation routes. Thunder Bay (formerly the twin cities of Port Arthur and Fort William) is located at the head of the Great Lakes navigation system and is the transshipment point for western wheat. Its population exceeded 110,000 by the early 1970s. Sudbury (pop. 155,000) is the centre of a major mining area, as are such communities as Timmins, Kirkland Lake, and Geraldton. Sault Ste. Marie is both an important lake-navigation port and a centre of large steel and paper industries.

Urban growth

Agricultural settlement is more frequent in southern Ontario, and about 80 percent of the region's farms are family-owned. The standard farm lot, determined in the 1790s, was an oblong of 100 or 200 acres, and despite a recent increase in the number of large farms, the average farm size was about 160 acres in 1966. Fields and townships were laid out in a square grid pattern, with roads one mile apart. In areas of French settlement, however, the long, narrow fields typical of French-Canadian strip farming may be seen. Villages grew up at water-power sites, convenient distribution points, and around early garrison centres. Kingston, the first important town, combined all three of these advantages.

Urbanization and industrialization are rapidly transforming the contemporary Ontario countryside; although the population density of southern Ontario as a whole grew from 39 persons per square mile in 1861 to 152 per square mile in 1961, the farming population dropped from 31 percent of the labour force in 1911 to 7 percent in 1961, with the absolute total also falling toward the end of the period, dropping below the half-million mark during the 1960s. Total farm acreage has also diminished,

due to the disappearance of farms on marginal lands, the conversion of land to recreational use, the great expansion of urban areas, and the growth of commercial farming operations. Urban growth has been confined almost entirely to the south-central and southwestern parts of the province, except in Ottawa, the national capital, in the east whose population approached the half-million mark by the 1970s. The Toronto metropolitan complex—the so-called "Golden Horseshoe"—sprawls along the Lake Ontario shore from Oshawa (population 100,000) to St. Catharines (population 110,000), and includes the major port and industrial city of Hamilton (population 499,000). Metropolitan Toronto, whose total population exceeds the 2,600,000 mark, is Canada's second largest city, with a hinterland that embraces not only much of the province but a good part of the country. It has a phenomenal rate of growth. Other important urban concentrations in western Ontario include Windsor, London, and Kitchener–Waterloo, all with populations above the 200,000 mark.

THE PEOPLE

Until the end of the War of 1812, Ontario was peopled chiefly from the United States, by Loyalists, frontier farmers, and Quakers and Mennonites from Pennsylvania—the latter forming a nucleus for German settlement in Waterloo County. For the remainder of the 19th century most of the immigrants were Protestants from the British Isles, although both Irish and Scottish Catholics came in large numbers. The first wave of British immigration, between 1815 and 1850, altered the original American character of the province. The second phase of European immigration, from 1896 to 1914, included sizable numbers from Germany, Scandinavia, Russia, Poland, the Ukraine, and Italy. Their arrival coincided with the first great mining discoveries in northern Ontario, and, as a result, the composition of that region's population became much less British in character than the remainder of the province. After World War II, Ontario's immigrants were mainly attracted to the industrial areas of the province, being drawn by the demands of a rapidly expanding economy. Although people from the British Isles still formed the largest single group (about 44 percent), they were no longer in the majority, and the cities of Ontario became more cosmopolitan.

Immigration from Europe

The change in urban populations has been quickened by the growth of suburbs. In the case of Toronto, the city proper has grown little since 1921 because of the stability of its boundaries. Metropolitan Toronto, however, has grown from about 1,000,000 people in 1951 to over 2,600,000 in 1971. Since the movement to the suburbs was largely by citizens of British origin, there has been an extraordinary change in the ethnic composition of the population. Whereas the city was 80 percent British in 1931, the proportion dropped to 52 percent by 1961 and to 43 percent by 1966. The largest single increase was in Toronto's Italian community, which now numbers well over 200,000, and the German, Polish, and Ukrainian communities also grew substantially. Most Ontario cities have mirrored these changes to some degree. Political disruption in Hungary in the 1950s, and in Czechoslovakia in the 1960s, brought thousands of immigrants to the province, and there has been a marked increase in Portuguese and Greek immigration. The general result has been to alter the cultural pattern of life considerably and to enliven it decidedly. Also, Roman Catholics have become, at 30 percent, the largest single religious denomination; they are followed by members of the United Church at 26 percent and Anglicans at 18 percent.

There are no Eskimos indigenous to Ontario, and the small Indian population is almost exclusively rural. Blacks first arrived as slaves to Loyalist immigrants and, in the 19th century, as escaped slaves from the southern United States. Immigration continued during the 20th century, especially from Detroit, Michigan. The black population is also small and is concentrated largely in Toronto, where most are probably of West Indian origin.

As well as immigration from abroad, Ontario has benefitted from population movements within Canada. Approximately a quarter of the net migration into Ontario came from other parts of the country. After Newfoundland joined the confederation in 1949, Newfoundlanders began to move to Ontario in increasing numbers, and there has been a continuing movement from Nova Scotia and New Brunswick. There are about 500,000 French-speaking Canadians living in Ontario, who were drawn from Quebec since the 19th century by the lumber industry and railroads of the north, the farms of the east, and the Cornwall industrial area. Ottawa contains the largest concentration of French-Canadians, and there are sizable communities in Windsor, Toronto, and the Niagara Peninsula.

Emigration

Balanced against large-scale immigration is the fact that, as with Canada as a whole, there has long been substantial emigration from the province. As early as the 1850s, Ontarians were attracted by the westward movement in the United States, and outflow to the south has been important ever since. In the 20th century, much concern has been caused by what appears to be a disproportionate loss of professionally and technically skilled people to the United States. Similarly, many thousands of Ontarians have migrated to the Canadian West—to Manitoba in the 1870s and to Alberta and Saskatchewan after the turn of the century.

The result of the relationship between immigration and emigration is that natural increase has been the more important cause of population growth. The population more than doubled between 1901 and 1951 and reached 7,703,106 in 1971. Despite this growth, Ontario's proportion of the total Canadian population dropped from 41 percent in 1901 to 33 percent in 1951. This reflected not only the opening up of the Canadian and American, West and British Columbia, but also the fact that Ontario's birthrate declined to a 1937 low of 16.9. The birthrate later rose to a high of 26.8 in 1957, but it has since declined to 17.5 in 1969. The combination of birthrate trends and immigration meant that in the 1950s the province absorbed almost 39 percent of Canada's total increase in population, and by the mid-1960s it represented more than a third of the country's total population.

The province is overwhelmingly urban, with 80 percent of its citizens living in metropolitan areas. The population is relatively young, with only 8 percent of the total being over age 65 and 38 percent under age 20. This has meant an enormous expansion of educational facilities and has accentuated the need for continuing economic growth. Although the Ontario death rate of 7.5 per 1,000 is slightly above the national average of 7.3, it is well below death rates for comparable Western countries.

THE ECONOMY

Ontario's strategic central location with respect to the Canadian provinces, its proximity to U.S. markets and coal supplies, its cheap power, its large and skilled labour force, its abundant natural resources and diversified transportation system, and its general attractiveness to both domestic and foreign investment, have made its economy the most productive in Canada. As with all mature economies, the bulk of employment and output is concentrated in the manufacturing and service industries.

Agriculture. With the annual value of agricultural production exceeding $1,000,000,000 during the 1960s, Ontario is second only to the three Prairie Provinces as a Canadian farming region. The most important cash crop is tobacco, but most farms are concerned with dairying or livestock. As a result of this specialization, corn (maize) acreage has more than doubled since 1951. Nevertheless, it is still necessary to import corn from the United States for livestock feed. The Niagara Peninsula is the chief fruit-producing region, while Kent and Essex counties in western Ontario and the Holland Marsh north of Toronto are the major areas of vegetable production.

Mining and forestry. Ontario is Canada's leading mining province; the value of its mineral production rose from under $200,000,000 in 1946 to over $1,000,000,000 20 years later. The province mines 51 percent of the world's nickel; this metal alone accounts for more than a third of the total value of mineral production. Copper

production, whose annual value approaches $300,000,000, is the second most important metal, followed by iron, zinc, gold, uranium, and silver. The iron reserves alone are estimated at 3,000,000,000 tons.

Although the forest products industry does not rank with agriculture and mining in terms of value of production, it is still one of the most important branches of the industry in Canada. Over 600,000,000 cubic feet of lumber were produced annually by the early 1970s, nearly 80 percent of which was for pulpwood; pulp and paper manufacture dominate the industry in Ontario.

Manufacturing. Ontario is the leading manufacturing province in Canada, employing half of the country's workers and generating the same proportion of the total value of production. Historically, this pre-eminence derived from the milling, farm implement, furniture, and textile industries of the 19th century. The Canadian protective tariff of 1878 encouraged a domestic steel industry at Hamilton and the location of many American branch plants in the province. Since the United States-Canada automobile agreement of 1965, Canadian automobile production—almost all of which is in Ontario—has jumped by 270 percent, and the impact on related industries also approaches the phenomenal. Other leading industries include textiles, food processing, and the manufacture of industrial machinery, electrical goods, farm implements, rubber and synthetics, aircraft, and furniture.

The automotive industry

Labour. In 1973 the total labour force was estimated at over 3,250,000 persons, almost 75% higher than 1961. There has been a noticeable trend to the service industries that now employ over three-fifths of the working force, only 5% of which remains in agriculture and the extractive industries. Since World War II, the most remarkable change, however, has been the increasing participation of women. In 1951, 26% of women over 15 years of age were in the labour force, in 1961, 33%, and by 1969, 38% were working. Trends in the distribution of labour are shown in the Table.

Trends in Distribution of Labour Force in Ontario
(percentage)

	1961*	1973†
Agriculture	7.1	3.5
Forestry	0.7	1.5
Fisheries	0.1	
Mining	1.8	
Total primary industries	9.7	5.0
Manufacturing	26.9	28.9
Construction	6.4	4.9
Total secondary industries	33.3	33.8
Transportation, communications, utilities	8.2	7.8
Trade	15.5	16.3
Finance, insurance, real estate	4.1	5.1
Services (includes public administration)	27.1	31.9
Total tertiary (service) industries	54.9	61.2

*Excludes 2.16% whose occupations were unspecified or undefined at the 1961 census. †Employed, rather than labour force, in March, 1973; *i.e.*, excludes unemployed and certain others classified in 1961 figures.

Management of the economy. Federal economic policy remains the chief agent of control acting upon the private sector, but the provincial government has played an increasingly important role. Since 1962 the province has been divided into ten economic regions for planning purposes, and planning advice is given to the Cabinet by the Ontario Economic Council. The provincial government has stimulated industrial research and development with the Sheridan Park Project near Toronto, where the research facilities of more than 30 industries are now located. Industrial disputes are regulated by the Ontario Labour Relations Board; minimum-wage laws have set wage levels throughout the province, and government-sponsored marketing boards have influenced agricultural production.

TRANSPORTATION

Bulk cargoes, chiefly consisting of mining and forestry products and prairie grains, are moved to the United States or overseas by the Great Lakes–St. Lawrence waterway system. Seagoing carriers bring imports from abroad by the same route: roughly two tons move out of this sytem for every incoming ton.

Road system

The 80,000 miles of provincial roads carry Ontario's 3,000,000 or so motor vehicles, plus the heavy annual influx of tourist traffic. The basic road pattern, laid out in the 1790s, is an east–west highway from the Quebec border to Windsor and a north–south expressway from Toronto. The Ontario section of the Trans-Canada Highway runs for 1,400 miles from Montreal through Ottawa to the Manitoba border. Capital and maintenance costs are high because the region's heavy snowfall and extreme temperature range make constant repairs necessary.

Ontario has over 10,000 miles of standard-gauge railroad track, about 40 percent of which is in the north. It is crossed by two transcontinental lines and is bisected by one provincially owned railroad with its northern terminus at Moosonee on James Bay. Although there has been a recent reduction in passenger mileage in southern Ontario due to lack of revenue, this region's rail network—centred around Toronto—is Canada's most elaborate.

Toronto is also the focus for the province's air traffic. Its International Airport is now the country's largest and is a main centre of operations for both domestic and international flights. Other important airports are located at Ottawa, London, Sault Ste. Marie, and Thunder Bay.

ADMINISTRATION AND SOCIAL CONDITIONS

The provincial government consists of the lieutenant governor, appointed by the governor general in council, whose functions are purely formal; the Executive Council, or Cabinet; and the elected Legislative Assembly of 117 members. The province is represented in the federal government by 24 senators and 85 elected members of the House of Commons. Manhood suffrage, except for Indians living on reservations, has obtained since 1888. Women were given the vote in 1917, and in 1954 Indians were enfranchised.

Rural political domination

Since 1963, urban constituencies elect the majority in the legislature, but rural areas are still over-represented. The Cabinet, headed by the premier (frequently called prime minister in Ontario), contains about 20 ministers, and it, too, tends to be dominated by rural interests. Northern Ontario is habitually under-represented, as is the province's Catholic population.

Ontario has three political parties—the Conservatives, the Liberals, and the New Democratic Party. The latter, founded in 1961, represents an amalgamation of the Socialist Co-operative Commonwealth Federation and sections of the trade-union movement. Stability and one-party dominance, however, has long been the mark of Ontario politics, and there were only six changes of government in the hundred years following the establishment of federation in 1867. The Liberals were in power from 1872 to 1905, the Conservatives from 1905 to 1919, and —after the fall of a United Farmers government in 1923 —the Conservatives returned until 1934. Although the Liberals held sway from 1934 to 1943, the Conservatives have held office since. It is conceivable that the changing character of Ontario society might alter the tradition of one-party rule, but the existence of three parties competing for votes has so far prevented that possibility. In 1963 and 1967, the Conservatives won the elections, although they received less than half of the vote, because the majority was split between the other two parties.

Municipal government in Ontario consists of elected township and county councils in rural areas and of elected councils in cities and towns. Urban growth, the demand for enhanced services, and the movement of taxpayers to the suburbs promoted the federation of the city of Toronto and its 12 suburbs in 1954. The 13 components retained their existence as political entities, but certain general powers and responsibilities were transferred to a second tier of government. In 1965 the federation was adjusted, subsequently being composed of the city proper and five boroughs with roughly equal populations. These changes have brought an efficient centralization of public services, have enhanced the borrowing capacity of the

Municipal federation

municipality, and have created a uniform tax system. Similar regional governments were instituted for Ottawa–Carleton in 1969 and for Niagara in 1970.

The province is constitutionally responsible for the administration of justice. Although the federal government appoints the Supreme Court and district and county court judges, the province appoints magistrates and juvenile and family court judges. A legal-aid system provides free counsel based on assessment of means, and compensation is given to innocent victims of criminal acts and to those injured attempting to prevent a crime. The Ontario Provincial Police, with a strength of 4,500, is the third largest deployed force in North America.

Government expenditure per capita is about Can$160. From 1955 to 1962, Ontario expenditures increased by 105 percent, and in 1962 the province had its first Can$1,000,000,000 budget. By the early 1970s, total expenditure was more than Can$3,000,000,000.

Education costs have risen sharply to over 40 percent of the provincial budget because of increased enrollment in elementary and secondary schools and increased assistance to higher education. Undergraduate and graduate schools are rapidly expanding, and 14 universities, though legally autonomous, receive both operating and capital grants from the Department of University Affairs, which has a measure of control over university planning, intake, and curricula. The province has also established 18 colleges of applied arts and technology to provide postsecondary technical education.

In the field of social legislation, the province has entered into important shared-cost arrangements with the federal government. An Ontario Health Services Insurance Plan, for example, provides comprehensive medical insurance to all residents, and is free for those with no taxable income. Ontario receives 50 percent of the cost of its welfare services from federal sources under the Canada Assistance Plan of 1966.

Health and welfare

Poverty among Indians, marginal farmers, and city dwellers is of continuing concern, but the proportion of the population on welfare has remained fairly constant at a relatively low level. Real wages have risen steadily and exceed the national average for manufacturing and other industrial personnel. The cost of living, however, is high compared to most of the other provinces, and substantial discrepancies exist between male and female wages, although legislation is now in force against some forms of such wage and job discrimination.

CULTURAL LIFE AND INSTITUTIONS

Toronto tends to dominate the province's cultural activities, although its pre-eminence is by no means exclusive. The first notable school of Canadian painting—that of the Group of Seven—was founded there in the 1920s. The capital city is the centre of Canada's English-language theatre, as well as the headquarters of national radio and television broadcasting. The city supports two symphony orchestras, two choirs, national opera and ballet companies, and many art galleries and museums.

Cultural institutions are also growing in other urban centres. Ottawa boasts the National Arts Centre, a symphony orchestra, museums, and art galleries. London is developing an artists' colony, while Stratford is already established as a fine dramatic arts centre and site of the annual Shakespearean Festival. The latter, first launched in 1952 in this pleasant but previously culturally undistinguished western Ontario town, has had a substantial impact upon the cultural life of the province. There is also an annual George Bernard Shaw Festival at Niagara-on-the-Lake. Upper Canada Village near Morrisburg is a re-creation of a 19th-century Ontario community, and Ft. Henry at Kingston and Ft. George at Niagara Falls have been preserved.

THE FUTURE

Ontario's remarkable growth in wealth and population has brought not only greater educational, cultural, and recreational opportunities for its people but also many of the difficulties of other highly urbanized areas of North America. In 1970 the provincial government disclosed its "Design for Development: the Toronto-Centred Region." This plan contained an estimate that, by the year 2000, at least 8,000,000 of a total provincial population of 13,000,000 would live in the area of the Toronto conurbation. Since Ontario is already experiencing serious pollution, traffic congestion, and the complex social dislocations incidental to rapid urbanization, it is likely that public concern will increasingly be directed to the environmental and social consequences of further growth. It seems probable that the planning function of the provincial government will be enlarged and that efforts will be made to locate both industry and population in eastern and northern Ontario to a greater degree than has yet been the case. With a large and self-conscious French-speaking minority, Ontario has committed itself to the task of reaching agreement with Quebec and the other provinces on a satisfactory restructuring of the Canadian union. In 1967 Premier John Robarts took the initiative by calling the Confederation of Tomorrow Conference of provincial premiers in Toronto, and his successors will continue to bear a heavy responsibility for the provision of leadership in the quest for national unity.

BIBLIOGRAPHY

History: No complete history of Ontario has yet been written. E.G. FIRTH (ed.), *Profiles of a Province: Studies in the History of Ontario* (1967), is the most useful single source.

Landscape: L.J. CHAPMAN and D.F. PUTNAM, *The Physiography of Southern Ontario*, 2nd ed. (1966), is excellent. The Canadian Shield region is described in JOHN WARKENTIN (ed.), *Canada: A Geographical Interpretation* (1968). For settlement patterns, see JACOB SPELT, *The Urban Development in South-Central Ontario* (1955).

People: The peopling of Ontario may be followed in two volumes ed. by J.M.S. CARELESS, *Colonists and Canadiens, 1763–1867* (1971), and *The Canadians, 1867–1967* (1967).

Economy: The best statistical source for the economy, as well as for other aspects of Ontario, is the *Canada Year Book;* W.T. EASTERBROOK and H.G.J. AITKEN, *Canadian Economic History* (1956), is a useful introduction. W.G. DEAN, *Economic Atlas of Ontario* (1969), is a magnificent compilation in graphic form.

Transportation: G.P. DE T. GLAZEBROOK, *A History of Transportation in Canada*, 2 vol. (1964), is a standard source. P. CAMU *et al.*, *Economic Geography of Canada* (1964), gives a good functional treatment.

Administration and social conditions: F.F. SCHINDELER, *Responsible Government in Ontario* (1969), is the only good treatment of the structure of government. Local government is treated by T.J. PLUNKETT, *Urban Canada and Its Government: A Study of Municipal Organization* (1968); education by ROBIN HARRIS, *Quiet Evolution: A Study of the Educational System of Ontario* (1967). JOHN PORTER, *Vertical Mosaic: An Analysis of Social Class and Power in Canada* (1965), is a major work of scholarship.

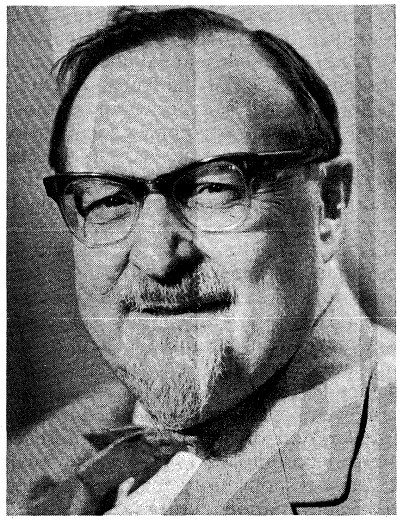
(S.F.W.)

Oparin, Aleksandr Ivanovich

The 20th-century idea that there can be a natural explanation of the origin of life pervaded the scientific career of Aleksandr Ivanovich Oparin. By drawing on the insights of chemistry, he extended the Darwinian theory of evolution backward in time to explain how simple organic and inorganic materials might have combined into complex organic compounds, and how the latter might have formed the primordial organism.

Oparin, the youngest of three children, was born in 1894 in a small village near the Volga River, north of Moscow. When he was nine, his family moved to Moscow because there was no secondary school in their village. While majoring in plant physiology at Moscow State University, Oparin was influenced by K.A. Timiryazev, a Russian plant physiologist, who had known Darwin. The indirect effect of Darwin upon Oparin's thinking can be found in many of the latter's writings.

In his postdoctoral days Oparin was influenced also by A.N. Bakh, a botanist. Bakh left Russia at the time of the Revolution but later returned. Despite the financial difficulties of the times, the Russian government established a biochemical institute in his honour in 1935 in Moscow; Oparin helped to found it and has been its director since 1946.



Oparin, 1970.
Tass—Sovfoto

Concept of primordial organism

At a meeting of the Russian Botanical Society in the spring of 1922, Oparin first introduced his concept of a primordial organism arising in a brew of already formed organic compounds. He stated a number of premises that were not popular at the time. For example, according to his hypothesis, the earliest organisms were heterotrophic —*i.e.*, they obtained their nutrition ready-made from compounds that had already been formed in variety and profusion by what are in the laboratory quite ordinary means. Thus, at that early stage, these first organisms did not need to synthesize their own food materials in the way that present-day plants do. Oparin also emphasized that a high degree of structural and functional organization is characteristic of the living state, a point of view that is in opposition to the idea that "life" is essentially molecular. He was also farsighted in his observation that living organisms, as open systems, must receive energy and materials from outside themselves; they cannot, therefore, be limited by the second law of thermodynamics, which is applicable to closed systems in which energy is not replenished.

When Oparin first proposed his hypothesis, the prevailing view was that the first organisms could make all of their own organic compounds, and so the negative reaction to his proposal was almost universal. With continued retesting, however, his concept has come to be accepted in its main outlines. Although the possibility of a natural origin of life had been promulgated for at least 2,500 years, a specific formulation had to compete with vitalistic points of view in modern times. Also, organic chemistry, necessary for Oparin's hypothesis, had not been sufficiently developed by the time of Pasteur.

Oparin's various novel premises can be shown to be closely related to one another. What had been missing was an explanation of (1) how populations of large, complex molecules of largely predetermined structure could have arisen in contrast with the widely held view that the first proteins would have been random in structure and (2) an adequate explanation of how a first cell-like system might reproduce. When experimental answers to these questions arose from another laboratory, Oparin acknowledged them in a forthright manner. These answers consisted essentially of (1) ordered coupling of amino acids due to their differing shapes and distribution of electric charge and (2) the formation of buds on microscopic droplets followed by growth of separated buds and cyclical repetition of the process. In attempting to test his basic hypothesis, Oparin dealt with coacervate droplets, which are microscopic units assembled typically from gelatin and gum arabic, as models of early cells. His experiments showed that enzymes (biological catalysts) could function more efficiently within the boundaries of these artificial cells than they could in ordinary aqueous solution. This helped emphasize the fact that complete cells are important for the action of enzymes and metabolism.

The heterotroph hypothesis

The heterotrophic hypothesis for the origin of life has gained wide attention through Oparin's efforts. He organized the first international meeting of the origin of life in Moscow in 1957 at which representatives from 16 countries participated. A second conference was held in 1963, and at a third in Pont-à-Mousson, France, in 1970, Oparin was acclaimed president of the newly formed International Society for the Study of the Origin of Life.

Although he is best known for his contributions to studies of the origin of life, Oparin also devoted considerable effort to enzymology and to the closely related subject of industrial biochemistry. His wide interests are reflected in the title of the volume prepared in honour of his 70th birthday, *Problems in Evolutionary and Industrial Biochemistry*. But at the start of the 1970s, the centre of his interest remained at the A.N. Bakh Institute, where, under his direction, a number of research workers were concerned with the problems of the origin of life. Oparin's decorations include the Order of Lenin, Hero of Socialist Labour, the Bakh Prize, and the Mechnikov Gold Medal.

BIBLIOGRAPHY. Oparin's definitive work is *The Origin of Life on Earth*, 3rd rev. ed. (1957). His other books (in English translation) include: *Enzyme Action in the Living Cell* (1934); *Life: Its Nature, Origin, and Development* (1961); *The Chemical Origin of Life* (1964); and *Genesis and Evolutionary Development of Life* (1968). S.W. FOX (ed.), *The Origins of Prebiological Systems and Their Molecular Matrices* (1965), treats technical and historical aspects; and J.D. BERNAL, *The Origin of Life* (1967), discusses related sciences and some aspects of the main problem.

(S.W.F.)

Opera

The English word opera is an abbreviation of the Italian phrase *opera in musica* ("work in music"). It names a theatrical form consisting of a dramatic text (libretto, or "little book") combined with music, usually singing with instrumental accompaniment. Besides solo, ensemble, and choral singers and a group of instrumentalists, the forces performing opera since its inception have often included dancers. A complex, often costly variety of musicodramatic entertainment, opera has attracted audiences for nearly five centuries. Although its supporters have greatly outnumbered its detractors, it has been the target of intense adverse criticism.

Charles de Saint-Évremond, a French man of letters in the 17th century, called it "a bizarre thing consisting of poetry in music, in which the poet and the composer, equally standing in each other's way, go to endless trouble to produce a wretched result." The 18th-century English statesman and writer Lord Chesterfield wrote to his son: "As for operas, they are essentially too absurd and extravagant to mention. I look upon them as a magic scene contrived to please the eyes and the ears at the expense of the understanding." At the opposite extreme of reaction to opera, it has been said that the mere existence of such a masterpiece as Mozart's *Le nozze di Figaro* (*The Marriage of Figaro*) suffices to justify Western civilization.

The unique characteristics of opera

Although the characteristic of opera that most clearly separates it from other theatrical forms is that its principals sing rather than speak their lines, to approach it or criticize it as simply one variety of the musical art is to misjudge it. Its multiple creators almost always have intended an opera to be a lofty and eloquent form of theatrical performance. What commonly differentiates it from other varieties of musicodramatic theatre such as operetta (literally "small opera") and musical comedy is sobriety of workmanship, density of texture, and (even in operas with comic and farcical librettos) accompanying seriousness of musical tone. On the other hand, some lighter works—by Jacques Offenbach, Johann Strauss the Younger, Gilbert and Sullivan, Kurt Weill, George Gershwin, and a few others—make neat categorization impossible.

In the preparation of an opera performance, many individual artists and artisans, sometimes spread out across a

century and more, necessarily are or have been involved. The first, unintentional recruit is likely to have been the writer of the original story. Then comes the librettist, who puts the story or play into a form suitable for musical setting and singing, and the composer, who sets that libretto to music. Architects and acousticians have built an opera house suited or adaptable to performances that demand a sizable stage, a pit to house an ensemble of instrumentalists, and a reasonably large audience. A producer-director has to specify the work of designers, scene painters, costumers, and lighting experts. The producer, conductor, and musical staff have to work for long periods with chorus, dancers, orchestra, and extras as well as the principal singers to prepare the performance—work that may last anywhere from a few days to many months. All this does not even take into account the part played by the administrative staff.

Once the complete operatic score—the final libretto and music—is available, what must rule all of those involved is dedication to fulfilling the wishes of the librettist and the composer. Overemphasis or underemphasis of any larger component of an operatic performance can be as damaging to it as off-pitch singing or false entries by instrumentalists. More than one desirable balance among the constituents of performance is often possible. What is certain is that one or another of them must be decided upon, worked toward, and achieved.

The article is divided into the following sections:

I. The early centuries

ITALIAN ORIGINS

Works in antiquity had combined poetic drama and music. The plays of the ancient Greek dramatists Aeschylus, Sophocles, and Euripides employed choral music in a manner that certainly reflected related usages in earlier times. During the Middle Ages, biblical dramas were commonly accompanied by some music, being known under various labels, including mystery and miracle plays. These and other related musicodramatic forms may or may not have become collateral ancestors of opera: their descent seems most certain in some 17th-century operas on religious subjects performed in Rome and at several places in Germany. Musical historians and musicologists continue to debate opera's ancestry. The earliest universally accepted direct ancestors of opera appeared in 16th-century Italy. Purely nonreligious works of edifying drama with music, they included intermezzos and *intermedii* played between the acts of spoken dramas, with which their purported subject matter often claimed a tenuous connection, and staged ballet. The latter, Italian by birth, achieved a complex, quasi-operatic state in the court ballet (*ballet de cour*) danced in France late in the 16th century and throughout the 17th: it approached ever closer to opera in the *comédie-ballet* evolved by Molière and Jean-Baptiste Lully in the 1660s, beginning with *Le Mariage forcé* (*The Enforced Marriage*, first performed 1664).

16th century. Musicians, singers, poets, playwrights, and enthusiasts of the literary, musical, and theatrical arts had long cherished a desire for some more formally constituted and more stable form of drama with music, especially in Italy. One response to that expressed desire was the 16th-century "madrigal comedy," the singing of dramatic or semidramatic lines (often farcical, and most often with story and characters borrowed from the traditional commedia dell'arte as it had become formalized during the 16th century) in a linked series of more or less discrete madrigals and other varieties of polyphonic song.

16th-century "madrigal comedy"

But polyphony—the musical texture created by simultaneous, largely unaccompanied singing of interwoven melodies—was by nature alien to theatrical drama: it made extremely difficult, when not impossible, the delineation of individual characters through clearly understandable text words. This is noticeable even in the most celebrated of the madrigal comedies, Orazio Vecchi's paean to the "double Parnassus" of poetry and music, *L'Amfiparnaso* (Modena, 1594).

The gestation of opera required the simultaneous availability of a dramatic literary style and a musical texture suitable for incorporation into a new theatrical unity. The essential literary materials had begun to appear in Italy in such chivalresque epics as Lodovico Ariosto's *Orlando furioso* (published complete in 1532) and Torquato Tasso's *Gerusalemme liberata* (1575), both of which were to be mined for subjects by innumerable opera librettists. More immediately decisive in setting the first direction of opera was one sort of poetic drama: the shorter pastoral writings of 16th-century Italian poets, notably Tasso's *Aminta* (1573) and Giovanni Battista Guarini's *Pastor fido* (*The Faithful Shepherd*, completed in 1590). Idylls or eclogues that had sprung up in the 15th century, like the *Orfeo* of the Italian poet Poliziano (Politian) with music by Germi (one of the earliest examples, which was staged at Mantua in 1472, with solo song, chorus, and spoken dialogue), were seized on, adapted, and imitated by the men who had begun to evolve the musical texture essential to the birth of opera and who found apt subjects in the loves, joys, and sorrows of Arcadian shepherds and shepherdesses, often with the intervention of gods, demigods, and heroes.

Until the 1950s it was generally accepted that opera originated with a *camerata* (a sort of humanistic discussion group) that met in the late 1570s and early 1580s in the Florentine palace of Giovanni Bardi, count of Vernio. In 1953, however, the Roman musicologist Nino Pirrotta showed that the Bardi Camerata, far from having furthered innovation or interested itself in musical drama, was predominantly conservative, often acted in defense of the polyphonic madrigal, and showed no sympathy with the new combinations that would shortly produce opera. In fact, that literary–musical texture was evolved at Florence, but largely among a group of intellectuals, artists, and dilettantes who met informally in the palace of the theatrical theoretician Jacopo Corsi during the 1590s. This latter group also included Emilio de' Cavalieri, the composer, impresario, and choreographer who was to write what is often called the first oratorio, *Rappresentatione di anima et di corpo* (*The Representation of Soul and Body*, an acted form unlike later oratorios); the singer-composer Jacopo Peri; and—although they attended infrequently—both the poet Tasso and the composer Claudio Monteverdi. Still active at the time, though a little out of favour, was the singer-composer Giulio Caccini. Corsi and his friends were by no means the first creators of solo vocal lines with instrumental accompaniment, and they shaped their musicotheatrical creations partly in the mistaken belief that their performances were reviving ancient Greek procedures. What, in fact, they did was to take hints from the French court ballet and simultaneously discard polyphony in favour of monody (or homophony)—accompanied singing or recitation on musical tones (recitativo) of one melody at a time. Thus, they insured both the relative comprehensibility of the words (which to them seemed much more important than the accompanying music) and the use of at least some instrumental support.

Florentines who "created" opera

An important "manifesto" of the monodic innovators was a collection of short vocal pieces with thorough-bass

accompaniment (instrumental chords in sequence as accompaniment to melody) by Caccini, published in 1602: *Le nuove musiche*, a title that often has been extended to cover the novel musical texture itself. The interaction of these and other Italians with the texture of monody was what finally led, after some false starts, to the emergence not only of opera more or less as it is known today but also of the cantata and the oratorio.

The honour of being deemed "the first opera" usually is given to a setting by Peri of *Dafne* by the Renaissance pastoral poet Ottavio Rinuccini. It was staged at the Palazzo Corsi in Florence during the pre-Lenten Carnival of 1597–98. The text, divided into a prologue and six scenes, was published in 1600 and therefore survives, but neither Peri's music (the prologue and one aria excepted) nor that of an almost contemporary setting of the same text by Caccini can now be recovered. The earliest surviving opera is also Peri's: a setting of Rinuccini's pastoral *Euridice*, likewise in a prologue and six scenes, which was performed at the Palazzo Pitti in Florence on October 6, 1600. More unusual, the musical score also was issued in 1601 and reprinted several times thereafter.

17th and 18th centuries. However significant historically, the pioneering operas of Peri and Caccini were tentative in both style and structure; further, neither of the founding fathers of opera seems to have possessed notable dramatic talent.

Monteverdi. Within ten years of the premiere of Peri's *Dafne* at Florence, Mantua heard an opera that is a masterpiece and still is staged frequently. This was *La favola d'Orfeo* (*The Fable of Orpheus*), a setting by Monteverdi of a poetic text by Alessandro Striggio the Younger. Presented during the Carnival of 1607 (the libretto was published then, the score in 1609), it soon was presented elsewhere in Italy. In *Orfeo*, the accompanying instruments come into their own as a dramatic element: the score contains more than two dozen pieces for increased (though not precisely determinable) numbers of instruments. It not only introduces, as a preluding toccata, the idea of the operatic overture but also achieves some sectional unity by repeating brief instrumental numbers (ritornellos). More important, Monteverdi uses recitative expressively and gives it an organizational function by repetitions and developments in predetermined patterns.

Monteverdi's *Orfeo*

Monteverdi continued to compose operas for more than 35 years; meanwhile, the new manner of musicodramatic entertainment spread to other Italian cities. Rome probably first heard an opera as early as 1606, Bologna before 1610. Continuing to employ librettos based on Italian interpretations of Greek and Roman myth, legend, and pseudo-history, literary men and composers rapidly swelled the number of operas heard. At Venice in 1642, the 75-year-old Monteverdi created his masterpiece *L'incoronazione di Poppea* (*The Coronation of Poppea*). Gian Francesco Busenello's superior libretto carried a new note of realism into opera, particularly in the development of human character; it was to have a prolonged, important line of descent, and Monteverdi translated it blazingly into music. Throughout that first period of operatic history, the importance given by composers to emotional drama, to instrumental music, and to structural stability increased along with the capabilities (and pretensions) of singers and the magnificence and complexity of stage settings, stage machinery, and costuming.

The first opera house

Venice. The inauguration early in 1637 of the first opera house open to the general public, the Tron family's Teatro di San Cassiano at Venice, was another decisive factor in establishing opera. That action removed opera from the exclusive hands of royalty and nobility and placed it within reach of all but the poorest sectors of the Italian urban population.

A pupil of Monteverdi, Pier Francesco Caletti-Bruni, known as Francesco Cavalli (1602–76), became the most popular opera composer of his era by furnishing the opera houses of Venice with some 40 operas between 1639 and 1669. A highly talented but not always fastidious composer, Cavalli reacted to the librettos he used with dramatic force and directness. The most renowned of his operas was *Giasone* (*Jason*, 1649), whose libretto by Giacinto Andrea Cicognini included farcical episodes. His chief Venetian rival and successor was Pietro Antonio (often called Marc' Antonio) Cesti (1623–69), about a dozen of whose nearly 100 operas have survived. Some notion of the extravagance to which imitation of Louis XIV's spectacles at Versailles had driven the production of opera elsewhere can be gained from descriptions of the Cesti opera *Il pomo d'oro* (*The Golden Apple*, 1667), composed for the wedding of Emperor Leopold I and Margarita Teresa of Spain in Vienna. Constructed in a prologue and five acts (the third and fifth of which have been lost), with 48 characters, it contained 66 scenes requiring 24 stage settings making use of complex stage machinery. Ballets occurred in every act, and a grand triple ballet brought the opera to its conclusion. *Il pomo d'oro* provided numerous purely instrumental introductions and interludes, gave relatively little importance to the chorus, skipped rapidly over the essential storytelling recitative, and concentrated on arias and duets that often were most sensuous and almost feminine in allure.

Opera in Venice and Rome

Venetian opera continued to flourish in the works of such greatly talented theatrical composers as Cavalli, Cesti, and Giovanni Legrenzi (1626–90). In some details, these Venetian operas reflected the pressures exerted by the tastes and wishes of the paying audiences for which they were designed. Not so lavish with choral interpolations as their Roman contemporaries, the Venetians demanded and received complex, strong librettos calling for large casts and special lavishness in staging. They also began to develop the sensuous melodic profiles that have come to be thought of as particularly "Italian." Furthermore, they all but separated the solo aria from the surrounding recitative. They also frequently prefaced and followed solos with purely instrumental music, so continuing the orchestra's elevation from a purely accompanying role. After the middle of the 17th century, the Venetian operatic style began to decline. Among the later Venetian operatic composers of talent and fame were Antonio Lotti (*c.* 1667–1740), Carlo Francesco Pollaroli (1653–1722), Antonio Vivaldi (*c.* 1678–1741), and Baldassare Galuppi (1706–85), who is often referred to somewhat loosely as "the father of opera buffa."

Rome. Several Italian cities soon developed recognizably indigenous operatic styles. At Rome, for example, a group of composers tended toward unified structure, gave ensemble and choral song expanded roles, and increased the difference between the solo (aria) and the Florentine type of continuous recitative by allowing arias to interrupt dramatic progress in order to express or comment upon emotional moods. Less emphatic about stage magnificence than their Venetian counterparts, such Roman composers as Stefano Landi (*c.* 1590–*c.* 1655), Domenico Mazzocchi (1592–1655), Luigi Rossi (1597–1653), and Michelangelo Rossi (1600?–70) also permitted comic episodes to lighten prevailingly tragic stories. They concentrated attention productively on instrumental overtures and on overture-like pieces preceding acts or sections of acts. Two Roman composers—Domenico Mazzocchi's brother Virgilio (1597–1646) and Marco Marazzoli (1619–62)—often are cited as having created the first completely comic opera, *Chi soffe speri* (*He Who Suffers, Hopes*, 1639). Its libretto was written by Cardinal Giulio Rospigliosi, who was to be elevated to the papacy in 1667 as Clement IX. The invited guests at its first performance, in the Palazzo Barberini, included the English poet John Milton and Giulio Mazarini, the future Cardinal Mazarin, statesman to Louis XIV.

Naples. With the 18th century the centre of Italian opera shifted to Naples, where so great a variety of styles evolved that the term Neapolitan opera eventually covered operas that dominated most of Italy and many foreign centres of operatic activity. With some exceptions, the earliest unmistakably Neapolitan operas changed their focus back from the music to the words. Two of its instigators were dramatic poets: Apostolo Zeno (1668–1750), born a Venetian, and the Roman Pietro Trapassi, known as Metastasio (1698–1782)—perhaps the greatest of the 18th-century librettists. Continu-

Contributions of Zeno and Metastasio

ing the custom of basing librettos on Greco-Roman legend and pseudo-history (but dispensing almost entirely with classical mythology), Zeno and Metastasio wrote texts of formal beauty and linguistic clarity, preferring solemn, usually tragic subjects (opera seria) in three acts to comic episodes and characters. The aria came to dominate, and the use of chorus declined.

The term Neapolitan opera also came to indicate harmonically naïve, melodious lighter operas in the gallant tone of the Rococo style; the rich development of the bel canto styles (where beautiful singing *per se* was predominant), signifying supreme vocal agility and smoothness that was supplied first by castratos, men who had been castrated before puberty in order to preserve the high ranges of their boyish voices; and the appearance of the *centone* or pasticcio, a libretto set to a score made up of music borrowed either from scores (then uncopyrighted or otherwise legally protected) of several composers or from several operas by a single composer. The use of the orchestra also became limited. But perhaps the most discussed, and often senselessly maligned, feature that particularly designated Neapolitan opera was the aria da capo, an aria in three sections, the third part repeating the first. It had appeared in northern Italy early in the 17th century but was employed with comparative infrequency there. Some Neapolitan operas, however, consisted of up to 20 da capo arias separated by a minimum of story-advancing recitativo secco (narrative passages in which voice is accompanied only by a thorough bass).

Domination of the aria da capo

A masterly operatic composer of the transitional style who bridged the era between the Baroque and the preclassical Neapolitan style was Alessandro Scarlatti (1660–1725). In his many operas Scarlatti triumphed, by the strength of musical imagination, over librettos intended to provide vehicles for phenomenally trained singers in the prevailing pattern and the consequent reduction of dramatic interest to a narrow minimum. Talented and influential among Scarlatti's contemporaries were such composers as Nicola Antonio Porpora (1686–1768), Leonardo Vinci (1690–1730), and Leonardo Leo (1694–1744).

Attempts at reform. In 1720 the Venetian composer-poet-statesman Benedetto Marcello (1686–1739) published a mordant satire on the increasingly rigid and undramatic conventions that had taken hold of opera seria: *Il teatro alla moda, o sia metodo sicuro e facile per ben comporre ed eseguire opere italiane in musica* ("The Theater à la Mode, or The Secure and Easy Method of Composing and Performing Italian Operas"). The distress that it and other criticisms brought resulted in an improved genre, still in effect opera seria but showing attempts at reform of its mannerisms. The principal operas against which that reform later evolved were the often melodically seductive works of Gaetano Latilla (1711–88), Giuseppe Sarti (1729–1802), Antonio Sacchini (1730–86), Johann Christian Bach (1735–82), Antonio Salieri (1750–1825), and Mozart's *Idomeneo, rè di Creta* (1781) and *La clemenza di Tito* (*The Clemency of Titus,* 1791). Representative composers within the short "reform" movement itself were Niccolò Jommelli (1714–74) and Tommaso Traetta (1727–79). A more intellectually rigorous reformation was undertaken consciously by Christoph Willibald Gluck (see below *The "reform"*) in collaboration with the librettist Ranieri de' Calzabigi, beginning with *Orfeo ed Euridice* (1762).

Comic opera. Comic opera meanwhile had expanded from its shadowy existence within and between the acts of opera seria. From the early, tentative efforts of several 17th-century Roman and Florentine composers, it had moved into a bustling, rude, independent vitality of its own, often in the form of satirical opera buffa (Italian: "comic opera"), generally shaped in two acts rather than the usual three of opera seria. Expelled from the precincts of opera seria by the librettos of Zeno and Metastasio, the comic spirit had taken refuge in such an expanded intermezzo as *La serva padrona* (*The Maid Mistress,* 1733), by Giovanni Battista Pergolesi. When it matured, the style borrowed back some of the more serious emotional qualities of opera seria, often including "serious" roles among those of the comedians. This led to a hybrid nature in many operas, including two works using librettos derived from the plays of Pierre de Beaumarchais—*Il barbiere di Siviglia* (*The Barber of Seville,* 1782), by Giovanni Paisiello, and Mozart's *Le nozze di Figaro* (1786)—as well as *Il matrimonio segreto* (*The Secret Marriage,* 1792), by Domenico Cimarosa.

One of the determining characteristics of this mixed style was the elaboration of ensemble numbers concluding acts. These operas dispensed almost entirely with the magnificent display and grandeur of staging increasingly required of opera seria. Another of the drawbacks of the mixed style was well summed up by the 20th-century Italian musicologist Andrea Della Corte:

> With few exceptions among the great composers dedicated to instrumental music or to teaching, almost all the serious composers also collaborated in the comic theatre. Not so the literary men. The best dramatists were not equally tempted by this tendency.

The natural result was that the large majority of opera buffa librettos remained inferior to the serious texts by Metastasio and his imitators and successors, though not necessarily less workable in their own way.

EARLY OPERA IN FRANCE

Opera was imported into France from Italy before 1650, but it long failed to take firm hold there with royal and other audiences, at first having to compete on unequal terms with the spoken drama (often with musical interludes) and the ballet. The *Pomone* (1671) of Robert Cambert, to a pastoral libretto by Pierre Perrin, is commonly called the first French opera. Its premiere almost certainly inaugurated the Académie Royale de Musique (now the Paris Académie de Musique or Paris Opéra) on March 3, 1671. Only fragments of the music of *Pomone* still exist.

Lully's important role

Opera really did not become a French art until the time of Jean-Baptiste Lully (1632–87). This highly talented, very shrewd, and dictatorial man borrowed freely from both the spoken French drama and the court ballet. Though himself an Italian, he played down the extended, formalized Italian aria in favour of shorter, more instantly captivating "airs." He formed recitative after the declamatory manner of the Comédie-Française theatre company and also evolved the "French overture" (one movement with a slow and a fast section) as distinct from the "Italian overture" (one movement with a fast, slow, and another fast section). His operas and those of his imitators and followers assigned great importance to dancing, choruses, instrumental interludes, and a dazzlingly complex stage setting. Lully became the virtual dictator of music in France partly because of the strengths of his literary collaborators: first the dramatist Molière (1662–73) in *comédie-ballet,* then the exceedingly able librettist Philippe Quinault (1635–88).

The pervasive Lullyan style, altered surprisingly little except in the direction of still more imposing grandeur, attained its culmination in the magnificent operas of Jean-Philippe Rameau (1683–1764), especially in his *Hippolyte et Aricie* (1733; libretto by Simon-Joseph de Pellegrin), *Les Indes galantes* (*The Courtly Indies,* 1735; libretto by Louis Fuzelier), and particularly *Castor et Pollux* (1737; libretto by Pierre-Joseph-Justin Bernard), which was performed at the Paris Opéra 254 times in 48 years. Except for the special instance of *Les Indes galantes,* which was billed as a *ballet héroique,* Rameau's chief operas were each divided into a prologue and five acts, a pattern that many later French composers favoured. Rameau confirmed the still-enduring insistence of French operatic composers on setting their language to music with such probity and clarity that it can be understood properly when sung. His operatic works are regarded widely as the apogee of 18th-century French opera.

EARLY OPERA IN GERMANY AND AUSTRIA

Although Heinrich Schütz composed *Dafne,* the first known opera with a German text, and heard it played at Torgau on April 23, 1627, the active history of opera in Germany began with the Italian composers residing

there. A remarkable Venetian composer-diplomatist-ecclesiastic, the Abbé Agostino Steffani, carried much of his native city's early operatic manner to Munich, Hanover, and other German centres, beginning his operatic production with *Marco Aurelio* (Munich, 1681), and continued thereafter to compose operas for 28 years. In his use of both Italian and French procedures, particularly in handling overture and recitative, Steffani evolved a sort of international Italian style that clearly influenced other "transplanted" composers.

For the next 100 years the influence of Italian opera was so pervasive that even native German composers adopted the Italian operatic style and used texts in Italian.

Singspiel. The German word *Singspiel* was originally used for all sorts of opera. The earliest known entertainments so designated were composed by a pupil of Heinrich Schütz, Johann Theile (1646–1724). One of them, *Adam und Eva*, eminently "serious" as to story, inaugurated the Hamburg Opera in 1678. During the mid-18th century the term singspiel came to be reserved for what the English called "ballad opera," the French *opéra-comique*: light, usually comic operas including spoken dialogue. The comic singspiel of the 18th century was born with *The Devil to Pay* (London, 1731) and its sequel, *The Merry Cobbler* (London, 1735), English ballad operas with texts by Charles Coffey. These had pasticcio scores capitalizing, not very successfully, on the great popularity of *The Beggar's Opera* (1728), which had a text by John Gay and an assembled (pasticcio) score brought together by John Christopher Pepusch. The Coffey texts having been translated into German, scores were composed to them by J.C. Standfuss (died 1759?) as *Der Teufel ist los* (Leipzig, 1752) and *Der lustige Schuster* (Lübeck, 1759); they later were restaged as arranged by Johann Adam Hiller (1728–1804), who also composed several other singspiels and brought to culmination the so-called Leipzig School. Both Berlin and Vienna inevitably took up the singspiel, examples of which, composed in those cities and elsewhere, included *Der neue krumme Teufel* (Vienna, 1752, music lost), by Joseph Haydn; Mozart's *Die Entführung aus dem Serail* (*The Abduction from the Seraglio*, Vienna, 1782) and *Die Zauberflöte* (*The Magic Flute*, Vienna, 1791); *Doktor und Apotheker* (Vienna, 1786), by Karl Ditters von Dittersdorf; Beethoven's *Fidelio* (Vienna, 1805), which, like *Die Zauberflöte*, immeasurably transcends the common artistic scope of the singspiel; and *Die Zwillingsbrüder* (*The Twin Brothers*, Vienna, 1820), by Schubert.

Popularity of *The Beggar's Opera*

Opera seria. The most important opere serie composed in Germany during the early 18th century were created for the Hamburg Opera, at which both Reinhard Keiser (1674–1739) and, for a brief interval, the young George Frideric Handel worked. Keiser composed more than 125 operas, mostly to German texts. Of Handel's large operatic output, only two works with Italian texts—*Almira* and *Nero* (both 1705, the second now lost)—were staged during his Hamburg stay. Keiser doggedly tried to attract the widest possible public. His operas often succeeded in charming audiences, but most of them have vanished, and those that survive appear to possess only superficial allure, although they are historically of interest for their skillful exploitation of the orchestra and of solo instruments tellingly used during arias.

Handel went from Italy to England in 1710. In London, with the opera *Rinaldo* (1711), he began 30 years of stubborn dedication to the by then moribund (but still dominant) traditions of Neapolitan opera seria. He created, however, a dozen or more of the most inspired operas of the first half of the century, including *Giulio Cesare* (*Julius Caesar*, 1724), *Tamerlano* (1724), *Rodelinda* (1725), *Sosarme* (1732), *Orlando* (1733), *Ariodante* (1735), and *Alcina* (1735). Handel transcended the formal style of opera seria by his melodic inspiration and harmonic daring. He even managed an immense variety of characterization within the cramped style with which he had to comply.

Handel's operatic style

German by births, but almost wholly Italianate by disposition, Johann Adolph Hasse (1699–1783) successfully carried on the Metastasian traditions of opera seria in a plethora of operas to Italian texts. The intensely Italian sensuousness of his best melodies, supported by some attractive adventurousness in harmonic placement and in instrumentation, did almost as much as Handel's operas to prolong past its true prime the glory of the Neapolitan style.

II. From the "reform" to grand opera

THE "REFORM"

Christoph Willibald Gluck (1714–87) has become an ambivalent figure in the evolution of opera: he has been loosely and often incorrectly categorized, and both praised and condemned, for mistaken reasons. His operas to Italian texts that he composed up to about 1756 were conventional settings of Metastasian librettos. After settling in Vienna in 1750, though not abandoning the composition of traditional opere serie in Italian, Gluck began to react to the French operatic styles popular there. At first he merely added a few new numbers to trivial one-act opéras-comiques brought to Vienna from Paris, which he arranged and conducted for the court at Schönbrunn and Laxenburg. Then he began to compose similar operas in French.

Thanks to the enthusiasm of the superintendent of the imperial Vienna theatres, Conte Giacomo Durazzo, Gluck had been absorbing the example of the outstanding French dancer-choreographer Jean-Georges Noverre (1727–1810). Seminal in Noverre's call for reform was the insistence that a ballet should not be left as a simple collection of unconnected episodes but should be shaped into a mimed dance drama. Gluck then composed the ballet *Don Juan* (1761), the earliest of his scores to place him among the great composers. In the same year a very talented poet-librettist-adventurer, Ranieri de' Calzabigi, reached Vienna from Paris. Falling in with the anti-Metastasian intellectuals spearheaded by Durazzo, Calzabigi thereafter brought his acquaintance with Rameau's stately operas to the writing of three librettos for Gluck, for whose signature he also drew up the renowned preface to the publication of the Calzabigi–Gluck *Alceste* in 1769. That dedication is the central document of "operatic reform," summing up one side of the unending debate between supporters of the primacy of the opera libretto and supporters of the primacy of the music. It emphasized that superfluous, florid da capo arias were to be dispensed with: simplicity of expression and emotional truth were to take their place.

The influence of *Alceste*

Although some of the most accomplished composers of opera certainly would not have agreed that, as Calzabigi–Gluck stated, the "true office" of music is "serving poetry," and though a strict obedience to all the precepts of the *Alceste* dedication would have impoverished much of the richest operatic music (Gluck himself did not follow them with iron strictness), the pronouncement unquestionably summoned opera back temporarily to its best condition: as musicotheatrical drama. The most obvious, as well as the greatest, results of the attitudes it expressed were the magnificent Calzabigi–Gluck operas first staged in Vienna: *Orfeo ed Euridice* (1762), *Alceste* (1767), and *Paride ed Elena* (*Paris and Helen*, 1770). The two earliest of these became even more stately and Rameau-like when Gluck reconstituted them to French librettos for Parisian audiences.

A little below *Orphée* and the French *Alceste* in austere dramatic force stands the somewhat mixed, only partially "reformed" *Armide* (Paris, 1777). But Gluck was to produce his best results in *Iphigénie en Aulide* (*Iphigenia in Aulis*, 1774; libretto adapted from Racine's tragedy by Bailli du Roullet) and in his masterpiece, *Iphigénie en Tauride* (*Iphigenia in Tauris*, 1779; libretto adapted from Euripides by Nicholas-François Guillard).

Gluck's extraordinary power at his best derives from a reasonable, pliable adherence to the precepts of the *Alceste* dedication; a lean sparseness of means, particularly of harmonic density, with the result that the smallest shifts arrive with great power; and the ability to make the most of the dramatic strengths of the often excellent librettos he used. Yet Gluck did not really "reform" opera, for opera is an incorrigible art form. Its mood and

manners shift with changes in society, dramaturgy, techniques of composition, and taste.

Followers and imitators of the Gluck reform

Gluck had some immediate followers and imitators, notably Antonio Salieri, who gave lessons to Beethoven, Schubert, and Liszt, was friendly with both Haydn and Rossini, and was ridiculously accused of having poisoned Mozart. Another, belated, Gluckian was Gluck's onetime "rival" in a Parisian polemic "war," Niccolò Piccinni (1728–1800), whose masterpiece, *Didon* (*Dido*, 1783; libretto by Jean-François Marmontel, Fontainbleau), joined a peculiarly Italianate melodiousness to Rameau-like solemnity and the massive simplicities of Gluck's style. Finally, there was Antonio Sacchini, remembered chiefly for his *Oedipe à Colone* (*Oedipus at Colonus*, 1786; libretto by Nicolas-François Guillard). On the highest level, however, the Gluckian "reform" produced only his own masterpieces, although it led indirectly to certain of the "international," though very French, operas of Gasparo Spontini (1774–1851), particularly *La Vestale* (*The Vestal Virgin;* Paris, 1807), and Luigi Cherubini, particularly *Médée* (1797). In part, Gluck also influenced the vast masterwork of Hector Berlioz, *Les Troyens* (*The Trojans*, composed 1856–58).

OPERA IN ENGLAND

Just as immediate acceptance of opera had been made difficult in France by the entrenched ballet and the 17th-century drama of Jean Racine and Pierre Corneille, so it was delayed in England by the court masque, an aristocratic 16th- and 17th-century entertainment derived largely from ballet. Most often dealing with allegorical and mythical subjects, the masque mixed poetic text, instrumental and vocal music, dancing, and acting. The most familiar masque is *Comus* (text by John Milton and music by Henry Lawes), which was staged at Ludlow Castle in 1634. Many other embryonic operas were produced in the middle decades of the 17th century, often being more like plays with incidental music. The first (and, in the view of many, still the finest) English opera, *Dido and Aeneas*, by Henry Purcell (*c.* 1659–95), was originally sung about 1689 by the pupils at a girls' school in London. This musical masterpiece, with libretto by a future poet laureate, Nahum Tate, contains the earliest operatic aria (apart from "Lasciatemi morire" from Monteverdi's *Arianna*) still frequently heard: Dido's lament "When I am laid in earth." In this opera, Purcell succeeded in writing a real, albeit brief, music drama, breaking down the formal barriers between recitative and song.

The first English opera

England, however, was not ready for opera. Although later Purcell works, including *The Fairy Queen* (1692), *The Indian Queen*, and *King Arthur* (1691), have been called operas, they were actually suites of incidental music for plays. No other composer in England of Purcell's genius turned his attention to opera, and before many decades had passed after the scarcely noticed performance of *Dido and Aeneas*, the rage for Italian opera (particularly when the singers included a good castrato) barred that road to English composers. The arrival of Handel from Italy in 1710 decided the direction of opera in London for many decades. Beginning in 1711 with *Rinaldo* (libretto by Giacomo Rossi, indirectly derived from Tasso's epic *Gerusalemme liberata*) and continuing intermittently until *Deidamia* (libretto by Paolo Rolli) in 1741, Handel provided English audiences with his own remarkable variant of Neapolitan opera seria, acting as both composer and, most often, impresario. The greatest composer of opera in his age, Handel eventually outlasted his popularity in London's opera houses and turned to the creation of a magnificent series of oratorios set to English texts. His operatic reign was challenged at its height only by a faction that set up the very gifted Modenese composer Giovanni Bononcini in an unequal battle against him. Handel's operas all but vanished from the repertory in the 19th century, but after a burst of German interest in them in the 1920s, they began to be increasingly revived by some major opera houses, smaller opera companies, students, musicologists, and recording companies.

An event that contributed to the defeat of Handel as an opera impresario was the London production in 1728 of *The Beggar's Opera*. That bawdy, rollicking satire in English became phenomenally popular and spawned a family of imitations that finally accustomed audiences in London and elsewhere in the British Isles to hearing a staged play sung in the vernacular.

VIENNESE MASTERS

Italian opera buffa strongly attracted Viennese audiences, and Austrian composers were naturally influenced by it. Perhaps the most interesting of the Vienna-born composers of 18th-century comic opera was Karl Ditters von Dittersdorf, whose Italianate *Doktor und Apotheker* (1786; libretto by Gottlieb Stephanie), though successful and lively, was overshadowed by the contemporary works of Mozart.

Haydn, who lacked a strong theatrical bent, nonetheless composed about 20 musicodramatic scores: a singspiel, five short operas for marionettes, and several very Italianate opere buffe and opere serie for private performance in the Eisenstadt palace theatre of his employer-patrons, the Esterhazy princes. Several of Haydn's dramatically undistinguished operas have had modern revivals, including *Il mondo della luna* (*The World on the Moon*, 1777; libretto by Carlo Goldoni), *L'isola disabitata* (*The Deserted Island*, 1779; libretto by Metastasio), and *La fedeltà premiata* (*Faithfulness Rewarded*, 1780; libretto by Giovanni Battista Lorenzi).

Mozart. Vienna, however, was to be one of the centres of the operatic career of Mozart, one of the greatest masters of opera. Mozart began to write theatrical music when only ten years old and brought out the first of his important operas at Munich in 1781, when only 25. This was *Idomeneo*. Its libretto, by Giambattista Varesco, is an imitation of Metastasio's style. But Mozart rose above the conventional operatic patterns and filled them with richly expressive music so that the result is scarcely recognizable as an opera seria. As a work of musical (though not of dramatic) art, *Idomeneo* is as fine as Gluck's *Iphigénie en Tauride* and ranks as the supreme Italian opera seria of the late 18th century.

One year after *Idomeneo*, Mozart, with the versatility of his unique genius, wrote a masterly, charming singspiel to a German text: *Die Entführung aus dem Serail* (libretto by Christoph Friedrich Bretzner as edited by Gottlieb Stephanie). A sentimental farce full of immediately attractive music and graced with a fine part for a comic bass (Osmin), *Die Entführung* also contains in "Martern aller Arten," a soprano aria so extensive in plan and difficult to sing that it has challenged the foremost sopranos down to the present. The opera has been called the greatest of all truly comic singspiele, and it is notable for the seriousness with which it treats the relationship between its two principal characters and for the human nobility of its "moral."

Mozart's next completed opera—except for *Der Schauspieldirektor* (*The Impresario*, 1786), a trifling one-act comedy—is one of the treasures of Western civilization, the greatest of all seriocomic operas, *Le nozze di Figaro* (libretto by Lorenzo da Ponte, after Pierre de Beaumarchais's *Le Mariage de Figaro*, 1786), produced at Vienna. In addition to its purely musical beauty, this work shows Mozart to be a creator of individual characters of almost Shakespearean calibre; he goes far beyond the opportunities offered by his able librettist and creates, by musical means, believable, rounded human beings, often employed in ensembles as well as in solos and in elaborately constructed finales.

Production of *Figaro* at Vienna

In 1787 Mozart's next opera, written for Prague, was *Don Giovanni* (libretto by da Ponte, based on earlier Don Juan librettos and other writings related to plays by Tirso de Molina, Thomas Corneille, and others). The 19th century tended to regard *Don Giovanni* as the greatest opera ever composed, in part because musical elements in it foretold operatic romanticism. Aspects of da Ponte's libretto disturb some 20th-century critics—particularly the grim, morally justifiable ending of what up to then has been a comedy, followed by the "vaudeville" postlude, during which the singers step out of character to

underline the "moral," in line with a then long-established convention. Musically, *Don Giovanni* shares many of the supreme virtues of *Le nozze di Figaro*, in beauty, in characterization, and in dramatic power.

In his last collaboration with Da Ponte, Mozart created another opera buffa, *Così fan tutte* (1790). Most musical opinion today considers it to be an opera of flawless workmanship reconciled with the dramatic claims of a seemingly artificial and cynical libretto, which in fact exposes the foibles of mankind. Farcical productions too often destroy its knife-edge balance between artificiality and reality. Too richly scored, too erotic, and too intense for Viennese taste of the time, it was not a success with the easy-going Viennese, but it now ranks as one of Mozart's greatest stage works.

In 1791, returning to the singspiel in German, Mozart composed *Die Zauberflöte* (libretto by Emanuel Schikaneder), an allegorical and Masonic opera with a seemingly nonsensical but in fact elaborately significant libretto in strong contrast with *Così fan tutte*'s cynicism about women. Here, Mozart created some of the most radiantly beautiful music ever composed, assigning it lavishly to both the serious and the comic, both the admirable and the vicious characters. George Bernard Shaw once said that the two arias given to the benevolent Sarastro in *Die Zauberflöte* were the only music he knew that would not sound out of place in the mouth of God—and they are but two of many numbers in this opera that have helped to place Mozart's theatrical works, together with his concertos, at the very apex of his astonishing and endlessly varied output.

Also in 1791 Mozart composed, to a Metastasio libretto slightly revised, another outdated opera seria, *La clemenza di Tito*. Created in 18 days for festivities surrounding the coronation of Emperor Leopold II as king of Bohemia in Prague, *La clemenza* provides numerous examples of Mozart's musical mastery, none of his mastery of dramatic creation through music.

Beethoven. Like *Die Zauberflöte*, Beethoven's *Fidelio* (1805, revised 1806 and 1814) rose above the limitations of the singspiel pattern, becoming something bigger and grander. In most of this, his only opera, Beethoven is musically at nearly his greatest. The libretto has never satisfied anyone entirely; also, some of the writing for the voices is instrumental rather than truly vocal. Beethoven lacked Mozart's theatrical sensibility and his ability to mix the comic–frivolous with the solemn and near tragic, as the text required. Yet the grandeur of much of the *Fidelio* music and the noble humanity of the central character (a wife, Leonore, who disguises herself as a young man, Fidelio, in order to rescue her husband, at terrible risk, from political incarceration in a dungeon) irradiates the opera from the moment of Leonore's first appearance. Its theme of the triumph of the human spirit over oppression has helped to place *Fidelio* among the world's most popular operas.

FRANCE, 1752–1825

The *Guerre des Bouffons*

Political changes and that intensity of intellectual discussion that has always played a role in determining the nature of artistic production in Paris led, in the early 1750s, to one of those polemic "wars"—this one called the *Guerre* (or *Querelle*) *des Bouffons* (War of the Buffoons)—that delight the French. This was a mainly literary confrontation of the solemn past of opera seria and *tragédie lyrique* with the farce and sentiment of opera buffa, though many of the writers saw it in nationalistic terms. It had a happy outcome: the subsequent composition by both French composers and resident foreigners of excellent examples of opéra comique, which became a French amalgam of the English ballad opera, the German singspiel, and the Italian opera buffa.

In 1752, the year of the first battles of the *Guerre des Bouffons*, Jean-Jacques Rousseau staged at Fontainebleau his one-act comic opera *Le Devin du village* (*The Village Soothsayer*). The libretto was his own. In the score he had brought together, in the pasticcio manner, melodies reflecting the very popular romances and vaudevilles being heard at the Paris fairs. It pleased battlers on both sides of the operatic war, being very French in manner and sentiment but Italian in being through-composed (continuously set) and employing recitative. Rousseau hoped to establish this combination as a standard for French comic opera, but his plan was not immediately successful. In 1755, however, a Naples-trained Italian, Egidio Duni, settled in Paris and began to compose (or perhaps, at first, merely to assemble) recognizably Rousseauesque opéras comiques. French and Belgian composers gladly adapted the new variety of opéra comique (not always with comic subject matter) and soon established its reign in the Paris and provincial theatres (nearly all of them simultaneously composed other sorts of musicodramatic works as well). Among them, several names stand out, together with the finest or most renowned of their operas. One of the most interesting of these opéra comique composers was François-André Danican (1726–95), called Philidor, also a famous chess player, who wrote about 20 works in the evolving manner.

More sentimental—in fact, tending toward the tenderly tearful—was Pierre-Alexandre Monsigny (1729–1817), never thoroughly trained in music but endowed with the ability to create winning melodies and to exploit for dramatic purpose the timbres of individual instruments. Probably the finest of the 18th-century composers of opéra comique was a Belgian, André Grétry (1741–1813), who most happily balanced the French and Italian styles. He was a very original and extremely productive composer over a 30-year period spanning the French Revolution.

Étienne-Nicolas Méhul (1763–1817), who used opéra comique conventions including the retention of spoken dialogue, also had a career spanning the Revolution. A devoted Gluckian, Méhul had composed numerous works in many genres when, in 1807, he produced his masterpiece, *Joseph*, which is a rarity among operas in several ways: its libretto by Alexandre Duval is derived from the Bible, a source of drama usually reserved for the oratorio; it has no female characters; and it mixes the most solemn classicism with the sentiment of the popular romances.

Boieldieu's immense success

Early in the 19th century, French opéra comique achieved a new equilibrium of classical tint in the best works of François-Adrien Boieldieu (1775–1834). He won truly astonishing and, to several composers (Berlioz included), maddening popularity with *La Dame blanche* (*The White Lady*, 1825; libretto by Eugène Scribe, based upon Sir Walter Scott's novels *Guy Mannering* and *The Monastery*). There were 1,669 performances of this work at the Opéra-Comique, Paris, between 1825 and 1926.

Le Pré aux clercs (*The Field of Honour*, 1832; libretto by François de Planard), the most accomplished opera of Louis-Joseph-Ferdinand Hérold (1791–1833), all but equalled the popularity of *La Dame blanche*; it had received 1,600 performances at the Paris Opéra-Comique by 1939. Hérold's other outstanding success was *Zampa* (1831; libretto by Anne-Honeré Mélesville), so much like the work of Carl Maria von Weber that it became vastly popular in Germany. An extraordinarily prolix composer, Hérold never succeeded in working out a dependable, unified manner of his own, with the result that his French operas are structurally weak. Opéra comique after Boieldieu became more Italianized, reflecting very largely Rossini's influence.

ITALY IN THE FIRST HALF OF THE 19TH CENTURY

The splendid musical achievements of the classicial Viennese style during the late 18th century and early 19th threatened to leave Italy, opera's native home, to one side of the operatic highroad. Two accidents—one the voluntary expatriation to northern Italy of a German, Johann Simon Mayr, the other the unpredictable eruption of a genius, Gioacchino Rossini—saved the day for Italian opera in Italy and outside it.

Mayr, known in Italy as Giovanni Simone Mayr, composed nearly 70 operas in Italian between his first (1794) and his last (1815). He appears to have been influenced deeply by Mozart, and his intensely keen dramatic sense, together with the extraordinary pliability with which he employed the conventions of opera seria and his varied

use of the orchestra (particularly of solo woodwinds and horns within it), would have made him a major composer of opera had he not lacked major gifts as a melodist. Many of his operas were for a long time extremely popular throughout Italy, and his immediate influence was beneficial, particularly on the practice of his most famous pupil, Gaetano Donizetti, and on another, less talented, but still admirable operatic composer, Saverio Mercadante.

The operas of Nicola Zingarelli (1752–1837) and of Ferdinando Paer (1771–1839) were transitional, between Classical and grand opera in mode and manner. Zingarelli's conventional opere buffe displayed a genuine humour and some liveliness of musical imagination; his most enduringly performed work, however, was an opera seria, *Giulietta e Romeo* (1796; libretto by Giuseppe Maria Foppa). He is now remembered chiefly as a teacher of Vincenzo Bellini. Paer, who composed more than 40 operas, worked mostly in Vienna, Dresden, and Paris; his musical style changed with his surroundings.

Rossini. The production at Venice in 1810 of the first performed opera of Rossini, *La cambiale di matrimonio* (*The Bill of Marriage;* libretto by Gaetano Rossi), announced a new operatic genius. Into the genteel, often charming atmosphere of lingering 18th-century operatic manners, Rossini brought genuine originality marked by rude wit and humour and an entire willingness to sacrifice all "rules" of musical and operatic decorum. Both his opere buffe and his now sadly neglected opere serie soon became so popular throughout Italy and then throughout the Western world that they all but blotted out his unfortunate contemporaries—Donizetti and Bellini excepted.

Rossini's originality and wit

Rossini's dazzling career marked the zenith of the bel canto style, a singer-dominated manner of composition (and at times improvisation) that played to audiences' delight in vocal agility, smoothness of voice, and long, florid phrasing. From the period of Rossini's greatest Italian triumphs (he had a second career in Paris) and of Donizetti and Bellini come the names of now legendary voices such as Isabella Colbran (Rossini's first wife), Giuditta Pasta, Maria Malibran, Giovanni David, Giovanni Battista Rubini, Domenico Donzelli, Antonio Tamburini, and Luigi Lablache. For appearances by these singers, composers altered their scores; when they sang, they interpolated extraneous arias that displayed their prowess. Rossini tried to insist that his operas be sung as he himself composed or revised them, but it was a losing battle. The polished artistry and extreme technical training and technique of such singers, as well as their extraordinarily wide ranges, have left performance of the bel canto operas bristling with nearly insoluble problems for latter-day singers.

Rossini's most famous opera is *Il barbiere di Siviglia* (*The Barber of Seville,* 1816; based on the libretto by Cesare Sterbini after the 18th-century play by Beaumarchais), the most nearly flawless of all opere buffe. Several others among his comedies rank only a little lower in musical invention, genuine comic brio, and opportunities for trained singers of vocal display and for farcical characterization: *L'Italiana in Algeri* (*The Italian Girl in Algiers,* 1813; libretto by Angelo Anelli), *Il Turco in Italia* (*The Turk in Italy,* 1814; libretto by Felice Romani); *La cenerentola* (*Cinderella,* 1817; libretto by Jacopo Ferretti), and the half Italian opera buffa and half French opéra comique *Le Comte Ory* (*Count Ory,* 1828; libretto by Scribe and Charles-Gaspard Delestre-Poirson). Rossini prefaced several of these operas with swift, witty overtures that have held a place in the repertory of symphony orchestras.

Perfection of Rossini's comic operas

His style began to have a more serious bent with *Otello* (*Othello,* 1816; libretto by Francesco di Salsa), the opera semiseria (a serious opera with a happy ending), *La gazza ladra* (*The Thieving Magpie,* 1817; libretto by Giovanni Gherardini), and *Armida* (1817; libretto by Giovanni Schmidt), all of which show the composer adapting his florid style to more dramatic, and often eloquent, purposes. But it was only in his Parisian pieces, such as *Semiramide* (1823; libretto by Gaetano Rossi), *Le Siège de Corinthe* (*The Siege of Corinth,* 1826; a revision of the earlier opera, *Maometto II* [1820]; libretto by Alexandre Soumet and Luigi Balocchi), and *Guillaume Tell* (*William Tell,* 1829; libretto by Étienne de Jouy and Hippolyte Bis), his last opera, that his talent for works on a bigger scale, presaging Parisian grand opera, found its full flowering. Some of these later operas owe their revival in the middle of the 20th century to the appearance of a few singers able to project meaningfully their difficult vocal lines.

Donizetti. Gaetano Donizetti first composed a series of very Rossinian, well-made, and largely undistinguished operas but gradually developed dramatic strength, a latent gift for memorable melody, and a forceful deployment of the orchestra for theatrical drama. In 1830, the year after Rossini's farewell to operatic composition, Donizetti produced at Milan the first of his forward-tending, less Rossinian, dramatically remarkable operas: *Anna Bolena* (*Anne Boleyn*), with a libretto by Felice Romani, who worked with so many opera composers of the time. It immediately placed him with Bellini as an inevitable successor to Rossini. What became clear only in retrospect was that it also showed him to be the most important immediate predecessor—in some sense, teacher—of Giuseppe Verdi. Donizetti clung to the long, legato (smoothly flowing) melodies and the ornamented vocal lines of bel canto, but he also unmistakably foreshadowed Verdi's dramatic vigour and many of the younger man's compositional methods. Several unconscious borrowings from Donizetti have been noted by students of Verdi's operas.

Like Rossini, but unlike Bellini, Donizetti moved freely back and forth between serious and comic subjects. He composed about 70 stage works in 25 years. After the success of *Anna Bolena,* with speed and facility that remain astonishing, he composed numerous operas of enduring quality. They include the sentimental comedy *L'elisir d'amore* (*The Elixir of Love,* 1832; libretto by Felice Romani); *Lucrezia Borgia* (1833; libretto by Romani) and *Maria Stuarda* (*Mary Stuart,* 1834; libretto by Giuseppe Bardari); the popular *Lucia di Lammermoor* (1835; libretto by Salvatore Cammarano was derived from Sir Walter Scott's *The Bride of Lammermoor*)—an opera that reflects Donizetti's acquaintance with the music of Bellini; *Roberto d'Evereux* (1837; libretto by Cammarano); the delightful opéra comique *La Fille du régiment* (*The Daughter of the Regiment,* 1840; libretto by Jules-Henri Vernoy de Saint-Georges and Jean-François-Alfred Bayard); the grand opera *La Favorite* (1840; libretto by Alphonse Royer, Gustave Vaëz, and perhaps Scribe); the opera semiseria *Linda di Chamounix* (1842; libretto by Gaetano Rossi); and—judged by many to be Donizetti's masterwork—the ever fresh and vivid opera buffa, *Don Pasquale* (1843; libretto by Giacomo Ruffini and Donizetti).

Bellini. Altogether different from either Rossini or Donizetti was Vincenzo Bellini. His operas have come to seem the natural habitat of bel canto, of the unchallenged supremacy of vocal melody in amazingly long-breathed and highly decorated lines. Only his first student opera contains even a trace of humour. He and his librettists filled their collaborations with intensely amorous and other subjective emotion, ethical confrontations, and usually tragic involvements (of the seven finest among his ten operas, only two—*La sonnambula* and *I puritani*—conclude happily for the principal characters). Bellini cultivated with meticulous care his unrivalled, native gift for convincingly melancholy melody, especially in arias and duets; he gave much less attention to ensembles, choruses, and the expressive potentialities of the orchestra. His orchestra, in fact, might have been what it was had Haydn, Mozart, Mayr, and Rossini never existed.

Bellini's innate gift of melody

Beginning in 1827 with *Il pirata* (*The Pirate,* libretto by Felice Romani, who thereafter supplied all of Bellini's librettos except that for *I puritani*), Bellini made his presence felt throughout Italy and then gradually throughout Europe and the Americas. In 1831 two of Bellini's enduring masterworks were produced: the pastoral opera semiseria *La Sonnambula* (*The Sleepwalker*) and the heroic tragedy *Norma.* Bellini's previously faith-

ful public temporarily deserted him when, in 1833, he gave it *Beatrice di Tenda.* In the year of his death after his final removal to Paris, he won another triumph with an opera very loosely connected with Cromwellian times in England, *I puritani* (*The Puritans;* libretto by conte Carlo Pepoli).

Unlike Rossini and Donizetti, Bellini exercised little or no influence upon the style of his successors: the most noticeable of his compositional means, his exclusively personal sort of melody, could only be debased in imitation by others. With these three men, both the late period of bel canto and the second period of opera buffa drew to a close. After the onset of Donizetti's crippling illness in 1847, the Italian opera houses in Italy, Paris, London, and elsewhere could look to Giuseppe Verdi.

III. Grand opera and beyond

FRENCH GRAND OPERA

Nineteenth-century Paris was to foster and witness the birth of "grand opera," an international style of large-scale operatic spectacle employing historical or pseudo-historical librettos and filling the stage with elaborate scenery and costumes, ballets, and phalanxes of supernumeraries. Dispensing almost entirely with the delicacies of bel canto, it vastly enlarged both the orchestra itself and its role in the dramatic happenings. Grand opera naturally had roots in the past, particularly in the Venetian "machine operas" of the 17th century, such as Cesti's *Il pomo d'oro,* as well as in the stately scores of Rameau and Gluck. But the immediate drive toward this new style of opera was instituted in Paris by Italian expatriates: Luigi Cherubini, who spent the last 54 years of his life in France, and Gasparo Spontini, whose most impressive operas were designed for Paris.

Cherubini was a greatly learned composer in almost all musical forms who won the admiration of Beethoven. His two most imposing operas were the embryonic grand opera *Médée* (1797; libretto by François-Benôit Hoffman) and a *comédie lyrique, Les Deux Journées* (*The Two Days,* 1800; libretto by Jean-Nicolas Bouilly). *Les Deux Journées* became something like a national German opera under its German title, *Der Wasserträger* (*The Water Carrier*). Spontini, in his French operas, ranged far beyond Cherubini and his other contemporaries in his demands for complex staging, finally reaching a sort of splendid megalomania. Daniel-François-Esprit Auber brought out a nearly total grand opera; *La Muette de Portici* (*The Mute Girl of Portici,* also known as *Masaniello,* 1828; libretto by Scribe). The popularity of *La Muette* became phenomenal in both France and Germany. This opera remains unique on several counts. Its title character neither sings nor speaks, the role being performed by a mime. A performance of it at Brussels on August 25, 1830, set off disorders that led to the separation of Belgium from The Netherlands. Eighteen months after the premiere of Auber's opera, the appearance of Rossini's *Guillaume Tell* showed that master of opera buffa and bel canto responding to the new genre. Auber's later operas include several charming comedies, among them *Fra Diavolo* (1830; libretto by Scribe).

Meyerbeer. The final, official birth of grand opera occurred in 1831, with the first French opera of another Parisian expatriate, the German Giacomo Meyerbeer: *Robert le Diable* (libretto by Scribe and Germain Delavigne). The popularity of this work became a sort of frenzy (by August 1893 it had been sung 751 times at the Paris Opéra). Although Meyerbeer's operas are rarely performed in the later 20th century, he remains a controversial figure. Using an expanded, powerful orchestra, with much emphasis placed on individual instrumental colours, requiring almost every kind of singing, filling huge stages with dazzling pageantry, employing characters who pretend to be actual figures from history, four of his operas held their leading positions even through the Wagnerian revolution and into the early 20th century. Besides *Robert le Diable,* they were *Les Huguenots* (1836; libretto by Scribe with the collaboration of Émile Deschamps), *Le Prophète* (1849; libretto by Scribe), and the posthumously staged *L'Africaine* (libretto by Scribe). The author of all of these, Eugène Scribe, was the most phenomenally productive librettist of his time, writing (with the help of various collaborators) a huge number of librettos for many composers, including Auber, Boieldieu, Cherubini, Donizetti, Gounod, Halévy, Meyerbeer, Rossini, and Verdi. He was, in fact, a major force in the evolution of French grand opera.

Imitators of Meyerbeer's successes naturally sprang up immediately. Later, numerous men totally unlike him stylistically—including Berlioz, Wagner, and Verdi—were influenced unwittingly by his practices. The first of the imitators was Fromental Halévy, whose works included at least one grand opera that could almost be mistaken for Meyerbeer's: *La Juive* (*The Jewess,* 1835; libretto by Scribe). After the times of Meyerbeer and Halévy, grand opera began to respond to new musical and intellectual currents, evolving into a variety of mixed forms.

Berlioz. Like most of Hector Berlioz's other compositions, his three operas stand apart from the mainstream of historical evolution. When first staged at the Paris Opéra in the shadow of *Robert le Diable* and *La Dame blanche,* his first opera, *Benvenuto Cellini* (1838; libretto by Léon de Wailly and Auguste Barbier), was a complete failure. The second, the lighthearted *Béatrice et Bénédict* (his own libretto, based upon Shakespeare's *Much Ado About Nothing*), finally was given its premiere at Baden-Baden in 1862 by Franz Liszt. And Berlioz's masterpiece, *Les Troyens* (his own libretto), is based on Virgil's *Aeneid* and divided into *La Prise de Troie* (*The Capture of Troy*), two acts, and *Les Troyens à Carthage* (*The Trojans at Carthage*), three acts. It was not performed complete during his lifetime: he heard only the second part as staged in Paris in 1863. Mid-20th-century complete (or very nearly complete) performances of *Les Troyens,* notably in London, showed it to be a great, noble, idiosyncratic work not without traces of grand opera, but in seriousness and scope much closer to the Wagner of *Der Ring des Nibelungen.* Berlioz's operas, like his other music, are distinguished by the individual arch of his melody, his revolutionary orchestration, and the dramatic thrust of the whole.

Modern revival of *Les Troyens*

Offenbach. Even more popular than Auber as a purveyer of light operatic comedy was Jacques Offenbach, a German émigré to Paris who supplied the Second Empire and the early years of the Third Republic with a long series of very tuneful, witty, and satiric operettas of deliberate frivolity. Remembered among them are *Orphée aux enfers* (*Orpheus in the Underworld,* 1858; libretto by Hector Crémieux and Ludovic Halévy), *La Belle Hélène* (*Beautiful Helen,* 1864; libretto by Henri Meilhac and Halévy), *Barbe-Bleue* (*Bluebeard,* 1866; libretto by Meilhac and Halévy), *La Vie Parisienne* (*Parisian Life,* 1866; libretto by Meilhac and Halévy), *La Grande-Duchesse de Gérolstein* (1867; libretto by Meilhac and Halévy), and *La Périchole* (1868; libretto by Meilhac and Halévy). Left incomplete at Offenbach's death was his major serious opera, *Les Contes d'Hoffmann* (*The Tales of Hoffmann;* libretto by Jules Barbier and Michel Carré, after their play of the same name based on tales by the German Romantic writer E.T.A. Hoffmann). Recitatives replacing the original dialogue were provided by Ernst Guiraud, and the opera was staged posthumously in 1881. This fantasy involving supernatural interventions rapidly became a worldwide favourite.

GERMAN ROMANTIC OPERA

Weber. Romanticism, part philosophical, part literary, part aesthetic, made one of its first appearances, and certainly its earliest overt appearance, in opera, in three works composed between 1821 and 1826 by Carl Maria von Weber. Beginning with his masterpiece, *Der Freischütz* (*The Freeshooter,* 1821; libretto by Friedrich Kind), Weber successfully challenged the outdated dictatorship of Spontini at Berlin. For the Italian's stiff grandeurs he substituted, in singspiel form, tender sentiment, grisly horrors, manly choruses, moral nicety, and music of extraordinary instrumental and vocal allure. *Der Freischutz* illustrates the German Romantic writers' love

Elements of German Romanticism

for dark forests, the echoes of hunters' horns, the threatening supernatural, the frustrations of pure young love. Its popularity in Germany and elsewhere was enormous.

Weber smarted under the anti-Romantic criticism of *Der Freischütz* as a mere singspiel (a work with spoken dialogue) rather than a musically continuous opera. His next major composition, *Euryanthe* (1823; libretto by Helmina von Chézy), was something like a proto-grand opera and therefore contained no spoken dialogue. Almost since its premiere, writers have attacked the remarkable silliness (on paper) of its libretto, but most of them have never witnessed the work in performance and therefore cannot judge how the libretto works on stage with Weber's fine score. His last opera, *Oberon*, or *The Elf King's Oath* (1826; libretto, in English, by James Robinson Planché), returns to the singspiel form. Like *Euryanthe*, it has not held the stage, and again the libretto has been blamed. The overtures to all three of these operas are still played frequently, and whatever future opinion of the operas themselves may be, *Der Freischütz* opened the floodgates of musical Romanticism in Germany.

Spohr, Lortzing, and others. Louis Spohr (1784–1859), a violinist, conductor, and composer of instrumental music, sounds pallidly Romantic if compared with Weber, but certain of his harmonic innovations taught something to Wagner, of whose early operas he was a defender. Heinrich August Marschner (1795–1861), more Romantic by nature than Spohr, borrowed sufficiently from Weber's style to serve as one bridge to Wagner. He displayed talent as orchestrator and melodist, and he applied his gifts to intensely Romantic and equally Teutonic librettos. The finest of his now unheard operas is *Hans Heiling* (1833; libretto by Eduard Devrient).

The other German-language composers of opera active during the Weber–Spohr–Marschner period were less important. Albert Lortzing composed several operas that have been likened to genre painting. He travelled in the direction of operetta in his popular sentimental comedies, such as *Zar und Zimmermann* (*Tsar and Carpenter*, 1837; his own libretto) and *Der Waffenschmied* (*The Armourer*, 1846; his own libretto). The same direction was taken by Friedrich, Freiherr von Flotow, whose operetta-like *Martha* (1847; libretto by Friedrich Wilhelm Reise) continues to be performed frequently. This trend toward operetta as a less intense variety of Romanticism continued in *Die lustigen Weiber von Windsor* (1849; libretto by Salomon Hermann Mosenthal, based on Shakespeare's *Merry Wives of Windsor*), the major success of Otto Nicolai, and in the extremely popular works of Franz von Suppé, a Dalmatian of Belgian ancestry. It culminated in operetta on the highest level of musical accomplishment in the masterworks of Johann Strauss the Younger. Many of Strauss's operettas are known now only by their overtures and waltzes, but one of them, *Die Fledermaus* (1874; libretto by Carl Haffner and Richard Genée), has never left the stage for long. Only the finest opéras comiques and opéras bouffes of Auber and Jacques Offenbach match Strauss's elegance, wit, humour, musical invention, and scrupulous workmanship.

Strauss's operettas

VERDI

When, in 1839, an opera called *Oberto, conte di San Bonifacio* (libretto by Antonio Piazza, revised by Bartolomeo Merelli and Temistocle Solera) was staged at the leading Italian opera house, the Teatro alla Scala (La Scala) at Milan, its first audiences received it reasonably well. Rossini had not offered a new opera for ten years, Bellini was dead, and Donizetti was composing for Paris, so the debut of a new talent was welcome. Those early audiences, however, could not know that *Oberto* had opened the active career of the greatest of all later Italian composers of opera, Giuseppe Verdi. Verdi's second opera, *Un giorno di regno* (*King for a Day*, 1840; libretto by Felice Romani), was a failure and was to remain his only comedy for 53 years. It was followed by *Nabucodonosor*, known as *Nabucco* (1842; libretto by Solera), which displayed the emergence of a musical dramatist of enormous vigour and rich melodic invention.

Verdi long suffered from his inability to obtain librettos worthy of his special talents, but each of the six operas that he wrote between *Nabucco* (1842) and *Macbeth* (1847) includes scenes and numbers of great power and immediately winning, memorable melody. Even *Macbeth* (libretto by Francesco Maria Piave, revised in 1865 by Verdi to a libretto in French), although it is marked both dramatically and musically by passages of astonishing vitality, has structural weaknesses.

Another period of lesser achievement by Verdi stretched from 1847 to 1851, the best of his five operas of those years having been *Luisa Miller* (1849; libretto by Salvatore Cammarano). Meanwhile, Verdi was on the way to becoming a public symbol of the risorgimento, the Italian movement of rebellion against foreign domination and toward political unification, both because of the patriotic emphasis in several of his librettos and because of his staunchly liberal public character (he was eventually to become a true national hero).

Beginning in 1851, Verdi produced three masterpieces, having found three librettos that fired aspects of his genius. The first of them was *Rigoletto* (libretto by Piave), in which his abundant creation of melody was at the service of his gift for musical characterization. *Rigoletto* was followed, less than two years later, by *Il trovatore* (*The Troubadour*, 1853; libretto by Cammarano), perhaps unmatched among his operas for its profusion of strong and memorable melodies.

Less than two months after the premiere of *Il trovatore* came *La traviata* (1853; libretto by Piave, after Alexandre Dumas *fils*'s *La Dame aux camélias*). At first a failure, it later came to be accepted as a masterpiece. It also established a composer's right to set librettos dealing with contemporary life and with characters not of exalted station. By comparison with the thunderous melodrama of *Il trovatore*, *La traviata* seems an intimate, quiet, almost chamber-music opera. The musical portrait of Violetta, the tubercular courtesan heroine, remains extraordinary for its depiction of the effects of love and sorrow on her character.

Production of *La traviata*

After *La traviata* came a comparative failure, a grand opera composed in French for Paris, *Les Vêpres siciliennes* (*The Sicilian Vespers*, 1855; libretto by Scribe and Charles Duveyrier). It was succeeded only two years later by *Simon Boccanegra* (1857; revised 1881), a gloomy opera of great power. Then came *Un ballo in maschera* (*A Masked Ball*, 1859; libretto by Antonio Somma, after Scribe's libretto for Auber's 1833 opera *Gustave III, ou Le Bal Masqué*), no less gloomy and powerful but including, in the page Oscar (sung by a soprano), a ray of light and youthful humour, and *La forza del destino* (*The Force of Destiny*, 1862; libretto by Piave), a kaleidoscopic mixture of the tragic and the farcical, with touches of matter not far from operetta and opera buffa.

For Paris, and again to a libretto in French, Verdi next composed *Don Carlos* (1867; libretto by François-Joseph Méry and Camille du Locle, revised by Verdi in 1884 to an Italian translation and again in 1887). This long opera, and particularly its fourth act, is majestic and subtle, its various musical confrontations—many of them duets—displaying a depth of characterization hitherto unknown in Italian or French opera, in spite of faults in the libretto.

By 1869 Verdi's fame had become so international that the Khedive of Egypt invited him to compose an opera for Cairo to mark the opening of the new Cairo Opera House (and possibly the opening of the Suez Canal). In fact, the canal began to operate in 1869, but the opera received its premiere at Cairo only in 1871. This was *Aida* (libretto by Antonio Ghislanzoni, based on a scenario by Auguste Mariette, the French Egyptologist, and Camille du Locle, with the collaboration of Verdi). The masterly libretto and its four well-delineated principal characters evoked from Verdi a Meyerbeerian opera of such unfailing melodic, orchestral, and dramatic richness that many have called *Aida* his masterpiece. For pageantry, combined as it is with harmonic, melodic, and instrumental skills and convincing, if generalized, characterization, it remains

unrivalled—and probably has been sung more often than any other opera.

Verdi's Shakespearean operas

In 1869 the public and the writers on opera assumed that Verdi would continue to produce a new opera every few years. But 16 elapsed before the premiere of his next opera, *Otello* (1887; libretto by Arrigo Boito). Verdi's varied, intensely dynamic, compressed, and tragic score was the result not only of his ripened genius but also of nearly 50 years of operatic practice. Many critics consider it the finest tragic opera ever composed.

In the following six years, rumours grew that the aged Verdi not only was composing still another opera but that it was to be a comedy. The comic masterpiece *Falstaff* (libretto by Boito, derived largely from Shakespeare's *Merry Wives of Windsor* and *Henry IV*) was performed in 1893 at La Scala, where Verdi's first opera had been staged more than 53 years before. An opera buffa with serious overtones, *Falstaff* always has been praised by critics and enthusiasts, but it has never become a true popular favourite.

Arrigo Boito not only wrote the librettos of Verdi's last two operas but was also himself a composer, as well as a poet, polemicist, and man of letters. He completed only one opera, *Mefistofele* (1868; his own libretto, derived from Goethe's *Faust*). It was at first a failure and then became more popular. His command of technical musical resources was vast. Unfortunately, he placed them at the service of lofty but diffuse philosophical concepts and ideals mostly beyond his range of expressiveness. *Mefistofele*, for instance, is more impressive and admirable than theatrically convincing.

WAGNER AND HIS SUCCESSORS

Richard Wagner (1813–83) is a unique figure in the history of both opera and music. A concentrated egoist gifted with a powerful, tenacious, and at times stubbornly confused intellect, he wrote both the music and librettos of his operas. He began his career, except for a youthful attempt, *Die Feen* (*The Fairies*, completed, 1834; first performed, 1888), with two grand operas mixing the influences of Meyerbeer, Marschner, and Weber: *Das Liebesverbot* (*The Ban on Love*, 1836) and *Rienzi* (1842). In 1843, with *Der fliegende Holländer* (*The Flying Dutchman*), he began to develop what was to become an extremely personal, powerful manner of operatic construction. Turning to mythic legend for his subjects and making unacknowledged bows to the operas of Weber and Marschner, while dispensing with the trappings of grand opera, he composed an intensely German, Romantic opera. In it he instituted the use of brief melodic and other motifs as materials for evolving a more-or-less continuous web of music in which the separate numbers of earlier opera appeared only when the libretto demanded them. Already, at the age of 30, he was giving harmony, in very unclassical guise, a central constructive role in the creation of both drama and characterization.

Wagner's personal operatic style

Patiently and challengingly elaborating a vast, interlocked system of theories in many published books and essays, Wagner continued the ripening of his style in two large, transitional operas, *Tannhäuser* (1845) and *Lohengrin* (1850). *Tannhäuser* again displays some grand-opera characteristics (particularly in the revision of it that Wagner prepared for the 1861 Paris performance); *Lohengrin*, the last of Wagner's serious operas peopled by human beings of recognizable dimensions, has been called the Romantic opera par excellence.

The earliest example of what Wagner called "music drama" (a term he preferred to "opera") was the monumental *Tristan und Isolde* (1865), the libretto of which illustrates his obsession with the idea of man's redemption through woman's love. *Tristan und Isolde* advances harmonic language. The score is woven in a harmonic idiom so advanced chromatically that it speeded the destruction of orthodox concepts of harmony. *Tristan* requires singers possessed of powerful voices capable of penetrating a vastly enlarged orchestra. It came to be regarded as the greatest German opera of the late 19th century, and its influence upon compositional methods and techniques continued into the 20th century.

In *Die Meistersinger von Nürnberg* (*The Mastersingers of Nürnberg*, 1868) he partly deserted his most "advanced" style because central episodes in the libretto required self-contained numbers. Warmhearted, overflowing with young love and the bitter wisdom of age, *Die Meistersinger* ranks with Verdi's *Falstaff* among late-19th-century comic operas. From 1853 until 1874 Wagner worked intermittently on the four poems and the scores of *Der Ring des Nibelungen*. It is an epic, based on Teutonic myths, of such proportions and implications that it cannot be summarized. Musically, Wagner uses leitmotiv—constantly recurring fragments—and weaves them into a large, elaborate pattern. Performed in its entirety, and without intermissions, the *Ring* would last about 12 hours; many listeners, almost bewitched, would have been and now would be willing to attend such a continuous performance, so compelling is the musical power of this unique, all but inhuman, masterpiece.

The last of Wagner's operas, *Parsifal* (1882), introduced few structural elements not used in *Tristan und Isolde* and *Der Ring des Nibelungen*. Wagner called it *Ein Bühnenweihfestspiel*—a sacred festival drama—and it is heavy with religious and ethical messages. It perfectly illustrated both his musicodramatic theories and the unsmiling solemnity with which he approached operatic composition and demanded, successfully, that his audience absorb the results. *Parsifal* closed a career unparalleled in its aggressive demands on society and on opera goers, a career that would have seemed a form of madness if it had not produced some of the most lastingly impressive of all operatic creations and many pages of music of the greatest beauty.

Operatic composers influenced by Wagner

Curiously, Wagner's influence has been felt more in the evolution of post-Romantic harmony than in the constructive practices of later operatic composers. An adaptation of his leitmotiv usage marked the delightful fairy-tale opera *Hänsel und Gretel* (1893; based upon the tale by the Brothers Grimm), by Engelbert Humperdinck. Wagner's early theories about both libretto and music played a constructive part in the excellent comic opera *Der Barbier von Bagdad* (*The Barber of Bagdad*, 1858), by Peter Cornelius, and his mature style was wholly adopted in *Guntram* (1894), the first opera of Richard Strauss. Otherwise, the Wagnerian revolutions (in contrast with his at times unrevolutionary practice) are clearly seen in the operas of some ardent French Wagnerians such as Ernest Reyer and Vincent d'Indy.

LATER OPERA IN FRANCE

Gounod. The history of French opera contemporary with and later than Berlioz includes many talented composers and stageworthy works, but it degenerates quickly into a catalog of pieces of considerable charm and some originality seldom made arresting by the appearance of operas unmistakably of the first rank. Charles Gounod, who composed many operas, had a unique gift for sentimental melody but an uncertain sense of what is theatrically viable. In his ever-popular *Faust* (1859; libretto by Jules Barbier and Michel Carré), his talents were most creatively gathered together, but *Faust* took no place in the steady evolution of operatic styles. Among Gounod's other operas, the best, mixing in different proportions the virtues and flaws of *Faust*, are probably *Mireille* (1864; libretto by Carré, derived from Frédéric Mistral's Provençal poem *Mirèio*) and *Roméo et Juliette* (1867; libretto by Barbier and Carré).

Bizet. The works of Georges Bizet are on a higher level in both their vigour and variety. He began to display his considerable ability with *Les Pêcheurs de perles* (*The Pearl Fishers*, 1863; libretto by Eugène and Carré) and *La Jolie Fille de Perth*, (1867; libretto by Jules-Henry Vernoy de Saint-Georges and Jules Adenis, based on Sir Walter Scott's *Fair Maid of Perth*). In 1875 Bizet produced his masterpiece, *Carmen* (libretto by Henri Meilhac and Ludovic Halévy, after a tale by Prosper Mérimée). Its then savage realism, broad but convincing characterization, and dazzling pseudo-Spanish ambience shocked its first audiences and strongly influenced Italian *verismo* (see below *Later opera in Italy: Verismo*). A

Production of *Carmen*

lonely masterpiece, *Carmen* remains one of the steady props of operatic repertory everywhere.

Thomas and others. Ambroise Thomas, as prolific as Gounod but not as gifted in emotional or musical persuasion, had composed many operas when Paris first welcomed his *Mignon* (1866; libretto by Barbier and Carré), probably his best opera. Two years later he composed *Hamlet* (1868; libretto by Barbier and Carré). Thomas, like Gounod, interlarded his operas with florid, often essentially undramatic and "showy" arias for a new type of lyric-coloratura soprano. One of the most frequently heard of such vacuous arias is the "Bell Song" from Léo Delibes' *Lakmé* (1883; libretto by Edmond Gondinet and Philippe Gille). Although Camille Saint-Saëns composed numerous operas, the only work by him to remain in the repertory is *Samson et Dalila* (libretto by Ferdinand Lemaire), originally sung under Franz Liszt's aegis in 1877 in German. It is lusciously melodic but so lacking in drama as to seem half oratorio.

Massenet. Phenomenally popular when first composed were many of the operas of Jules Massenet, who had a surer sense of the stageworthy than Saint-Saëns but often made saccharine what to many had seemed too sweet in the earlier French opera of the 19th century. At his best however, Massenet was something better than "the daughter of Gounod": in *Manon* (1884) he produced not only one of the most enduringly popular of operas but also a stylistically unflawed reflection of the tragico-sentimental 18th-century novel by the Abbé Prévost on which Henri Meilhac and Philippe Gille had based the libretto. Much the same qualities have kept alive Massenet's *Werther* (1892; libretto by Édouard Blau, Paul Milliet, and Georges Hartmann, derived from Goethe's *Leiden des jungen Werthers*), first performed at Vienna in a German translation. Some other operas by Massenet, particularly *Thaïs* (1894; libretto by Louis Gallet, after the novel by Anatole France), are important for their sensuous portrayals of seductive female characters.

Gustave Charpentier's *Louise* (1900; libretto by the composer) has remained in opera house repertories because of its loving, romanticized portrait of "Bohemian" Paris, the sentiment and surface allure, and the popularity of Louise's hymn to love, "Depuis le jour."

Debussy. Claude Debussy, the greatest French composer after Berlioz and a decisive influence upon 20th-century music, completed only one opera: *Pelléas et Mélisande* (1902), an almost verbatim setting of Maurice Maeterlinck's play. Like *Tristan und Isolde*—which, having helped to shape Maeterlinck's drama, inevitably and against Debussy's will also shaped some of his compositional procedures—*Pelléas* remains unique. Listeners immune to the attraction of its aristocratic sensuality often find it monotonous. Nonetheless, it is a masterwork in its wholly apt amalgamation of text and score, the inescapable rightness for its quiet dramatic purposes of Debussy's individual harmonies, the marvelous manner with which he made the sounds of Maeterlinck's French an integral element in a shimmering orchestral web. In *Pelléas*, the Wagnerian ideal of continuous music without separate numbers was attained. Although *Pelléas* remains one of the handful of important operas composed in the 20th century, it has had few descendants. One of those few is Paul Dukas' *Ariane et Barbe-Bleue* (*Ariadne and Bluebeard*, 1907)—like *Pelléas*, an almost verbatim setting of a Maeterlinck play. It, however, is notably noisy where *Pelléas* is quiet, and it lacks Debussy's thematic invention.

The achievement of *Pelléas*

Ravel. Maurice Ravel wrote one opera, *L'Heure espagnole* (*The Spanish Hour*, 1911; libretto by Maurice Legrand), and a *fantaisie lyrique* (really a ballet-pantomime-opera), *L'Enfant et les sortilèges* (*The Child and the Enchantments*, 1925; text by Colette). The former is opéra bouffe in Spanish dance rhythms overlaid with vocal lines that seem indebted to the works of Richard Strauss. The latter reverses the orchestra's domination over the voices in *L'Heure espagnole*. It is an edifying and hilarious fantasy about a child being punished for his mistreatment of his toys and other objects. Perhaps only its uncomfortable mixture of genres has kept it from widespread performance.

Les Six. Of the professedly anti-Debussyan, anti-Impressionist group known as Les Six, three have places in the history of opera: Arthur Honegger, Darius Milhaud, and Francis Poulenc. Honegger, a Swiss, employed a somewhat Teutonic "neoclassical modern" idiom in thickly dissonant, heavily percussive operas that have failed to hold the stage. Milhaud's once sensational *opéras-minutes* (1927–28), three brief one-act compositions, were first staged in German translation in Germany. They parody Greek myths and have music composed in a jazz manner that inevitably has lost effectiveness with the passing decades. Milhaud also composed large modern versions of grand opera, but his persistently undramatic and often busily indecisive musical style have kept them from popularity. Poulenc composed one comic monodrama and one serious opera that seem likely to endure. The first, *Les Mamelles de Tirésias* (1947), is a wildly nonsensical opéra bouffe, the sardonic music of which is humorously appropriate to the text by the French poet Guillaume Apollinaire. The second, *La Voix humaine* (*The Human Voice*, 1959; text by Jean Cocteau), has as its only visible character a distraught young woman conversing by telephone with her lover. Poulenc's only large serious opera, *Dialogues des Carmélites* (*Dialogues of the Carmelites*, 1957; libretto by Georges Bernanos), appears to be the most impressive French opera since *Pelléas et Mélisande*. It makes telling operatic use of Poulenc's instantly identifiable style of melody, harmony (only very gently dissonant), and rhythm to tell a moving and tragic story of nuns martyred during the French Revolution.

Tragic and sardonic themes

LATER OPERA IN ITALY

Viewed broadly, the story of Italian opera contemporary with and following Verdi parallels that of French opera after Berlioz, in the appearance of talented composers and stageworthy operas as well as in quickly becoming a tally of pieces of considerable charm and some degree of originality responding to shifting musical techniques and manners. It includes only two or three composers of the first rank.

Verismo. Amilcare Ponchielli, shining clearly in light reflected from Verdi, composed one opera that remains in the international "standard" repertory—*La Gioconda* (1876; libretto by Arrigo Boito). The general turn toward realism began on a Roman stage in 1890 with Pietro Mascagni's dazzlingly successful one-act opera *Cavalleria rusticana* (*Rustic Chivalry;* libretto by Guido Menasci and Giovanni Targioni-Tozzetti). It soon evoked the descriptive term *verismo* (realism) and set a vogue for raw, violent, melodramatic librettos matched to music clearly descended from the middle-period operas of Verdi. Mascagni went on composing operas for 50 years after *Cavalleria*, but none won a permanent place on the stage. The best of them is *L'Amico Fritz* (*Friend Fritz*, 1891; libretto by Nicolo Despure). And two years after the premiere of *Cavalleria rusticana*, an equally successful product of *verismo* was staged in Milan: *I Pagliacci* (*Clowns*, 1892; libretto by the composer), by Ruggero Leoncavallo, who had no more staying power than Mascagni, though he produced operas for the remainder of his life.

Puccini. The Italian counterpart of Massenet in France made his first important contribution to the operatic stage the year after *I Pagliacci*. He was Giacomo Puccini, whose work is characterized by emotional directness of appeal and colourful, rich orchestration; the opera was *Manon Lescaut* (1893), based on the novel by the Abbé Prévost from which the libretto of Massenet's *Manon* had been derived. Puccini established himself unmistakably as the most important post-Verdian Italian operatic composer with *La Bohème* (1896; libretto by Giuseppe Giacosa and Luigi Illica, after Henri Murger's *Scènes de la vie de bohème*). It, too, remains "standard," as do *Tosca* (1900; libretto by Giacosa and Illica) and *Madama Butterfly* (1904; libretto by Giacosa and Illica), which again capitalized upon Puccini's attraction to, and ability to characterize in music, sorrowing, attractive young women. Returning closer to violent

Establishment of Puccini's reputation with *La Bohème*

verismo, Puccini (who always had to struggle to find librettos germane to his purposes) next composed an opera of the American "Wild West," *La fanciulla del west* (*The Girl of the Golden West*, 1910; libretto by Guelfo Civinini and Carlo Zangarini).

Fumbling for apposite literary materials, Puccini next proposed to write an operetta for Vienna; the outcome was a mixed, uncertain operetta-like piece, *La Rondine* (1917; libretto by Giuseppe Adami), produced by Monte Carlo, instead of Vienna, because of World War I. It was followed by a trio of one-act operas given its premiere at the Metropolitan, in 1918. *Il trittico* (*The Triptych*) consisted of the veristic and powerful *Il tabarro* (*The Cloak*; libretto by Adami), the sweetly sad, all-female *Suor Angelica* (*Sister Angelica*; libretto by Giovacchino Forzano) and the opera buffa of medieval Florence, *Gianni Schicchi* (libretto by Forzano). *Gianni Schicchi*, in its sarcastic humour and musical vitality, is the finest Italian operatic comedy after *Falstaff*.

Puccini died before finishing *Turandot* (libretto by Adami and Renato Simoni based on the Italian writer Carlo Gozzi's fable of the same name). It was produced posthumously in 1926. *Turandot* shows Puccini taking note of the then-recent developments in harmony, while giving them an Eastern flavour. The bloodthirsty story with a happy ending (for the two characters who least deserve it) alienates some opera goers, but many of those who accept *Turandot* as the brilliant, extremely melodious, highly pictorial representation of a legend consider it the finest of Puccini's operas. The music of *Turandot* was completed after the composer's death by the Italian composer Franco Alfano.

Puccini's contemporaries. The other Italian operatic composers of the late 19th century and early 20th displayed neither the brash originality of the young Mascagni and Leoncavallo nor the varied theatrical genius of Puccini. They included Alfredo Catalani, whose best known opera, shaped with extreme melodic refinement and mildly interesting orchestral commentary, is *La Wally* (1892; libretto by Luigi Illica). Another of these minor men was Umberto Giordano (1867–1948), whose bombastic *Andrea Chenier* (1896; libretto by Illica) and *Fedora* (1898; libretto by Arturo Colautti) are still staged. Francesco Cilea overtly copied the more sentimental aspects of Puccini in his biggest success, *Adriana Lecouvreur* (1902; libretto by Colautti). The 20th-century music critic Donald Jay Grout made an accurate description of nearly all of the post-Verdian Italian operas except *Cavalleria rusticana*, *I Pagliacci*, and the best of Puccini when he wrote of *Adriana Lecouvreur* as "expertly contrived music of a lyrical-tragic sort . . . unadventurous harmonically or rhythmically, but good theatre and grateful for the singers."

Busoni and Wolf-Ferrari. Of much greater interest (though not of equal popularity) were two contrasted half-Italian, half-German composers active early in this century: Ferruccio Busoni and Ermanno Wolf-Ferrari. Busoni, a learned musician, wrote, in a then-advanced harmonic idiom and using his own librettos, four operas that have had scattered enthusiasts but no large public: *Die Brautwahl* (*Choice of a Bride*, 1912); two commedia dell'arte parodies—*Turandot* (1917; libretto after Gozzi) and the equally short *Arlecchino* (*Harlequin*, 1917)—and his major work, *Doktor Faust*, left incomplete at his death and completed by Philipp Jarnach (1925). Busoni's operas, and *Doktor Faust* in particular, are notable for their intellectual mastery, spiritual elevation, and other operatically peripheral virtues, but their almost total lack of dramatic cogency and musical allure have probably kept them from frequent performance.

At the other end of the operatic spectrum, Wolf-Ferrari possessed a fine talent for opera buffa of an especially light, airy sort, and also composed one of the rawest later examples of *verismo*. There is enormous charm in his comedies *Le donne curiose* (*The Curious Women*, 1903; libretto by Luigi Sugana), *I quattro rusteghi* (*The Four Ruffians*, 1906; libretto by Giuseppe Pizzolato), and *Il segreto di Susanna* (*The Secret of Susanna*, 1909; libretto by Enrico Golisciani). All three were first given in German translation. The melodrama *I gioielli della Madonna* (*The Jewels of the Madonna*, 1911; by Golisciani and Zangarini) has been well described as Donizetti plus *verismo*.

The "big three" Italian composers of the 1920s and 1930s—Ottorino Respighi, Gian Francesco Malipiero, and Alfredo Casella—all composed operas and opera-like works of considerable musical interest but little theatrical vitality. What historical influence they have had has operated largely outside the opera house.

RUSSIAN OPERA

After long subjection to imported Italian, French, and German composers, opera in Russia by Russians asserted itself in two well-known works by Mikhail Ivanovich Glinka (1804–57): *Zhizn za tsarya* (*A Life for the Tsar*), also known as *Ivan Susanin* (1836; libretto by Baron Georgy Fyodorovich Rosen), and *Ruslan i Lyudmila* (*Ruslan and Lyudmila*, 1842; libretto by Valeryan Fyodorovich Shirkov and others). Basically old-fashioned Italianate operas, they—*Ruslan* in particular—determined the nature of much future Russian music because of Glinka's approximations of Slavic folk music, his pre-Wagnerian use of a tentative leitmotiv technique, and the clarity and shifting colours of his orchestration. Glinka's operas, weakened by evidence of his lifelong dilettantism, remain in the repertory only in the Soviet Union, but their stylistic importance was decisive.

Glinka as father of Russian nationalism

Almost as influential as Glinka in shaping future Russian opera was his much less successful disciple Aleksandr Sergeyevich Dargomyzhsky. His *Rusalka* (1856; his libretto, after a fairy tale by Pushkin) illustrated his strong emphasis on a declamation midway between recitative and aria, as well as his musical amateurism. Even more influential, although left incomplete at Dargomyzhsky's death, was *Kamenny gost* (*The Stone Guest*, an integral setting of Pushkin's short Don Juan play; completed by César Antonovich Cui and Nikolay Rimsky-Korsakov and staged posthumously in 1872). It is couched in what were then advanced harmonic terms, powerful in characterization, but musically all but sterile. Dargomyzhsky remains more interesting to historians than to opera goers.

The operas of Aleksandr Borodin, Rimsky-Korsakov, and Modest Mussorgsky are still performed. Borodin's incomplete *Knyaz Igor* (*Prince Igor*, his own libretto; completed and edited by Rimsky-Korsakov and Aleksandr Glazunov) was staged posthumously in 1890. It is dramatically shapeless but is splashed with Slavic and Oriental colours. Most of Rimsky-Korsakov's numerous fairy-tale operas have the nature of brilliantly illustrated books, but what may be his finest work is "the Russian *Parsifal*," *Skazaniye o nevidimom grade Kitezhe i deve Fevroni* (*The Legend of the Invisible City of Kitezh*, 1907; libretto by Vladimir Ivanovich), a work of marked emotional strength. Of his lighter works, the best known are *Snegurochka* (*The Snow Maiden*, 1882; his own libretto), *Sadko* (1898; libretto by the composer and Byelsky), and the fantastic opera buffa, *Zolotoy petushok* (*Le Coq d'or*, or *The Golden Cockerel*, 1909; libretto by Byelsky, after Pushkin). Like *Prince Igor*, Rimsky-Korsakov's operas contributed largely to what many music lovers came to consider typically "Russian" music, though the splashily coloured world they create was in reality Rimsky-Korsakov's own invention.

Mussorgsky. Mussorgsky composed all or part of several operas. Among them, *Khovanshchina* (to his own libretto; the score completed and orchestrated by Rimsky-Korsakov; posthumous premiere in 1886) bears a family resemblance to *Prince Igor*, particularly in its employment of real and simulated Orientalism, but is more serious and much more confident in tone. Mussorgsky's greatest achievement, and the most worthwhile Russian opera, is *Boris Godunov* (1874; his own libretto, based upon Pushkin's drama and a history of Russia by Nikolay Mikhailovich Karamzin). Boris, the guilty usurper of the throne, dominates this glittering but dour pageant in which the Russian people are present in remarkably forceful choral writing. Mussorgsky's ability to transmit

Greatness of *Boris Godunov*

textual points in very condensed music has possibly never been matched. Except at a few weak moments, he made a virtue out of amateurishness and naïveté, fearlessly extracting intense power and theatrical effectiveness from his newly developed techniques. The influence of *Boris Godunov* has been strong on numerous composers of opera both in Russia and elsewhere.

Tchaikovsky. The operatic practice of Peter Ilich Tchaikovsky was very different. For dramatic vigour he substituted clear characterization expressed lyrically, creating, among others, two highly idiosyncratic operas: *Eugene Onegin* (1879; libretto by the composer and Konstantin S. Shilovsky, after Pushkin) and the stronger melodrama *Pikovaya dama* (*The Queen of Spades*, 1890; libretto by Modest Ilich Tchaikovsky, after Pushkin). The personal emotion and the characterization of hero and heroine in *Eugene Onegin* are vivid. In all of Tchaikovsky's operas the highly subjective emotion that long made him the most often performed of orchestral composers is tellingly present. Many consider his other operas, containing much fine music, all unjustly neglected.

Stravinsky. Igor Stravinsky turned to opera three times during his long composing career, to near-opera more often. First came *Solovey* (*The Nightingale*, 1914; libretto by the composer and Stepan Nikolayevich Mitusov, after Hans Christian Andersen), which clearly reveals the influence of Rimsky-Korsakov, who had been Stravinsky's teacher. Next among his true operas came *Mavra* (1922; libretto by Boris Kochno, derived from Pushkin), an opera buffa in the unmistakable musical style that made Stravinsky the foremost composer of his era. Then a long period, marked by several near-operas (among them the urgent opera-oratorio *Oedipus Rex*, 1927) elapsed before the appearance of Stravinsky's full-length opera in English, *The Rake's Progress* (libretto by the poets W.H. Auden and Chester Kallman, after Hogarth's engravings, 1951), a neoclassical, austere, but compassionate work.

Prokofiev. Sergey Prokofiev composed numerous operas both in his "modern" musical manner and in a harmonically less advanced "socialist realist" style after his return to the Soviet Union in 1934. Among the former, the best and most often staged are the opera buffa, *L'Amour des trois oranges* (*The Love for Three Oranges*, Chicago, 1921; his own libretto), in a musical style that might be called Rimsky-Korsakov updated, and the lurid opera of hallucination, *Angel of Fire* or *The Fiery Angel* (radio premiere 1954; his own libretto after a story by Valery Yakovlevich Bryusov). Of Prokofiev's Soviet operas, the most winning is the gay *Betrothal in a Monastery*, also known as *The Duenna* (1946; libretto by Mira Mendelson based on a play of that name by the 18th-century Irish-born dramatist Richard Brinsley Sheridan). The most ambitious is the massive *War and Peace* (1946; revised, condensed version, 1955; libretto by the composer and Mira Mendelson).

Prokofiev's Soviet operas

The best known Soviet opera outside its homeland, however, is a grim tale of sexual repression and violence by Dmitry Shostakovich originally called *Ledi Makbet Mtsenskogo Uyezda* (*Lady Macbeth of the Mzensk District*, 1934; libretto by the composer and Y. Priess, after a story by Leskov), later revised, after a long period of eclipse caused by government disapproval, as *Katerina Ismaylova* (1963). Numerous other Soviet operas have not been staged outside the Soviet Union or have proved substantially unexportable when so staged.

LATER OPERA IN GERMANY AND AUSTRIA

Strauss. Richard Strauss was greeted as the obvious "heir to Wagner" (and Liszt). His worldwide reputation was being established by his orchestral music and lieder (songs) before he turned to opera for the first time. But his pre-eminence among non-Italian composers of opera was established by two "shocking" one-act operas: *Salome* (1905; libretto taken from Oscar Wilde's drama, translated into German by Hedwig Lachmann) and *Elektra* (1909). With the latter work Strauss began a long and fruitful association with the poet and dramatist Hugo von Hofmannsthal as his librettist. Couched in a violent harmonic idiom, requiring huge orchestral forces and leading singers of great vocal power and stamina, *Salome* and *Elektra* seemed to many critics to be Straussian tone poems with added voices, but they soon became, and have remained, part of the standard repertory. They were followed by an altogether different sort of opera, *Der Rosenkavalier* (1911), again with a libretto by Hofmannsthal, a bittersweet comedy notable for the superb musical creation of the central character (the Marschallin) and for three-quarter rhythms that placed this Strauss alongside Johann the Younger as a composer of Viennese waltzes. It marks Strauss's invention of a subtle *parlando* (conversational) style all his own, which he also used to great effect in his later opera.

Strauss's collaboration with Hofmannsthal

Strauss composed ten operas after *Der Rosenkavalier*. All but one or two of them won wide popularity; none of them displays any constructive or musical characteristics not present in his earlier works. The most successful have been the "chamber opera" *Ariadne auf Naxos* (*Ariadne on Naxos*, 1912; revised 1916); the giant allegory *Die Frau ohne Schatten* (*The Woman Without a Shadow*, 1919), which some writers have called Strauss's masterpiece, whereas others denounce it as confused, even megalomaniac; and *Arabella* (1933), which closely resembles *Der Rosenkavalier* in many details. *Capriccio* (1942), his last opera, is an absorbing characterization of the old argument as to whether words or music should take precedence in opera.

Pfitzner and Hindemith. Several harmonically conservative German composers active during and just after Strauss's long career were much less gifted and also less successful. Probably the most notable of them were Hans Pfitzner and Paul Hindemith. The antimodern Pfitzner, partly a belated Wagnerian, composed several operas of melodically long-lined, subjective, at times mystical content which have not reached beyond the German-speaking countries. The best-known is *Palestrina* (1917; his own libretto), dealing with the great Italian composer, in which the austerities of 16th-century counterpoint are oddly mixed with Pfitzner's often abrupt dissonances.

Hindemith damagingly lacked theatrical insight, but was admired for his technical wizardry and lofty aims. He composed some satirical comedies, but came to be heard in opera houses only with the serious *Cardillac* (1926, revised 1952; libretto by Ferdinand Lion, after a tale by E.T.A. Hoffmann) and *Mathis der Maler* (*Mathias the Painter*, 1938; his own libretto on the life of Matthias Grünewald).

Schoenberg. If "modernism" seems to be struggling toward birth in the operas of Strauss, Pfitzner, Hindemith, and some others, it sprang fully armed from the music of three Viennese composers: Arnold Schoenberg, Alban Berg, and Anton von Webern, propagators and chief practitioners of what came to be called atonality and 12-tone composition. Webern composed no operas. Schoenberg's first theatrical works—the one-act *Erwartung* (*Expectation*, composed in 1909, performed in 1924; single-character libretto by Marie Pappenheim) and the one-act *"Drama mit Musik" Die glückliche Hand* (*The Hand of Fate*, 1924; his own libretto)—are extremely discordant, thickly and earnestly romantic, even expressionistic, and occasionally use *Sprechstimme*, a variety of vocal emission between speech and song that Schoenberg himself described as "the voice rising and falling relative to the indicated intervals, and everything being bound together with the time and rhythm of the music except where a pause is indicated." Their harmony is unremittingly chromatic.

Schoenberg's first theatrical works

Schoenberg's only comedy, the one-act *Von Heute auf Morgen* (*From Today to Tomorrow*, 1930; libretto by Max Blonda), is in strictly construed 12-tone texture; following the theories attending that technique, it therefore returned to separate-number construction. Schoenberg's largest opera, left incomplete at his death, was the powerful, oratorio-like *Moses und Aron* (1957; his own libretto).

Berg and others. The two operas of Alban Berg—*Wozzeck* (1925; libretto by the composer, after Georg Büchner's play *Woyzeck*) and the unfinished *Lulu* (1937;

his own libretto)—are among the most powerful, effective music dramas of the 20th century. Well described as "expressionistic, morbid, neurotic, hysterical" as to story, *Wozzeck* seamlessly joins an intensely learned and appropriate score to a melodrama of a poor soldier's helplessness at the hands of his fate. *Wozzeck* is such an intense work that audiences who might be expected to be alienated by its extreme dissonance and "tunelessness" have accepted it as the great opera it is. The unfinished *Lulu*, a part-tragic, part-comic drama of the hectic life and final murder of a nymphomaniac, adds film clips and spoken dialogue to the means employed in *Wozzeck*. Although Berg's score was elaborated from a single "tone-row" (arrangement of the 12 tones of a chromatic scale) and though the resultant tonal clashing is not bilked, many passages in *Lulu* appear to show Berg tending toward less dissonance.

Intensity of *Wozzeck*

The most appreciable Germanic musicodramatic composer outside the Viennese orbit has appeared to be Carl Orff (1895–), who has juggled many varieties of almost-operatic forms. Best known for his absorption of bawdy medieval student songs into a "scenic oratorio," *Carmina Burana* (1937), Orff worked that singular exploitation of repetitive rhythms, bare harmonies, *ostinati*, and crying vocal colours into a trilogy called *Trionfi* by adding to it *Catulli Carmina* (1943) and *Il Trionfo d'Afrodite* (1953). More conventionally operatic is his one-act comedy *Die Kluge* (*The Clever Girl*, 1943; his own libretto, after a tale by the Brothers Grimm). Probably his major effort in musical drama is another trilogy, consisting of *Antigonae* (1949), *Oedipus der Tyrann* (1959), and *Prometheus* (1968). In it a texture of spoken and declaimed texts is spaced out with incidental music of almost brutal force. Bernd Alois Zimmermann's opera *Die Soldaten* (*The Soldiers*, 1969) is the most successful of the multi-media operas.

Hans Werner Henze (1926–) is the most recent major operatic composer to come out of Germany. The works of his most likely to be encountered in the opera house are *Elegy for Young Lovers*, his first collaboration with the poets W.H. Auden and Chester Kallman, and *Der junge Lord* (*The Young Lord*), which satirizes German provincial life. In the 1970s he turned from opera to forms of musicodramatic experiment.

NATIONALIST OPERA

Czechoslovakia. The specifically Russian operas of Glinka, Dargomyzhsky, and the Five, as well as the non-operatic music they influenced, have parallels in other countries. In what is now Czechoslovakia, the national school effectively began with Bedřich Smetana, known, operatically speaking, outside his homeland almost exclusively for the vigorous, highly coloured folk comedy *Prodaná nevěsta* (*The Bartered Bride*, 1866; libretto by Karel Sabina), which determined many aspects of future Czech musical usage as clearly as Borodin and Mussorgsky had set Russian styles. Several of Smetana's other operas, both comic and tragic, remain on Czech stages—most notably the overtly patriotic *Dalibor* (1868) and *Libuše* (1881), both to librettos in Czech translation of originally German texts by Joseph Wenzig, and his comedy *Hubička* (*The Kiss*, 1876). The other leading Czech composer of Smetana's period, Antonín Dvořák, wrote nine operas but remained preponderantly an instrumental composer, never matching the older composer's stage success. Of Dvořák's mature operas, the best known outside Czechoslovakia is the melancholy fairy tale *Rusalka* (libretto by Jaroslav Kvapil), made attractive by his considerable melodic and harmonic gifts.

Production of *The Bartered Bride*

Leoš Janáček was specifically Moravian in musical background, far more harmonically advanced than Smetana and Dvořák, though less adroit technically. He was clearly a 20th-century composer, whose highly individual music, typified by a short-phrased melodic idiom used to catch the speech-rhythm of his native language, was rediscovered after World War II. A follower of Mussorgsky, whose much-discussed "dilettantism" his closely resembles, Janáček is now represented intermittently in non-Czech opera houses chiefly by *Její pastorkyňa* (*Her Foster Daughter*, 1904; changed to *Jenufa* for Janáček's 1916 revision; his libretto derived from a story by Gabriela Preissová), *Káťa Kabanová* (1921; libretto by Vincenc Červinky), *Příhody lišky bystroušky* (*The Cunning Little Vixen*, 1924; libretto by the composer), and *Věc Makropulos* (*The Makropoulos Case*, 1926; libretto by the composer), each of which has a character and milieu of its own while preserving the peculiarly individual setting of the Czech language.

Hungary and Poland. The most important Hungarian operas of the early 20th century, neither of them representing its composer at his finest, are the one-act *Duke Bluebeard's Castle* (1918; libretto by Béla Balázs), by Béla Bartók, and the ballad opera *Háry János*, by Zoltán Kodály (1926; libretto by Béla Paulini and Zsolt Harsányi), both of which have become more familiar in concert performance or excerpts than in staged productions. The most influential and popular of Polish nationalist operas, *Halka* (1854; libretto by Włodzimierz Wolski), was composed by Stanisław Moniuszko; he also wrote an admirable comedy, *The Haunted Manor* (1865; libretto by Jan Chichiński). Of 20th-century Polish operas, perhaps the most substantial is the much less popular *Król Roger* (*King Roger*, 1926; libretto by Jarosław Iwakiewicz) by Karol Szymanowski.

Spain. Spanish operatic (and generally musical) nationalism began with Felipe Pedrell, more influential as teacher-propagandist than as producing composer. Of his ten neglected operas, the most imposing were to have been contained in a trilogy, based on a Catalan libretto by Victor Balaguer, but only the first two sections, *Los Pirineos* (*The Pyrenees*) and *La Celestina*, were completed and only the first was staged (1902). Of the more familiar Spanish composers, both Isaac Albéniz and Enrique Granados composed operas of strongly Spanish colour that have lapsed from the repertory—Albéniz particularly in the comic one-act *Pepita Jiménez* (1896), Granados in the semi-veristic *Goyescas* (1916; libretto by Fernando Periquet y Zuaznabar), the score of which clearly reveals its origin in a suite of piano pieces. The best results of Spanish operatic nationalism (possibly because more than a little tinged with internationalism) are two very different operas by Manuel de Falla: the specifically Andalusian *La vida breve* (*Brief Life;* libretto by Carlos Fernández Shaw; first staged in French translation, 1913) and the one-act *El retablo de Maese Pedro* (*Master Peter's Puppet Show*, 1923; text by the composer, after a scene in *Don Quixote*), in effect a chamber opera for marionettes.

Other countries. Consciously and unintentionally nationalistic operas have been composed and staged in nearly all European countries, Latin America, Great Britain, the United States, and other parts of the world, but none of them has entered the international repertory, held the stage in its homeland, or affected importantly the morphology of opera. Sometimes viewed as an exception is George Gershwin's *Porgy and Bess* (1935; libretto by Dubose Heyward and Ira Gershwin), an uncertain crossbreed of folk opera and American musical comedy, which has had no recognizable descendant of high quality.

Chinese opera, mimed and choreographed in ways foreign to Western opera, has a repertory of traditional works. It seems unlikely in the extreme that this highly rigid and conventionalized art will ever be exported with success.

RECENT DEVELOPMENTS

England. English opera, which had languished for centuries, was revitalized by the theatrical talent of the eclectic Benjamin Britten, whose stage works show a remarkable sympathy for the human predicament expressed in readily accessible, deeply felt music. His masterpiece is the gloomy, forceful *Peter Grimes* (1945; libretto by Montague Slater). Among Britten's other operas to win widespread stagings are the chamber operas *The Rape of Lucretia* (1946; libretto by Ronald Duncan) and *Albert Herring* (1947; libretto by Eric Crozier), the all-male *Billy Budd* (1951; libretto by E.M. Forster and Crozier;

Benjamin Britten's works

based upon Herman Melville's unfinished story), the eerily effective *The Turn of the Screw* (1954; libretto by Myfanwy Piper, after the Henry James story), *A Midsummer Night's Dream* (1960; libretto by Britten and Peter Pears, from Shakespeare), three church operas to librettos by William Plomer (1964–68), and *Owen Wingrave* (1971; libretto by Myfanwy Piper). So far, less popular outside England are the idiosyncratic operas, to his own complex librettos, of Sir Michael Tippett: *The Midsummer Marriage* (1955), *King Priam* (1962), and *The Knot Garden* (1970).

United States. American contributions to international opera, after a 19th-century desert of imitation German and French operas, became much more numerous after World War II. It is possible here to mention only one composer and a few isolated operas that have evoked enduring resonance. The most often performed of contemporary operatic composers has been the Italo-American Gian Carlo Menotti. Using his own librettos, he has produced, in a variety of structural styles, a series of Puccini-derived melodramas and sentimental tragedies of considerable popular appeal, among them *The Medium* (1946), *The Consul* (1950), *Amahl and the Night Visitors* (composed for television performance, 1951), and *The Saint of Bleecker Street* (1954). He also wrote the libretto for the first, mildly successful, opera of Samuel Barber, *Vanessa* (1958; awarded 1958 Pulitzer Prize). Barber's second large opera, *Antony and Cleopatra* (1966; libretto derived from Shakespeare by Franco Zeffirelli), commissioned to inaugurate the second Metropolitan Opera House in New York, was a failure and vanished quickly from performance.

Menotti's works

A unique niche is occupied by the two operas that Virgil Thomson composed to texts by Gertrude Stein arranged by Maurice Grosser: the Spanish-tinted *Four Saints in Three Acts* (1934) and *The Mother of Us All* (1947), a delicious flow of invention around the figure of Susan B. Anthony. Their fragile but real durability has resulted from Thomson's singable, apt folk-based setting of texts that alternate among the apparently nonsensical, the satiric, and the emotionally moving. Within the United States—not to count the "workshop operas" and simplified semi-folk near-operas that many American composers recently have favoured—two of the most frequently performed recent American operas are the folklike "Western" *Ballad of Baby Doe* (libretto by John Latouche, 1956), by Douglas Moore (1893–1969) and the melodramatic, "Southern" *Susannah* (libretto by the composer, 1955) by Carlisle Floyd (1926–).

PROSPECTS

The existing audience for the standard repertory operas throughout the world is enormous. If to it be added the large number of people attending workshop, college and university, music school, amateur, and semi-amateur performances—both of the same repertory and of operas created for those special purposes—the statistics suggest that all is well with opera. Yet by the mid-20th century the art of opera had become largely a museum art. Constant repetitions of a relatively small group of operas of the past far exceed creation of new works that can be categorized with the finest of those mentioned in this article. What, then, are the future prospects for this art, now nearing four centuries of fertile existence?

The prospects for the creation of major new operas are poor. How could they be better when the two major components of opera—dramatic literature and music—are both in a period of fundamental crisis? Outside the opera house, audiences for the spoken theatre and for concert and recital have been oscillating widely while economic problems have multiplied swiftly. Some writers have seen the future of opera in new varieties of semi-operatic "happenings" and other mixed forms, others in the numerous pieces written and composed specifically for workshop and school staging. For the large paying audience, however, opera continues to mean more or less traditional performances of the operas of Mozart, Verdi, Wagner, Puccini, Richard Strauss, and a few others.

Revivals of long-neglected operas have meanwhile attracted audiences in many opera houses (notably, to speak only of the United States, with such companies as the New York City Opera and those in Chicago, Dallas, San Francisco, and Santa Fe). That trend, which freshens the repertory in the absence of new operas, has brought back most noticeably some of the bel canto masterpieces of Rossini, Donizetti, and Bellini. These revivals, in turn, have affected the otherwise tradition-bound activities of record-publishing companies, whose reluctance to issue operas in any idiom more recent than that of Britten has reflected general public taste. Ballet, which entered a period of notable flowering between the two world wars, has fallen heir to what might, under different conditions, have become a large audience for operas employing "advanced" harmonic and melodic idioms.

The bel canto revival

Until the harmonic and melodic idioms of contemporary composers and the taste of the large paying audience converge, until the crises in style and subject matter no longer prevent librettists from writing meaningfully for that same audience, the present situation in opera houses can be expected to continue with only an occasional foray into less immediately hospitable territory. Predictions about future developments in any art are especially hazardous, and the fact that opera has survived more than three and a half centuries of social, political, literary, musical, and other changes without losing either its audience or its particular glamour strongly suggests that it will evolve ways out of its current predicaments. What remains true is that if this greatest of musicodramatic forms is to approach its absorbing best, it must, in one of numerous possible ways, be what Livia Miragoli (1924) defined it as being:

> a work of unique art, then, in which the value of the literary element is higher or lower as it is better or less well adapted to being combined with music, the musical element as it is a better or a less good response to the contents of the text.

BIBLIOGRAPHY. Books on all aspects of opera are overwhelmingly numerous, particularly in Italian, German, French, and English. The following suggested list is confined to books written in English, a few indicative titles excepted. Unique in its coverage of opera (as well as of the spoken theatre, ballet, the cinema, and the circus) is the 12-volume Italian *Enciclopedia dello Spettacolo* (1954–68). Among books on opera in general, useful basic information and informed opinion may be found in: WALLACE BROCKWAY and HERBERT WEINSTOCK, *The World of Opera* (1962); EDWARD J. DENT, *Opera* (1940); DONALD JAY GROUT, *A Short History of Opera*, 2nd ed. (1965); JOSEPH KERMAN, *Opera As Drama* (1956); GUSTAV KOBBE, *Complete Opera Book*, rev. and enlarged by the EARL OF HAREWOOD (1954), particularly useful for its libretto stories; ALFRED LOEWENBERG (comp.), *Annals of Opera, 1597–1940*, 2nd ed., 2 vol. (1955), the basic source for data on premieres and the dates and places of important later productions; and ERNEST NEWMAN (ed.), *Stories of the Great Operas* (1927, reprinted 1948), *More Stories of Famous Operas* (1943), and *Seventeen Famous Operas* (1954), which extensively analyze both libretto and music.

Among books treating opera in individual cities and opera houses, reliable data may be found in: (German) ANTON BAUER, *Opern und Operetten in Wien* (1955); ARTHUR J. BLOOMFIELD, *The San Francisco Opera, 1923–1961* (1961); JOHN FREDERICK CONE, *Oscar Hammerstein's Manhattan Opera Company* (1966); RONALD L. DAVIS, *Opera in Chicago* (1966); QUAINTANCE EATON, *The Boston Opera Company* (1965); (Italian) CARLO GATTI, *Il Teatro alla Scala (1778–1963)*, 2 vol., the second, compiled by GIAMPIERO TINTORI, being a detailed chronology of operatic and other performances (1964); IRVING KOLODIN, *The Metropolitan Opera, 1883–1966* (1966); MARCEL PRAWY, *The Vienna Opera* (1970); HAROLD D. ROSENTHAL, *Two Centuries of Opera at Covent Garden* (1958); WILLIAM H. SELTSAM (comp.), *Metropolitan Opera Annals: A Chronicle of Artists and Performances* (1947; also two supplements, 1957, 1968); and (French) STEPHANIE WOLFF, *Un Demi-siècle d'opéra-comique, 1900–1950* (1953) and *L'Opéra au Palais Garnier, 1875–1962* (1962).

A 122-page bibliography, including periodical and monographic articles of importance, may be found in vol. 2 of the first, 2-volume, edition of DONALD JAY GROUT, *A Short History of Opera* (1947). A prime source of operatic events since 1950 is the volumes (with annual index) of the London periodical *Opera*.

(H.We.)

Operations Research

Operations research is the application of scientific method to the management of organized systems. It is called operational research in the United Kingdom, where it originated. Operations research attempts to provide those who manage organized systems with an objective and quantitative basis for decision; it is normally carried out by teams of scientists and engineers drawn from a variety of disciplines. Thus, operations research is not a science itself, but is the application of science to the solution of managerial and administrative problems, and it focusses on the performance of organized systems taken as a whole rather than on their parts taken separately. Usually concerned with systems in which human behaviour plays an important part, operations research differs in this respect from systems engineering, which, using a similar approach, tends to concentrate on systems in which human behaviour is not important. Operations research was originally concerned with improving the operations of existing systems rather than developing new ones; the converse was true of systems engineering. This difference, however, has been disappearing as both fields have matured.

The subject matter of operations research consists of decisions that control the operations of systems. Hence, it is concerned with how managerial decisions are and should be made, how to acquire and process data and information required to make decisions effectively, how to monitor decisions once they are implemented, and how to organize the decision-making and decision-implementation process. Extensive use is made of older disciplines such as logic, mathematics, and statistics, as well as recent scientific developments such as communications theory, decision theory, cybernetics, organization theory, the behavioral sciences, and general systems theory.

The Industrial Revolution

In the 19th century, what is coming to be called the First Industrial Revolution involved mechanization or replacement of man by machine as a source of physical work. Study and improvement of such work is the objective of industrial engineering. The contemporary Second Industrial Revolution is concerned with automation or mechanization of mental work. The primary technologies involved are: mechanization of symbol generation (observation by machines such as radar and sonar); mechanization of symbol transmission (communication by telephone, radio, and television); and mechanization of logical manipulation of symbols (data processing and decision making by computer). Operations research applies the scientific method to the study of mental work and provides the knowledge and understanding requisite to make effective use of men and machines to carry it out.

History. In a sense, every effort to apply science to management of organized systems, and to their understanding, was a predecessor of operations research. It began as a separate discipline, however, in 1937 in Britain as a result of the initiative of A.P. Rowe, superintendent of the Bawdsey Research Station, who led British scientists to teach military leaders how to use the then newly developed radar to locate enemy aircraft. By 1939 the Royal Air Force formally commenced efforts to extend the range of radar equipment so as to increase the time between the first warning provided by radar and the attack by enemy aircraft. When the scientists involved recognized that additional gains could be made if the time between the first warning and the deployment of defenses could be reduced, they began to study the communication system that connected detection centres to defenses. At first they analyzed physical equipment and communication networks, but later they examined behaviour of the operating personnel and relevant executives. As the number of early-warning stations was increased, it was observed that there was a substantial variation in performance between them, even when operated by the same group of test operators. Analysis revealed ways of improving the operators' techniques and also revealed unappreciated limitations in the network.

The scientists working on different aspects of this problem were brought together in September 1939 at Fighter Command headquarters. The section steadily extended its scope of activities beyond the use of radar and, by the time of the Battle of Britain, was consulted on an ever-widening variety of problems.

By the summer of 1941 it was decided to establish operations research sections very widely in the Royal Air Force. Similar developments took place in the Army and the Royal Navy, and in both cases radar again was the instigator. In the Army, use of operations research had grown out of the initial inability to use radar effectively in controlling the fire of antiaircraft weapons. In the course of studying this problem it was found that radar equipment that worked perfectly in testing laboratories often failed to operate in the field. Thus, since the traditional way of testing equipment did not seem to apply to radar gunsights, scientists found it necessary to test in the field under operating conditions, and the distinguished British physicist and Nobel Laureate P.M.S. Blackett organized a team to solve the antiaircraft problem. Blackett's Anti-aircraft Command Research Group included two physiologists, two mathematical physicists, an astrophysicist, an Army officer, a former surveyor, and subsequently, a third physiologist, a general physicist, and two mathematicians. In March of 1941 Blackett and some of the members of his group moved to the Coastal Command, where they became involved in radar detection of ships and submarines, and in May of that year various members of Blackett's original group formed the Operational Research Group of the Air Defense Research and Development Establishment, which later became the Army Operational Research Group. Thus, within two years after the beginning of the war, formal operations research groups had been established in all three of Britain's military services. In addition, operations research was later employed in civil defense activities; the group of 40 researchers assembled for this purpose included several Americans who subsequently continued their activities with the U.S. Air Force. This group's work was supplemented by studies on the effects of bomb explosions upon human beings.

Radar testing problem

Development of operations research paralleling that in Britain took place in Australia, Canada, among the Free French Forces, and, most significantly for future developments, in the U.S., which was the beneficiary of a number of contacts with British researchers. Sir Robert Watson-Watt, who with A.P. Rowe launched the first two operational studies of radar in 1937 and who claims to have given the discipline its name, visited the U.S. in 1942 and urged that operations research be introduced into the War and Navy departments. Reports of the British work had already been sent from London by American observers, and James B. Conant, then chairman of the National Defense Research Committee, had become aware of operations research during a visit to England in the latter half of 1940. Another stimulant was Blackett's memorandum, "Scientists at the Operational Level," of December 1941, which was widely circulated in the U.S. Service departments.

The first organized operations research activity in the United States began in 1942 in the Naval Ordnance Laboratory. This group, which dealt with mine warfare problems, was later transferred to the Navy Department from which it designed the aircraft mining blockade of the Inland Sea of Japan. In May of 1942 the Antisubmarine Warfare Operations Research Group, which reported to both the Army and Navy, was organized. This group was later expanded into the Operations Research Group on the staff of the Commander in Chief, U.S. Fleet. It dealt with submarine and antisubmarine warfare, aircraft and amphibious operations, and antiaircraft and new weapons analysis.

As in Britain, radar stimulated developments in the U.S. Air Force. In October 1942, all Air Force commands were urged to include operations research groups in their staffs. By the end of the war there were 26 such groups in the Air Force. In 1943 General George Marshall suggested to all theatre commanders that they form teams to study amphibious and ground operations.

At the end of the war a number of British operational research workers moved to government and industry.

Nationalization of several British industries was an important factor. One of the first industrial groups was established at the National Coal Board. Electricity and transport, both nationalized industries, began to use operations research shortly thereafter. Parts of the private sector began to follow suit, particularly in those industries with cooperative research associations; for example, in the British Iron and Steel Research Association.

The early development of industrial operations research was cautious, and for some years most industrial groups were quite small. In the late 1950s, largely stimulated by developments in the United States, the development of industrial operations research in Britain was greatly accelerated.

Industrial involvement in U.S.

Although in the United States military research increased at the end of the war, and groups were expanded, it was not until the early 1950s that American industry began to take operations research seriously. The advent of the computer brought an awareness of a host of broad system problems and the potentiality for solving them, and within the decade about half the large corporations in the U.S. began to use operations research. Elsewhere the technique also spread through industry.

Societies were organized, beginning with the Operational Research Club of Britain, formed in 1948, which in 1954 became the Operational Research Society. The Operations Research Society in America, formed in 1952, now has about 8,000 members. Many other national societies appeared; the first international conference on operations research was held at Oxford University in 1957. In 1959 an International Federation of Operational Research Societies was formed. By 1970 23 national societies had joined.

The first appearance of operations research as an academic discipline came in 1948 when a course in nonmilitary techniques was introduced at Massachusetts Institute of Technology. In 1952 a curriculum leading to a master's and doctoral degree was established at Case Institute of Technology. Since then about 30 major academic institutions in the United States have introduced programs. In the United Kingdom courses were initiated at the University of Birmingham in the early 1950s. The first chair in operations research was created at the newly formed University of Lancaster in 1964. In the years since, about a dozen chairs and about half as many departments have been created in Britain. Similar developments have taken place in most countries in which a national operations research society exists.

The first scholarly journal, the *Operational Research Quarterly*, published in the United Kingdom, was initiated in 1950. It was followed by the *Journal of the Operations Research Society of America* in 1952, which was renamed *Operations Research* in 1956. The International Federation of Operational Research Societies initiated the *International Abstracts in Operations Research* in 1961. In 1970 this journal contained abstracts from about 100 journals published in 21 countries.

Despite its rapid growth, operations research is still a very young scientific activity. Its techniques and methods, and the areas to which they are applied, can be expected to continue to expand rapidly. Most of its history lies in the future.

ESSENTIAL CHARACTERISTICS

The three essential characteristics of operations research are system orientation; use of interdisciplinary teams; and adaptation of scientific method to the conditions under which the research is conducted.

System orientation. The systems approach to problems recognizes that the behaviour of any part of a system has some effect on the behaviour of the system as a whole; and when individual components are performing well, the system as a whole is not necessarily performing equally well. An effort, for example, to assemble the best of each type of automobile part, regardless of make, will not necessarily result in a good automobile or even one that will run, because the parts may not fit together. It is the interaction between parts, and not the actions of any single part, that determines how well a system performs.

Thus operations research attempts to evaluate the effect of changes in any part of a system on the performance of the system as a whole and to search for causes of a problem that arises in one part of a system in other parts or in the interrelationships between parts; that is, in the structure of the system. In industry, a production problem may best be approached by a change in marketing policy; for example, the factory may fabricate a few profitable products in large quantities and many less profitable items in small quantities; long efficient production runs of high-volume, high-profit items may have to be interrupted for short runs of low-volume, low-profit items. An operations researcher might propose reducing the sales of the less profitable items and increasing those of the profitable items by placing salesmen on an incentive system that especially compensates them for selling particular items.

The interdisciplinary team. Scientific and technological disciplines have proliferated rapidly in the last 100 years. The proliferation, resulting from the enormous increase in scientific knowledge, has provided science with a filing system that permits a systematic classification of knowledge. This classification system is helpful in solving many problems by identifying the proper discipline to appeal to for a solution. But difficulties arise when more complex problems, such as those arising in large organized systems, are encountered. It is then necessary to find a means of bringing a diversity of disciplinary points of view together. Furthermore, since methods differ among disciplines, the use of interdisciplinary teams makes available a much larger arsenal of research techniques and tools than would otherwise be available. Hence, operations research is characterized by rather unusual combinations of disciplines on research teams (Blackett's group, for example), and by the use of research procedures, developed in one context, in other, very different contexts; for example, the use in economic forecasting of factor analysis, a technique that was originally developed for use in psychology.

Methodology. Until this century laboratory experiments were the principal and almost the only method of conducting scientific research. But large systems such as are studied in operations research cannot be brought into laboratories. Furthermore, even if systems could be brought into the laboratory, what would be learned would not necessarily apply to their behaviour in their natural environment, as shown by the World War II experience with radar. Experiments on systems and subsystems conducted in their natural environment ("operational experiments") are possible as a result of the experimental methods developed by the British statistician R.A. Fisher in 1923–24. For practical or even ethical reasons, however, it is seldom possible to experiment on large organized systems as a whole, even in their natural environments. This results in an apparent dilemma: to gain understanding of complex systems experimentation seems to be necessary but it cannot usually be carried out. This difficulty is solved by the use of models, representations of the system under study. Provided the model is good, experiments (called "simulations") can be conducted on it, or other methods (discussed below) can be used to obtain useful results.

Parts of a system model

A system model generally has two parts, of which the first is an equation that relates an appropriate measure of system performance (P) to variables that can be controlled (C), and to variables that are uncontrolled (U). In industry, for example, corporated management controls product prices, factory location, advertising expenditures, and other factors. Uncontrolled variables may include competitor, consumer, or supplier behaviour; the weather; and the economy. Therefore, the general structure of this equation is expressed by $P = f(C,U)$; meaning, system performance is a function of controlled and uncontrolled variables.

The second part of a system model consists of an explicit statement of the constraints within which the controlled variables can be manipulated. For example, if the total amount of time available on a machine is T and this is to be used to make three products, and the time allocated to

each is t_1, t_2, and t_3, two useful formulas can be stated: $0 \leq t_1 + t_2 + t_3 \leq T$. That is, the total time allocated (T) must be more than or equal to the sum of the separate product times (t_1, t_2, and t_3), which in turn must be more than or equal to zero. The second formula states that $t_1 \geq 0$, $t_2 \geq 0$, and $t_3 \geq 0$; that is, the amounts of time for each separate product time must be nonnegative.

Thus, a model consists of a performance function and a set of constraints. The problem to be solved can be stated in terms of the model; it is to find the values of the controlled variables (C) that, subject to the constraints and the conditions specified by the values of the uncontrolled variables (U), yield the best performance (P) of the system. These may be the values of C that either maximize or minimize P, as appropriate. If, for example, P is a measure of profit or return, the aim is to maximize it; if it is a measure of loss, cost, or risk, the aim is to minimize it. The function of P that is sought is called the "objective function." The best solution may be found exactly or approximately either by simulation or mathematical analysis. An explicit analytical procedure for finding the solution is called an "algorithm."

Even if a model cannot be solved, and many are too complex for solution, it can be used to compare alternative solutions. It is sometimes possible to conduct a sequence of comparisons, each suggested by the previous one, and each likely to contain a better alternative than was contained in any previous comparison. Such a solution-seeking procedure is called "heuristic."

A model is a simplified representation of the real world and, as such, includes only those variables relevant to the problem at hand. A model of freely-falling bodies, for example, does not refer to the colour, texture, or shape of the body involved. Furthermore, a model may not include all relevant variables because a small percentage of these may account for most of the phenomenon to be explained. Many of the simplifications used produce some error in predictions derived from the model, but these can often be kept small compared to the magnitude of the improvement in operations that can be extracted from them.

Operations-research methodology based on the use of models thus may be divided into five interdependent steps:

1. Formulating the problem: observing and analyzing the system operations to identify the relevant controlled variables, uncontrolled variables, and constraints; and formulating an appropriate measure of performance.
2. Constructing the model: assembling the variables, constraints, and measure of performance into a model. This consists primarily of finding the appropriate function (f) by which to relate the variables to the measure of performance.
3. Deriving a solution: solving the model exactly or approximately.
4. Testing the model and solution: taking steps to assure the adequacy of the representation and comparing the solution derived from it with performance that would otherwise be obtained to assure an improvement if the solution is implemented.
5. Implementing and controlling the solution: specifying who is to do what, when, and how to carry out the solution; checking results against predictions; and, if expectations are not realized, finding and correcting the cause of the deficiency.

PHASES OF OPERATIONS RESEARCH

Problem formulation. To formulate an operations-research problem, a suitable measure of performance must be devised, various possible courses of action defined (that is, controlled variables and the constraints upon them) and relevant uncontrolled variables identified. To devise a measure of performance, objectives are identified and defined, and then quantified. Objectives may be retentive, involving preservation of available resources or existing states, or acquisitive, involving attainment of additional resources or desired states. For example, a businessman may have the acquisitive objective of introducing a new product and making it profitable after one year. The identified objective is profit in one year, which is defined as receipts less costs, and would probably be quantified in terms of sales. In the real world, conditions may change with time. Thus, though a given objective is identified at the beginning of the time period, change and reformulation are frequently necessary.

Determining who makes decisions

To determine who actually makes relevant decisions and to identify the controlled and uncontrolled variables, detailed knowledge of how the system actually operates and of its environment is essential. Such knowledge is normally acquired through an analysis of the system, a four-step process that involves determining whose needs or desires the organization tries to satisfy; how these are communicated to the organization; how information on needs and desires penetrates the organization; and what action is taken, how it is controlled, and what the time and resource requirements of these actions are. This information can usually be represented graphically in a flow chart, which enables researchers to identify the controlled and uncontrolled variables that affect system performance.

Once the objectives, decision makers, their courses of action, and the uncontrolled variables have been identified and defined, a measure of performance can be developed and selection can be made of a function of this measure to be used as a criterion for the best solution. These two steps require extensive use of decision and value theory.

The type of decision criterion that is appropriate to a problem depends on the state of knowledge regarding possible outcomes. Certainty describes a situation in which each course of action is believed to result in one particular outcome. Risk is a situation in which, for each course of action, alternative outcomes are possible, the probabilities of which are known or can be estimated. Uncertainty describes a situation in which, for each course of action, probabilities cannot be assigned to the possible outcomes.

In risk situations, which are the most common in practice, the objective normally is to maximize expected (long-run average) net gain or gross gain for specified costs, or to minimize costs for specified benefits. A businessman, for example, seeks to maximize expected profits or minimize expected costs. Two objectives, not simply related, may be sought; for example, an economic planner may wish to maintain full employment without inflation. Finally, different groups within an organization may have to compromise their differing objectives, as when an army and a navy, for example, must cooperate in matters of defense.

In approaching uncertainty situations one may attempt either to maximize the minimum gain or minimize the maximum loss that results from a choice; this is the "minimax" approach. Alternatively, one may weigh the possible outcomes to reflect one's optimism or pessimism and then apply the minimax principle. A third approach, minimax regret, attempts to minimize the maximum deviation from the outcome that would have been selected if a state of certainty had existed before the choice had been made.

Each identified variable should be defined in terms of the conditions under which, and research operations by which, questions concerning its value ought to be answered; this includes identifying the scale used in measuring the variable.

Model construction. The type of model described above is called a symbolic model because symbols represent properties of the system. The earliest models were physical representations such as model ships, airplanes, tow tanks, and wind tunnels. Physical models are usually fairly easy to construct, but only for relatively simple objects or systems, and are usually difficult to change.

The next step beyond the physical model is the graph, easier to construct and manipulate, but more abstract. Since graphic representation of more than three variables is difficult, symbolic models came into use. There is no limit to the number of variables that can be included in a symbolic model, and such models are easier to construct and manipulate than physical models.

Symbolic models are completely abstract. When the

Symbolic models

symbols in a model are defined, the model is given content or meaning. This has important consequences. Symbolic models of systems of very different content often reveal similar structure. Hence, most systems and problems arising in them can be fruitfully classified in terms of relatively few structures. Furthermore, since methods of extracting solutions from models depend only on their structure, some methods can be used to solve a wide variety of problems from a contextual point of view. Finally, a system that has the same structure as another, however different the two may be in content, can be used as a model of the other. Such a model is called an analogue. By use of such models much of what is known about the first system can be applied to the second.

Despite the obvious advantages of symbolic models there are many cases in which physical models are still useful, as in testing physical structures and mechanisms; the same is true for graphic models. Physical and graphic models are frequently used in the preliminary phases of constructing symbolic models of systems.

Operations-research models represent the causal relationship between the controlled and uncontrolled variables and system performance; they must therefore be explanatory, not merely descriptive. Only explanatory models can provide the requisite means to manipulate the system to produce desired changes in performance.

Operations-research analysis is directed toward establishing cause and effect relations. Though experiments with actual operations of all or part of a system are often useful, these are not the only way to analyze cause and effect. There are five patterns of model construction, only two of which involve experimentation: inspection, use of analogues, operational analysis, operational experiments, and use of "artificial reality." They are considered here in order of increasing complexity.

Inspection. In some cases the system and its problem are relatively simple and can be grasped either by inspection or from discussion with persons familiar with it. In general, only low-level and repetitive operating problems, those in which human behaviour plays a minor role, can be so treated.

An example is the problem of the newsboy who must decide how many newspapers to order to maximize his expected profit. He buys a certain number of newspapers each day (n, the controlled variable) and either sells some but not all of them (if demand, d, is less than n); or all of them (if demand is equal to or greater than n). He pays an amount a for each paper he buys, and sells each paper for an amount b so that his profit per paper sold is $(b - a)$. Each paper not sold is returned and he receives an amount c so that his loss per paper returned is $(a - c)$. Demand varies from day to day, thus $p(d)$ is the probability that demand will equal d on a randomly selected day. Thus, a, b, c, and d are uncontrolled variables. The measure of performance to be maximized is P, the expected net profit per day, which, of course, can be negative.

The causal structure of this situation is apparent. From it can be deduced an equation whose solution produces the value of n that maximizes P.

Use of analogues. When the researcher finds it difficult to represent the structure of a system symbolically, it is sometimes possible to establish a similarity, if not an identity, with another system whose structure is better known and easier to manipulate. It may then be possible to use either the analogous system itself or a symbolic model of it as a model of the problem system. For example, an equation derived from the kinetic theory of gases has been used as a model of the movement of trains between two classification yards. Hydraulic analogues of economies and electronic analogues of automotive traffic have been constructed with which experimentation could be carried out to determine the effects of manipulation of controllable variables. Thus, analogues may be constructed, as well as found in existing systems.

Operational analysis. In some cases analysis of actual operations of a system may reveal its causal structure. Data on operations are analyzed to yield an explanatory hypothesis, which is tested by analysis of operating data. Such testing may lead to revision of the hypothesis. The cycle is continued until a satisfactory explanatory model is developed.

For example, an analysis of the cars stopping at urban automotive service stations located at intersections of two streets revealed that almost all came from four of the 16 possible routes through the intersection (four ways of entering times four ways of leaving). Examination of the percentage of cars in each route that stopped for service suggested that this percentage was related to the amount of time lost by stopping. Data were then collected on time lost by cars in each route. This revealed a close inverse relationship between the percentage stopping and time lost. But the relationship was not linear, that is, the increases in one were not proportional to increases in the other. It was then found that perceived lost time exceeded actual lost time, and the relationship between the percentage of cars stopping and perceived lost time was close and linear. The hypothesis was systematically tested and verified and a model constructed that related the number of cars stopping at service stations to the amount of traffic in each route through its intersection and to characteristics of the station that affect the time required to get service.

Operational experiments. In some situations it is not possible to isolate the effects of individual variables by analysis of operating data; it may be necessary to resort to operational experiments to determine which variables are relevant and how they affect system performance.

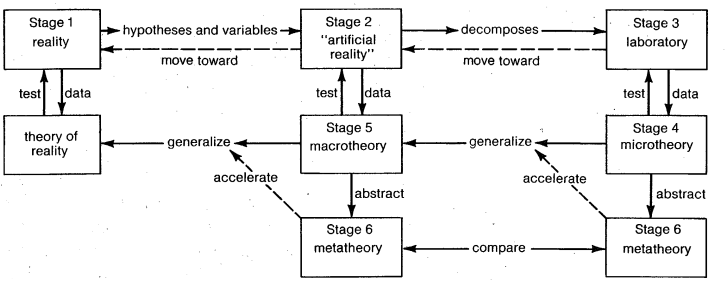
Such is the case, for example, in attempts to quantify the effects of advertising (amount, timing, and media used) upon sales of a consumer product. Advertising by the producer is only one of many controlled and uncontrolled variables affecting his sales. Hence, in many cases its effect can only be isolated and measured by controlled experiments in the field.

The same is true in determining how the size, shape, weight, and price of a food product affect its sales. In this case laboratory experiments on samples of consumers can be used in preliminary stages, but field experiments are eventually necessary. Experiments do not yield explanatory theories, however. They can only be used to test explanatory hypotheses formulated before designing the experiment, and to suggest additional hypotheses to be tested.

Testing hypotheses

Use of "artificial reality." In this last and most complex situation one cannot obtain even a good description of the system's operations, and experimentation on the system itself is ruled out by its size or character; such is the case, for example, in the problem of controlling large-scale social conflicts such as wars or strikes.

In this case model construction is a six-stage process (see illustration). In stage 1 the literature dealing with the problem (reality) is reviewed, relevant hypotheses extracted, and variables identified. In stage 2 an experimental game, called "artificial reality," is designed. The game is complex enough to permit testing most of the hypotheses formulated in stage 1. Behaviour observed in it must be capable of description in quantitative terms. In stage 3 the "artificial reality" game is broken down into the simplest possible experimental situations, which are used to generate laboratory behaviour. In stage 4 a simple theory, or microtheory is developed to explain the behaviour. When this microtheory is tested and found valid,



Model construction for experimentation on a complex system with the use of "artificial reality."

it is expanded to contain behaviour in more complex situations, and this process is repeated until a macrotheory is formulated that explains "artificial reality." In stage 5 "artificial reality" is modified in the direction of reality and stage 4 is repeated, yielding a fresh macrotheory. This process is repeated, and in stage 6 the sequences of micro and macrotheories are analyzed to find a metatheory, a theory of theory generalization: when a metatheory is found that can be applied to the whole sequence of macrotheories it can be used to extrapolate an explanation of the real problem (reality). This pattern is an elaborate way of pulling research up by its own boot straps.

The effect of data availability. It is sometimes necessary to modify an otherwise acceptable model because it is not possible or practical to find the numerical values of the variables that appear in it. For example, a model to be used in guiding the selection of research projects may contain such variables as "the probability of success of the project," "expected cost of the project," and its "expected yield." But none of these may be calculable with any reliability.

Even when the initial construction of a model has been easy, subsequent modification of it so that its variables can be estimated accurately and reliably may be very complex. This often involves a sequence of steps in which the model is complicated but made more usable.

Models not only assist in solving problems but are also useful in formulating them; that is, models can be used as guides to explore the structure of a problem and to reveal possible courses of action that might otherwise be missed. In many cases the course of action revealed by such application of a model is so obviously superior to previously considered possibilities that justification of its choice is hardly required.

In some cases the model of a problem situation may be either too complicated or too large to solve. It is frequently possible to divide the model into individually solvable parts and to take the output of one model as an input to another. Since the models are likely to be interdependent, several repetitions of this process may be necessary.

Deriving solutions from models. Procedures for deriving solutions from models are either deductive or inductive. With deduction one moves directly from the model to a solution in either symbolic or numerical form. Such procedures are supplied by mathematics; for example, the calculus.

Inductive procedures involve trying and comparing different values of the controlled variables. Such procedures are said to be "iterative" (repetitive) if they proceed through successively improved solutions until either an optimal solution is reached or further calculation cannot be justified. A rational basis for terminating such a process—known as the Las Vegas technique—involves the determination of the point at which the expected improvement of the solution on the next trial is less than the cost of the trial.

Such well-known algorithms as linear, nonlinear, and dynamic programming are iterative procedures based on mathematical theory. Simulation and experimental optimization are iterative procedures based primarily on statistics.

Simulation consists of calculating the performance of a system by evaluating a model of it for randomly selected values of variables contained within it. Most simulation in operations research is concerned with "stochastic" variables; that is, variables whose values change randomly within some probability distribution over time. The random sampling employed in simulation requires either a supply of random numbers or a procedure for generating them. It also requires a way of converting these numbers into the distribution of the relevant variable, a way of sampling these values, and a way of evaluating the resulting performance. A number of sampling and estimating procedures are available for this purpose.

Simulation in military games

A simulation in which decision making is performed by one or more real decision makers is called "operational gaming." Such simulations are commonly used in the study of interactions of decision makers as in competitive situations. Military gaming has long been used as a training device, but only relatively recently has it been used for research purposes. There is still considerable difficulty, however, in drawing inferences from operational games to the real world.

Experimental optimization is a means of experimenting on a system so as to find the best solution to a problem within it. Such experiments, conducted either simultaneously or sequentially, may be designed in various ways, no one of which is best in all situations.

Testing the model and the solution. A model may be deficient because it includes irrelevant variables, excludes relevant variables, contains inaccurately evaluated variables, is incorrectly structured, or contains incorrectly formulated constraints. Tests for deficiencies of a model are statistical in nature; their use requires knowledge of sampling and estimation theory, experimental designs, and the theory of hypothesis testing (see also STATISTICS).

Sampling-estimation theory is concerned with selecting a sample of items from a large group and using their observed properties to characterize the group as a whole. To save time and money, the sample taken is as small as possible. Several theories of sampling design and estimation are available, each yielding estimates with different properties.

The structure of a model consists of a function relating the measure of performance to the controlled and uncontrolled variables; for example, a businessman may attempt to show the functional relationship between profit levels (the measure of performance) and controlled variables (prices, amount spent on advertising) and uncontrolled variables (economic conditions, competition). In order to test the model, values of the measure of performance computed from the model are compared with actual values under different sets of conditions. If there is a significant difference between these values, or if the variability of these differences is large, the model requires repair. Such tests do not use data that have been used in constructing the model, because to do so would determine how well the model fits performance data from which it has been derived, not how well it predicts performance.

The solution derived from a model is tested to find whether it yields better performance than some alternative, usually the one in current use. The test may be prospective, against future performance; or retrospective, comparing solutions that would have been obtained had the model been used in the past with what actually did happen. If neither prospective nor retrospective testing is feasible, it may be possible to evaluate the solution by "sensitivity analysis," a measurement of the extent to which estimates used in the solution would have to be in error before the proposed solution performs less satisfactorily than the alternative decision procedure.

The cost of implementing a solution should be subtracted from the gain expected from applying it, thus obtaining an estimate of net improvement. Where errors or inefficiencies in applying the solution are possible, these should also be taken into account in estimating the net improvement.

Implementing and controlling the solution. The acceptance of a recommended solution by the responsible manager depends on the extent to which he believes the solution to be superior to alternatives. This in turn depends on his faith in the researchers involved and their methods. Hence, participation by managers in the research process is essential for success.

Operations researchers are normally expected to oversee implementation of an accepted solution. This provides them with an ultimate test of their work and an opportunity to make adjustments if any deficiencies should appear in application. The operations research team prepares detailed instructions for those who will carry out the solution and trains them in following these instructions. The cooperation of those who carry out the solution and those who will be affected by it should be sought in the course of the research process, not after everything is done. Implementation plans and schedules are pretested and deficiencies corrected. Actual performance of the solution is compared with expectations and, where diver-

gence is significant, the reasons for it are determined and appropriate adjustments made.

The solution may fail to yield expected performance for one or a combination of reasons: the model may be wrongly constructed or used; the data used in making the model may be incorrect; the solution may be incorrectly carried out; the system or its environment may have changed in unexpected ways after the solution was applied. Corrective action is required in each case.

Controlling a solution requires deciding what constitutes a significant deviation in performance from expectations; determining the frequency of control checks, the size and type of sample of observations to be made, and the types of analyses of the resulting data that should be carried out; and taking appropriate corrective action. The second step should be designed to minimize the sum of the costs of carrying out the control procedures and the errors that might be involved.

Since most models involve a variety of assumptions, these are checked systematically. Such checking requires explicit formulation of the assumptions made during construction of the model.

Effective controls not only make possible but often lead to better understanding of the dynamics of the system involved. Through controls the problem-solving system of which operations research is a part learns from its own experience and adapts more effectively to changing conditions.

PROTOTYPE PROBLEMS AND ASSOCIATED TECHNIQUES

As previously mentioned, many operational problems of organized systems have common structures. The most common types of structure have been identified as prototype problems, and extensive work has been done on modelling and solving them.

Though all the problems with similar structures do not have the same model, those that apply to them may have a common mathematical structure and hence may be solvable by one procedure. Some real problems consist of combinations of smaller problems, some or all of which fall into different prototypes. In general, prototype models are the largest that can be solved in one step. Hence, large problems that consist of combinations of prototype problems usually must be broken down into solvable units; the overall model used is an aggregation of prototype and possibly other models.

Solving allocation problems

Allocation. Allocation problems involve the distribution of resources among competing alternatives in order to minimize total costs or maximize total return. Such problems have the following components: a set of resources available in given amounts; a set of jobs to be done, each consuming a specified amount of resources; and a set of costs or returns for each job and resource. The problem is to determine how much of each resource to allocate to each job.

If more resources are available than needed, the solution should indicate which resources are not to be used, taking associated costs into account. Similarly, if there are more jobs than can be done with available resources, the solution should indicate which jobs are not to be done, again taking into account the associated costs.

If each job requires exactly one resource (*e.g.*, one person) and each resource can be used on only one job, the resulting problem is one of assignment. If resources are divisible, and if both jobs and resources are expressed in units on the same scale, it is termed a transportation or distribution problem. If jobs and resources are not expressed in the same units, it is a general allocation problem.

An assignment problem may consist of assigning men to offices or jobs, trucks to delivery routes, drivers to trucks, or classes to rooms. A typical transportation problem involves distribution of empty railroad freight cars where needed, or the assignment of orders to factories for production. The general allocation problem may consist of determining which machines should be employed to make a given product, or what set of products should be manufactured in a plant during a particular period.

In allocation problems the unit costs or returns may be either independent or interdependent; for example, the return from investing a dollar in selling effort may depend on the amount spent on advertising. If the allocations made in one period affect those in subsequent periods, the problem is said to be dynamic, and time must be considered in its solution.

Inventory. An inventory consists of usable but idle resources; these may be men, machines, materials, money, products, or facilities. An inventory problem involves determining how much of a resource to acquire, either by purchasing or producing it, and whether or when to acquire it to minimize the sum of the costs that increase with the size of inventory and those that decrease with increases in inventory. Costs of the first type include the cost of the capital invested in inventory, handling, storage, insurance, taxes, depreciation, deterioration, and obsolescence. Costs that decrease as inventory increases include shortage costs (arising from lost sales), production setup costs, and the purchase price or direct production costs. Setup costs include the cost of placing a purchase order or starting a production run. If large quantities are ordered inventories increase, but the frequency of ordering decreases, hence setup costs decrease. In general, the larger the quantity ordered the lower the unit purchase price because of quantity discounts and the lower production cost per unit resulting from the greater efficiency of long production runs. Other relevant variables include demand for the resource and the time between placing and filling orders.

Applications in inventory problems

Inventory problems arise in a wide variety of contexts; for example, determining quantities of goods to be purchased or produced, how many people to hire or train, how large a new production or retailing facility should be or how many should be provided, and how much fluid (operating) capital to keep available. Inventory models for single items are well developed and are normally solved with calculus. When the order quantities for many items are interdependent (as, for example, when there is limited storage space or production time) the problem is more difficult. Some of the larger problems can be solved by breaking them into interacting inventory and allocation problems. In very large problems simulation can be used to test various relevant decision rules.

Replacement and maintenance. Replacement problems involve items that degenerate with use, or the passage of time, and those that fail after a certain amount of use or time. Items that deteriorate are likely to be large and costly (*e.g.*, machine tools, trucks, ships, and home appliances). Nondeteriorating items tend to be small and relatively inexpensive (*e.g.*, light bulbs, vacuum tubes, ink cartridges). The longer a deteriorating item is operated the more maintenance it requires to maintain efficiency. Furthermore, the longer such an item is kept the less is its resale value and the more likely it is to be made obsolete by new equipment. If the item is replaced frequently, however, investment costs increase. Thus the problem is to determine when to replace such items and how much maintenance (particularly preventive) to perform so that the sum of the operating, maintenance, and investment costs are minimized.

In the case of nondeteriorating items the problem involves determining whether to replace them as a group, or to replace individuals as they fail. Though group replacement is wasteful, labour cost of replacements is greater when done singly; for example, the light bulbs in New York City subway stations are replaced in groups, to save labour. Replacement problems that involve minimizing the costs of items, failures, and the replacement labour are solvable either by numerical analysis or simulation.

The items involved in replacement problems may be people. If so, maintenance can be interpreted as training or improvements in salary, status, or fringe benefits. Failure can be interpreted as departure, and investment as recruiting, hiring, and initial training costs. There are many additional complexities in such cases; for example, the effect of one person resigning or being promoted on the behaviour of others. Such controllable aspects of the environment as location of work and working hours can

have a considerable effect on productivity and failure rates. In problems of this type, the inputs of the behavioral sciences are particularly useful.

Queuing. A queue is a waiting line and queuing involves dealing with items or people in sequence. Thus, a queuing problem consists either of determining what facilities to provide or scheduling the use of them. The cost of providing service and the waiting time of users is minimized. Examples of such problems include determining the number of check-out counters to provide at a supermarket, the number of runways at an airport, parking spaces at a shopping centre, or tellers in a bank. Many maintenance problems can be treated as queuing problems; items requiring repair are like users of a service. Some inventory problems may also be formulated as queuing problems in which orders are like users and stocks are like service facilities.

Sequencing and coordination. In queuing problems, the order in which users waiting for service are served is always specified. Selection of that order so as to minimize some function of the time to perform all the tasks, is a sequencing problem. The performance measure may account for total elapsed time, total tardiness in meeting deadlines and the cost of being late, and the cost of in-process inventories.

The most common context for sequencing problems is a "job-shop," a production facility that processes many different products with many combinations of machines. In this context account may have to be taken of such factors as overlapping service (*i.e.*, if a customer consists of a number of items to be taken through several steps of a process, the first items completing the initial step may start on the second step before the last one finishes the first), transportation time between service facilities, correction of service breakdowns, facility breakdowns, and material shortages.

Some processes consist of a network of unique operations that are carried out only once, some of which can be carried out in parallel, and others of which must be done in a prescribed sequence (*e.g.*, in constructing a building, in carrying out a research and development program, launching a new product, and assembling and testing complex equipment or systems). In such processes coordination consists of establishing starting times and due dates so the total cost of operations is minimized. Penalties for lateness and rewards for early completion may also have to be taken into account. Two techniques have been developed for handling this type of problem, project evaluation and review technique and critical path method.

CPM and PERT techniques

Critical path method (CPM) is an optimizing procedure applicable only to certainty-type formulations of such problems. Project evaluation and review technique (PERT) is applicable to risk- as well as certainty-type formulations but does not always yield optimal solutions. These techniques make it possible to determine labour needs, budget requirements, procurement and design limitations, and the effects of delays or speed-ups and communication difficulties.

Network routing. A network may be defined by a set of points or "nodes" that are connected by lines or "links." A way of going from one node (the "origin") to another (the "destination") is called a "route" or "path." Links, which may be one-way or two-way, are usually characterized by the time, cost, or distance required to traverse them. The time or cost of travelling in different directions on the same link may differ.

A network routing problem consists of finding an optimum route between two or more nodes in relation to total time, cost, or distance. Various constraints may exist, such as a prohibition on returning to a node already visited, or a stipulation of passing through every node once and only once.

Network routing problems commonly arise in communication and transportation systems. Delays that occur at the nodes (*e.g.*, railroad classification yards or telephone switchboards) may be a function of the loads placed on them and their capacities. Breakdowns may occur in either links or nodes. Much studied is the "travelling salesman problem," which consists of starting a route from a designated node that goes through each node (*e.g.*, city) once and only once, and returns to the origin in the least time, distance, or cost. This problem arises in selecting an order for processing a set of production jobs over a facility when the cost of setting up the facility for each job depends on which job has preceded it. In this case the jobs can be thought of as nodes each of which is connected to all of the others and setup costs are the analogue of distances between them. The order that yields the least total setup cost is therefore equivalent to a solution to the travelling salesman problem. The complexity of the calculations is such that even with the use of computers it is very costly to handle more than 20 nodes. Less costly approximating procedures are available, however. More typical routing problems involve getting from one place to another in the least time, cost, or distance. Both graphic and analytic procedures are available for finding such routes.

Competitive problems. Competitive problems deal with choice in interactive situations where the outcome of one decision maker's choice depends on the choice, either helpful or harmful, of one or more others. Examples of these are war, marketing, and bidding for contracts. Competitive problems are classifiable into certainty, risk, or uncertainty types depending on the state of a decision maker's knowledge of his opponent's choices. Under conditions of certainty, it is easy to maximize gain or minimize loss. Competitive problems of the risk type require the use of statistical analysis for their solution; the most difficult aspect of solving such problems usually lies in estimating the probabilities of the competitor's choices; for example, in bidding for a contract on which competitors and their bids are unknown.

The theory of games was developed to deal with a large class of competitive situations of the uncertainty type in which each participant knows what choices he and each other participant has; there is a well-defined "end-state" that terminates the interaction (*e.g.*, win, lose, or draw); and the payoffs associated with each end-state are specified in advance and are known to each participant. In situations in which all the alternatives open to competition, or some of their outcomes are not known in advance, operational gaming can sometimes be used. The military have long constructed operational games; their use by business is relatively recent.

Search problems. Search problems involve finding the best way to obtain information needed for a decision. Though every problem contains a search problem in one sense, situations exist in which search itself is the essential process; for example, in auditing accounts, inspection and quality-control procedures, in exploration for minerals, in the design of information systems, and in military problems involving the location of such threats as enemy ships, aircraft, mines, and missiles.

Errors in search problems

Two kinds of error are involved in search: those of observation and those of sampling. Observational errors, in turn, are of two general types: commission, seeing something that is not there, and omission, not seeing something that is there. In general, as the chance of making one of these errors is decreased, the chance of making the other is increased. Furthermore, if fixed resources are available for search, the larger the sample (and hence the smaller the sampling error), the less resources available per observation (and hence the larger the observational error).

The cost of search is composed of setup or design cost; cost of observations; cost of analyzing the data obtained; and cost of error. The objective is to minimize these costs by manipulating the sample size (amount of observation), the sample design (how the things or places to be observed are selected), and the way of analyzing the data (the inferential procedure).

Almost all branches of statistics provide useful techniques for solving search problems. In search problems that involve the location of physical objects, particularly those that move, physics and some fields of mathematics (*e.g.*, geometry and trigonometry) are also applicable.

A "reversed-search" problem arises when the search procedure is not under control but the object of the

search is. Most retailers, for example, cannot control the manner in which customers search for goods in their stores, but they can control the location of the goods. This type of problem also arises in the design of libraries and information systems, and in laying land and sea mines. These, too, are search problems, and solution techniques described above are applicable to them.

FRONTIERS OF OPERATIONS RESEARCH

Operations research is a rapidly developing application of the scientific method to organizational problems. Its growth has consisted of both technical development and enlargement of the class of organized systems and the class of problems to which it is applied. Therefore, its current frontiers lie in the techniques, organizational types, and classes of problems that it is presently exploring. In this section some of the more important of these frontiers will be identified.

Tactics and strategy compared

Strategic problems. Tactics and strategy are relative concepts; the distinction between them depends on three considerations: (1) the longer the effect of a decision and the less reversible it is, the more strategic it is; (2) the larger the portion of a system that is affected by a decision, the more strategic it is; and (3) the more concerned a decision is with the selection of goals and objectives, as well as the means by which they are to be obtained, the more strategic it is.

Strategy and tactics are separable only in thought, not in action. Every tactical decision involves a strategic choice, no matter how implicit and unconscious it may be. Since the strategic aspects of decisions are usually suppressed, an organization's strategy often emerges as an accidental consequence of its tactical decisions, not as a result of deliberate and conscious choice.

Operations research is becoming increasingly concerned with strategic decisions and the development of explicit strategies for organizations so as to improve the quality of their tactical decisions and make even the most immediate and urgent of these contribute to its long-run goals.

The system-design problem. Operations research has traditionally been concerned with finding effective solutions to specific operational problems. It has developed better methods, techniques, and tools for doing so. But operations researchers have found that too many of their solutions are not implemented and of those that are, too few survive the inclination of organizations to return to familiar ways of doing things. Therefore, operations researchers have gradually come to realize that their task should not only include solving specific problems but also designing problem-solving and implementation systems that predict and prevent future problems, identify and solve current ones, and implement and maintain these solutions under changing conditions.

The planning problem. Operations researchers have come to realize that most problems do not arise in isolation but are part of an interacting system. The process of seeking simultaneous interrelated solutions to a set of interdependent problems is planning. More and more operations-research effort is being devoted to developing a rational methodology for such planning, particularly strategic planning.

Most organizations resist changes in their operations or management. The organizational need to find better ways of doing things is often not nearly as great as is the need to maximize use of what it already knows or has. This is apparent in many underdeveloped countries that, while complaining about the lack of required resources, use what resources they have with considerably less efficiency than do most developed countries. Operations research, therefore, has been addressing itself more and more to determining how to produce the willingness to change.

Types of organization. Operations researchers have become increasingly aware of the need to distinguish between different types of organization because their distinguishing features affect how one must go about solving their problems. Two important classifications exist, the first of which is homogeneous–heterogeneous. Homogeneous organizations are those in which membership involves serving the objectives of the whole (*e.g.*, a corporation or military unit), while heterogeneous organizations are those whose principal objective it is to serve the objectives of its members (*e.g.*, a university or city). The second classification is uninodal/multinodal. Uninodal organizations are hierarchical organizations with a single decision-making authority who can resolve differences between any lower-level decision makers. Multinodal organizations have no such authority but have diffused decision making and hence require agreement among the several decision makers in order to reach conclusions.

Since current skills in operations research are largely restricted to homogeneous uninodal organizations, attempts are under way to develop methodologies adequate for improving the other three types of organization.

In order to solve any of the preceding problems more effectively, operations research requires a better understanding of human behaviour, individual and collective, than is currently available. Furthermore, what understanding the behavioural sciences claim to provide is seldom available in a form that lends itself to symbolic representation and hence to operations research methodology. Operations researchers, therefore, are increasingly working with behavioural scientists to develop behavioural theories that are expressible in a more usable form.

As the scope of problems to which operations research addresses itself increases, it becomes more apparent that the number of disciplines and interdisciplines that have an important contribution to make to their solution also increases. An attempt to provide such a higher-order integration of scientific activity is being made in the management sciences.

BIBLIOGRAPHY

History: J.G. CROWTHER and R. WHIDDINGTON, *Science at War* (1947); F.N. TREFETHEN, "A History of Operations Research," in J.F. MCCLOSKEY and F.N. TREFETHEN (eds.), *Operations Research for Management*, pp. 3–35 (1954); GREAT BRITAIN AIR MINISTRY, *The Origins and Development of Operational Research in the Royal Air Force* (1963).

General texts: R.L. ACKOFF and M.W. SASIENI, *Fundamentals of Operations Research* (1968); S. BEER, *Decision and Control* (1966); D.W. MILLER and M.K. STARR, *Executive Decisions and Operations Research* (1960); H.M. WAGNER, *Principles of Operations Research* (1969).

Methodology: R.L. ACKOFF, *Scientific Method: Optimizing Applied Research Decisions* (1962); C.W. CHURCHMAN, *Prediction and Optimal Decision* (1961); P.C. FISHBURN, *Decision and Value Theory* (1964), *Utility Theory for Decision Making* (1970); D.J. WHITE, *Decision Theory* (1969).

General techniques: B.V. DEAN, M.W. SASIENI, and S.K. GUPTA, *Mathematics for Modern Management* (1963); J.R. EMSHOFF and R.L. SISSON, *Design and Use of Computer Simulation Models* (1970); H. RAIFFA and R. SCHLAIFER, *Applied Statistical Decision Theory* (1961); T.L. SAATY, *Mathematical Methods of Operations Research* (1959); C.J. THOMAS and W.L. DEEMER, JR., "The Role of Operational Gaming in Operations Research," *Operations Research*, 5:1–27 (1957); K.D. TOCHER, *The Art of Simulation* (1963).

Specific techniques: (*Programming*) R.E. BELLMAN, *Dynamic Programming* (1957), and with S.E. DREYFUS, *Applied Dynamic Programming* (1962); S.I. GASS, *Linear Programming: Methods and Applications*, 3rd ed. (1969); R.A. HOWARD, *Dynamic Programming and Markov Processes* (1960); G. HADLEY, *Linear Programming* (1962), *Non-Linear and Dynamic Programming* (1964). (*Inventory*) K.J. ARROW, S. KARLIN, and H. SCARF, *Studies in the Mathematical Theory of Inventory and Production* (1958); G. HADLEY and T.M. WHITIN, *Analysis for Inventory Systems* (1963); F. HANSSMANN, *Operations Research in Production and Inventory Control* (1962). (*Replacement and Maintenance*) R.E. BARLOW and F. PROSCHAN, *Mathematical Theory of Reliability* (1965); D.R. COX, *Renewal Theory* (1962); J.J. MCCALL, "Maintenance Policies for Stochastically Failing Equipment: A Survey," *Management Science*, 11:493–524 (1965). (*Queuing*) D.R. COX and W.L. SMITH, *Queues* (1961); T.L. SAATY, *Elements of Queueing Theory, with Applications* (1961); L. TAKACS, *Introduction to the Theory of Queues* (1962). (*Sequencing and Coordination*) J.E. KELLEY, JR., "Critical-Path Planning and Scheduling: Mathematical Basis," *Operations Research*, 9:296–320 (1961); R.W. MILLER, *Schedule, Cost, and Profit Control with PERT: A Comprehensive Guide for Program Management* (1963); R.L. SISSON, "Sequencing Theory," in R.L. ACKOFF (ed.), *Progress in Operations Research*, vol. 1 (1961). (*Networks*) L.R. FORD, JR. and D.R. FULKERSON, *Flows*

in Networks (1962); E.L. LAWLER and D.E. WOOD, "Branch and Bound Methods: A Survey," *Operations Research*, 14:699–719 (1966); J.D.C. LITTLE *et al.*, "An Algorithm for the Traveling Salesman Problem," *Operations Research*, 11:972–989 (1963). (*Competitive*) N. HOWARD, "The Theory of Meta-Games," and "The Mathematics of Meta-Games," *General Systems*, 11:167–200 (1966); J.C.C. MCKINSEY, *Introduction to the Theory of Games* (1952); A. RAPOPORT, *N-Person Game Theory* (1970); J. VON NEUMANN and O. MORGENSTERN, *Theory of Games and Economic Behavior*, 3rd ed. (1944, reprinted 1953). (*Search*) R.L. ACKOFF and M.W. SASIENI, *Fundamentals of Operations Research*, ch. 14 (1968); B.O. KOOPMAN, "Theory of Search," *Operations Research*, 4:324–346 (1956), 4:503–531 (1956), and 5:613–626 (1957); J.B. MACQUEEN, "Optimal Policies for a Class of Search and Evaluation Problems," *Management Science*, 10:746–759 (1964).

Planning: R.L. ACKOFF, *A Concept of Corporate Planning* (1970); H.I. ANSOFF, *Corporate Strategy: An Analytic Approach to Business Policy for Growth and Expansion* (1965); J.C. EMERY, *Organizational Planning and Control Systems* (1969).

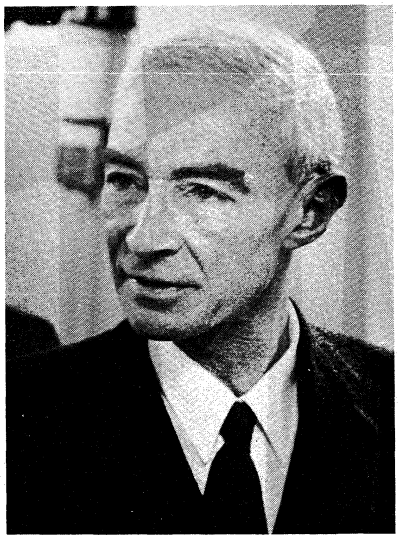
Implementation: J.H.B.M. HUYSMANS, *The Implementation of Operations Research* (1970).

(R.L.A.)

Oppenheimer, J. Robert

J. Robert Oppenheimer, American theoretical physicist and science administrator, directed the Manhattan Project, which produced the first atomic bomb during World War II. Accusations as to his loyalty and reliability as a security risk led to a government hearing that resulted in the loss of his security clearance and of his position as adviser to the highest echelons of the U.S. government. The case became a cause célèbre in the world of science because of its implications concerning political and moral issues relating to the role of scientists in government.

By courtesy of the Los Alamos Scientific Laboratory, New Mexico



Oppenheimer.

Oppenheimer was born on April 22, 1904, in New York City, where his father, who had emigrated from Germany at the age of 14, had made his fortune by importing textiles. During his undergraduate education at Harvard University, Oppenheimer excelled in Latin, Greek, physics, and chemistry, published poetry, and studied Oriental philosophy. After graduating in 1925, he sailed for England to do research at the Cavendish Laboratory at Cambridge University, which, under the leadership of Lord Rutherford, had an international reputation for its pioneering studies on atomic structure. At the Cavendish, Oppenheimer had the opportunity to collaborate with the British scientific community in its efforts to advance the cause of atomic research.

Training in physics

Max Born invited him to Göttingen University, where he met other prominent physicists, such as Niels Bohr and Paul Dirac, and where, in 1927, he received his doctorate. After short visits at science centres in Leiden and Zürich, he returned to the United States to teach physics at the University of California at Berkeley and the California Institute of Technology.

In the 1920s the new quantum and relativity theories were engaging the attentions of science. That mass was equivalent to energy and that matter could be both wave-like and corpuscular carried implications seen only dimly at that time. Oppenheimer's early research was devoted in particular to energy processes of subatomic particles, including electrons, positrons, and cosmic rays. Since quantum theory had been proposed only a few years before, the university post provided him an excellent opportunity to devote his entire career to the exploration and development of its full significance. In addition, he trained a whole generation of American physicists, who were greatly affected by his qualities of leadership and intellectual independence.

The rise of Hitlerism in Germany stirred his first interest in politics. In 1936 he sided with the republic, during the Civil War in Spain, where he became acquainted with Communist students. Although his father's death in 1937 left Oppenheimer a fortune that allowed him to subsidize anti-Fascist organizations, the tragic suffering inflicted by Stalin on Russian scientists led him to withdraw his associations with the Communist Party—in fact, he had never joined the party—and at the same time reinforced in him a liberal–democratic philosophy.

After the invasion of Poland by Nazi Germany in 1939, the physicists Albert Einstein and Leo Szilard warned the American government of the danger threatening all of humanity if the Nazis should be the first to make a nuclear bomb. Oppenheimer then began to seek a process for the separation of uranium-235 from natural uranium and to determine the critical mass of uranium required to make such a bomb. In August 1942 the U.S. Army was given the responsibility of organizing the efforts of British and American physicists to seek a way to harness nuclear energy for military purposes, an effort that became known as the Manhattan Project. Oppenheimer was instructed to establish and administer a laboratory to carry out this assignment. In 1943 he chose the plateau of Los Alamos, near Santa Fe, New Mexico, where he had spent part of his childhood in a boarding school.

The Manhattan Project

For reasons that have not been made clear, Oppenheimer in 1942 initiated discussions with military security agents that culminated with the implication that some of his friends and acquaintances were agents of the Soviet government. This led to the dismissal of a personal friend on the faculty at the University of California. In a 1954 security hearing he described his contribution to those discussions as "a tissue of lies."

The joint effort of outstanding scientists at Los Alamos culminated in the first nuclear explosion on July 16, 1945, at Alamogordo, New Mexico, after the surrender of Germany. In October of the same year, Oppenheimer resigned his post. In 1947 he became head of the Institute for Advanced Study at Princeton University and served from 1947 until 1952 as chairman of the General Advisory Committee of the Atomic Energy Commission, which in October 1949 opposed development of the hydrogen bomb.

On December 21, 1953, he was notified of a military security report unfavourable to him and was accused of having associated with Communists in the past, of delaying the naming of Soviet agents, and of opposing the building of the hydrogen bomb. A security hearing declared him not guilty of treason but ruled that he should not have access to military secrets. As a result, his contract as adviser to the Atomic Energy Commission was cancelled. The Federation of American Scientists immediately came to his defense with a protest against the trial. Oppenheimer was made the worldwide symbol of the scientist, who, while trying to resolve the moral problems that arise from scientific discovery, becomes the victim of a witch-hunt. He spent the last years of his life working out ideas on the relationship between science and society.

The Cold War having declined, Pres. Lyndon B. Johnson in 1963 formalized Oppenheimer's reinstatement by

presenting him the Fermi Award of the Atomic Energy Commission. He retired from Princeton in 1966 and died of throat cancer on February 18, 1967, in Princeton, New Jersey.

BIBLIOGRAPHY. MICHEL ROUZE, *Oppenheimer* (1962; Eng. trans., *Robert Oppenheimer: The Man and His Theories*, 1965), is the only biography (to 1961). Oppenheimer's philosophical ideas are expressed in his two books, *Science and the Common Understanding* (1954) and *The Open Mind* (1955). His scientific work is spread in many reports, published in particular in *Physical Review*, from 1928 to 1948. ROBERT JUNGK, *Heller als tausend Sonnen: Das Schicksal der Atomforscher* (1956; Eng. trans., *Brighter Than a Thousand Suns: Moral and Political History of the Atomic Scientists*, 1958), is the story of the dramatic events lived by the nuclear physicists during World War II and the Cold War. *In the Matter of J. Robert Oppenheimer: Transcript of Hearings Before Personnel Security Board, Washington, D.C. April 12, 1954 Through May 6, 1954* (1954), is the fundamental document on Oppenheimer's trial by the Atomic Energy Commission. Among the numerous publications on this trial (all of them being in favour of Oppenheimer), see J. and S. ALSOP, *We Accuse!* (1954); J. ALVIN KUGELMASS, *J. Robert Oppenheimer and the Atomic Story* (1953); PHILIP M. STERN, *The Oppenheimer Case: Security on Trial* (1969); PETER MICHELMORE, *The Swift Years: The Robert Oppenheimer Story* (1969). HAAKON CHEVALIER, the professor denounced by Oppenheimer, describes this issue in *The Man Who Would Be God* (1959) and *Oppenheimer: The Story of a Friendship* (1965).

(Mi.Ro.)

Optical Engineering

Optical engineering is the branch of technology concerned with the design, construction, and testing of optical equipment. The word optics, of Greek origin, relates to what is seen, and optical instruments historically have been aids to vision. With the advance of technology, however, and the introduction of other means for detecting light, such as photographic films and photoelectric devices, the meaning of the term has been broadened, and today several kinds of optical instrument are recognized, including (1) systems for projecting a beam of light in a desired direction, such as the searchlight and lighthouse; (2) aids to vision such as spectacles, microscopes, and telescopes; (3) optical recording devices, such as cameras; (4) apparatus for information transfer and display, such as movie projectors, planetariums, and television sets; (5) detectors of invisible radiation; (6) measuring devices, such as surveying instruments and rangefinders; (7) instruments for measuring and analyzing light, such as photometers, colorimeters, and spectroscopes.

Though simple lenses had been in use as magnifiers for over a thousand years, and though eyeglasses had been developed in the 14th century, it is generally agreed that optical engineering began in the early 17th century with the development of the first precision optical instrument, the telescope. The microscope was developed almost simultaneously. Because manufacturing problems were little understood, and available glass was poor, early instruments were primitive. Newton, despairing of making adequate lenses for a telescope, turned to mirrors instead, while Dutch microscopist Antonie van Leeuwenhoek in 1670 found that he could see more through a small glass bead used as a magnifier than he could through his microscope.

By the early 19th century, however, excellent lenses were being made, and by the time photography was introduced by L.-J.-M. Daguerre in 1839, the art of lens design was developing rapidly (see also OPTICS, PRINCIPLES OF; TELESCOPE; and PHOTOGRAPHY, TECHNOLOGY OF).

Function of the optical engineer. Until the specialized nature and national importance of the subject were appreciated during World War I, the engineering problems involved in the design of optical equipment were given insufficient attention. Though mechanical engineers, physicists, and mathematicians were commonly involved, it was considered revolutionary when Ernst Abbe, the director of research for the Zeiss optical company in Germany, about 1870 decreed that all Zeiss instruments had to be completely designed on paper before any construction could be started. Today, a number of universities have departments entirely given over to the study of optical engineering. Because of the increasing complexity of optical devices, prospective optical engineers usually receive a broad training that includes considerable mathematics and physics, together with courses in chemistry, mechanical engineering, and electronics, in addition to studies in vision, lenses, and optics.

Training of engineers

Optical engineering students ultimately specialize in either instrument design or lens design, since the problems in optical engineering fall naturally into these two distinct parts. The determination of the exact specification of the optical system itself is the concern of the lens designer, while the design of the supporting body, which also provides the necessary movements and controls, is the job of the instrument designer. The optical device may be as simple as a pocket lens or an opera glass, or it may be a large, complex structure requiring a special building to house it.

If the instrument is very simple, or the producing company very small, the lens designer may plan and work out the whole system according to the customer's requirements and employ a draftsman or instrument designer to complete the mount design and prepare detailed drawings for the factory. After assembly, the instrument will be returned to him for testing and release. When the instrument is complex or large, and sufficient staff is available, the optical engineer having overall responsibility for the project will prepare a preliminary layout of the system to meet the customer's requirements. The lens design department will then work out in detail a suitable optical system and will specify tolerances; that is, the permissible variations in dimensions. A mechanical engineer, and if necessary an electronics expert, will develop and prepare detailed drawings for the factory. Today the design and construction of the optical parts, and sometimes electronic parts, are increasingly subcontracted to companies specializing in these fields, the original contractor then being responsible for the preliminary design, for providing metal or plastic parts, and for performing the assembly and final testing of the whole instrument.

In a large company, the industrial design and sales departments are concerned with appearance, weight, convenience, durability, cost, and any other nonscientific factors that may enter into ensuring the total satisfaction of the customer.

DESIGN PROBLEMS IN OPTICAL ENGINEERING

The optical image. The work of designing an optical instrument is not basically different from that of designing any other apparatus, so far as mechanical and electronic features are concerned. The distinctive task of the optical engineer is to develop a suitable optical system to form an image of the required sharpness at a specified position. Images—*i.e.*, optical pictures—may be real or virtual. A real image is formed outside the system where it can be projected on a screen or received on a piece of film, while a virtual image can be seen only by looking into the instrument.

Real and virtual images

To the ancients, image formation was a mystery, and not until the 17th century did the explanation of it begin to become clear. The German astronomer Johannes Kepler assumed that every point in an object emits a bundle of light rays that are either refracted (bent) at a transparent surface, or reflected at a mirror, and come together again at another point called the "image" of the original object point (see Figure 1A). An object was regarded as an assemblage of millions of separate points each of which is independently "imaged" by the optical system. An ideal or perfect lens would focus all of the rays from an object point P at an exact point P′ in the image; and, furthermore, if the object points all lie in a plane perpendicular to the lens axis, as when viewing the wall of a house, then the image points would also lie in a plane, and the image of the wall would be flat. For a perfect lens, the image magnification ratio (image size to object size or h'/h) would be constant over the entire field.

Later workers, including the German mathematician and astronomer C.F. Gauss (1777–1855), in attempting

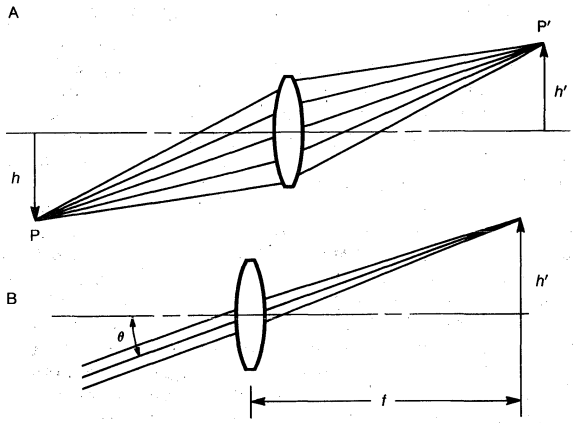


Figure 1: *Image formation.*
(A) Johannes Kepler's theory (see text). (B) Relation between focal length *(f)* and image size *(h')*; namely, $f = h'/\tan\theta$.

to develop a quantitative theory of image formation, introduced the concepts of lens power and focal length. The focal length of a lens, which determines the size of an image formed by the lens, is the distance from the centre of the lens to the point at which the image of a distant object is formed. More specifically, it is the ratio of the image height to the tangent of the angle subtended at the lens by the distant object (Figure 1B). A long-focus lens, therefore, forms a large image, while a short-focus lens forms a small image, of a distant object.

Types of lenses. A lens, named for its resemblance to the seed of the lentil plant, is a piece of transparent material, usually circular in shape, with two polished surfaces, either or both of which is curved and may be convex (bulging) or concave (depressed). The curves are almost always spherical; *i.e.*, the radius of curvature is constant. A lens has the valuable property of forming images of objects situated in front of it. Single lenses are used in spectacles, pocket magnifiers and reading glasses, projection condensers, signal lights, viewfinders, and on simple box cameras. More often a number of lenses made of different materials are combined together as a compound lens in a tube to permit correction of aberrations (see below). The simplest compound lens is a thin cemented combination of two single lenses (doublet) such as a small telescope objective (the lens nearest the object) or a component of an erector lens or eyepiece. Microscope objectives may contain as many as eight or nine elements or single lenses, some of which may be made of materials other than glass; *i.e.*, crystalline calcium fluoride to help the designer to bring all colours of light to a common focus. Photographic objective lenses may contain as few as two or as many as ten or more elements, while a so-called zoom or variable focal length lens may include as many as 18 or 20 elements in several groups, the different groups being movable along the axis by levers or cams at different rates to produce the desired change in focal length without a shift of the image plane. The design of a compound lens system is difficult and laborious; the process is outlined below.

Fluoride lens elements

Aberrations. A single lens is far from the perfect optical system envisaged by Kepler and Gauss. The cone of rays emitted by a single point in the object is not united in a perfect point by the lens but instead forms a small patch of light. In the 19th century the German scientist L. P. Seidel analyzed the defects of a single lens and identified seven distinct kinds of imperfection, now usually called aberrations, to be found in the image of a single object point. Two aberrations, field curvature and distortion, affect the position of the image point in relation to its ideal position in a flat plane; two chromatic (having to do with colour) aberrations result in displacements of the image point as the colour of the light is varied; and the remaining three aberrations (known as spherical aberration, coma, and astigmatism) are concerned with the size and shape of light patch in light of a single colour. When spherical aberration is present, rays from an object point passing through different zones of the lens come to a focus at different distances along the axis; if coma is present, the rays come to focus at different heights above the axis; and if astigmatism is present, the image of a point breaks up into two short focal lines, one being radial to the field and one tangential to it. Invariably, all three of these aberrations are present together and produce undesirable blurring of the image. Stopping the lens down to a smaller aperture (using only the central part of the lens) generally reduces the blurring, but it has little if any effect on the field curvature, distortion, or either of the chromatic aberrations.

Diffraction and resolving power. Another phenomenon that can have considerable effect on the sharpness of an image is diffraction, the slight bending of light rays around the edge of an obstacle. Because light consists of a train of waves, light rays are really only the paths along which light travels, and if a ray passes very close to the edge of an obstacle such as a diaphragm, which controls the size of the aperture opening, or a lens mount, the light is liable to bend slightly outwards and cause a blurring of the image. The magnitude of the blurring increases as the size of the lens aperture decreases, being almost negligible in a large lens but decidedly noticeable in a lens having a small opening, say less than about five millimetres (0.2 inches) in diameter. As diffraction is an inherent property of light, it cannot be eliminated by changing the structure of the lens. It must, however, be taken into account by the optical engineer when designing systems of very small aperture.

A frequently used measure of lens quality is the ability to form an image that is sharp enough to separate or "resolve" two very close dots or lines in an object. Resolving power depends on how much the various aberrations are corrected and on the diffraction of light at the rim of the lens aperture. The resolving power of a perfect telescope (expressed in seconds of arc) is about 4.5 divided by the diameter of the objective in inches, but very few astronomical telescopes are capable of resolving much less than one second of arc. The limiting factors are usually air currents and atmospheric turbulence, and not aberrations or diffraction.

Resolving power of lenses

Image brightness. A major goal in optical engineering is to achieve images that are as bright as possible. The brightness of the image of an extended object formed by a photographic lens is proportional to the brightness of the object divided by the square of the *f*-number of the lens; *i.e.*, by the square of the ratio of the focal length of the lens to its aperture, or $(2.8)^2$ for an $f/2.8$ lens. The image brightness in a telescope follows this law for an extended object such as the moon or a nebula, but for point objects such as stars the image brightness depends only on the area of the telescope objective. Thus, to see a star brightly against the continuous sky background, the telescope objective must have a large diameter and a high *f*-number, which leads to the requirement of a very long focal length.

Lens materials. Lenses are almost always made from some particular type of optical glass, although for work in the ultraviolet and infrared regions of the spectrum special crystals are used because glass is generally opaque to those wavelengths. Hundreds of different types of optical glass are manufactured today, distinguished by their ability to bend light (refractive index) and ability to separate light into different wavelengths (dispersive power). At first lenses were made from selected pieces of window glass or the glass used to make blown tableware. In the early 1800s, the manufacture of specially clear glass for lenses began in Europe. The glass was slowly stirred in the molten state to remove striations and irregularities, and then the whole mass was cooled and broken up into suitable pieces for lens making. The pieces were placed in molds of the approximate size of the lens, slowly remelted to shape, and carefully annealed; *i.e.*, allowed to cool slowly under controlled conditions. Various chemicals were added in the molten state to vary the properties of the glass: lead oxide, for example, was found to raise both the refractive index and the dispersive power. In 1884 it was discovered that barium oxide had the effect of raising the refractive index without increasing the disper-

sion, a property that proved to be of the greatest value in the design of photographic lenses known as anastigmats (no astigmatic aberration). In 1938 a further major improvement was achieved by the use of various rare-earth elements, and since 1950 lanthanum glass has been commonly used in high-quality photographic lenses.

The cost of optical glass varies considerably, depending on the type of glass, the precision with which the optical properties are maintained, the freedom from internal striae and strain, the number of bubbles, and the colour of the glass. Many common types of optical glass are now available in quite large pieces, but as the specifications of the glass become more stringent the cost rises and the range of available sizes becomes limited. In a small lens such as a microscope objective, the cost of the glass is insignificant, but in large lenses in which every millimetre of thickness may represent an additional pound in weight, the glass cost can be very high indeed (see also GLASS PRODUCTS AND PRODUCTION).

Plastic lenses

Lenses can be molded successfully of various types of plastic material, polymethyl methacrylate being the most usual. Even multi-element plastic lenses have been manufactured for low-cost cameras, the negative (concave) elements being made of a high-dispersion plastic such as styrene.

Other basic components of optical systems. These include mirrors, light sources, detectors, projection screens, reflecting prisms, dispersing devices, filters and thin films, and fibre-optics bundles.

Mirrors. Mirrors are frequently used in optical systems. Plane mirrors may be employed to bend a beam of light in another direction, either for convenience or to yield an image reversed left for right if required. Curved mirrors, concave and convex, may be used in place of lenses as image-forming elements in reflecting telescopes. All of the world's largest telescopes and many small ones are of the reflecting type. Such telescopes use a concave mirror to produce the main image, a small secondary mirror often being added to magnify the image and to place it in a convenient position for observation or photography. Telescope mirrors are commonly made parabolic or hyperbolic in section to correct the aberrations of the image. Originally telescope mirrors were made from polished "speculum metal," an alloy of copper and tin, but in 1856 Justus von Liebig, a German chemist, invented a process for forming a mirror-like layer of silver on polished glass, which was applied to telescope mirrors by the German astronomer C.A. von Steinheil. Today most mirrors are made of glass, coated with either a chemically deposited silver layer or more often one made by depositing vaporized aluminum on the surface. The aluminum surface is as highly reflective as silver and does not tarnish as readily.

A large astronomical mirror presents many problems to the optical engineer, mainly because even a distortion of a few millionths of an inch of the mirror under its own weight will cause an intolerable blurring of the image. Though many schemes for supporting a mirror without strain have been tried, including one to support it on a bag of compressed air, the problem of completely eliminating mirror distortion remains unsolved. A metal mirror, if well ribbed on the back, may be lighter than a glass mirror and therefore easier to handle, but most metals are slightly flexible and require just as careful support as glass mirrors. Since temperature changes can also cause serious distortion in a mirror, astronomers try to hold observatory temperatures as constant as possible.

Light sources. Many types of optical instruments form images by natural light, but some, such as microscopes and projectors, require a source of artificial light. Tungsten filament lamps are the most common, but if a very bright source is required, a carbon or xenon arc is employed. For some applications, mercury or other gas discharge tubes are used; a laser beam is often employed in scientific applications. Laser light is brilliant, monochromatic, collimated (the rays are parallel), and coherent (the waves are all in step with each other), any or all of these properties being of value in particular cases (see also LASER AND MASER).

Use of laser light

Detectors. The image formed by an optical system is usually received by the eye, which is a remarkably adaptable and sensitive detector of radiation within the visible region of the electromagnetic spectrum (see LIGHT). A photographic film, another widely used detector, has the advantage of yielding a permanent record of events. Since about 1925 many types of electrical detectors of radiation, both within the visible region and beyond it, have been developed. These include photoelectric cells of various kinds in which either a voltage or a resistance is modified by light falling on the device. Many new types of detector are sensitive far into the infrared spectrum and are used to detect the heat radiated by a flame or other hot object. A number of image intensifiers and converters, particularly for X-ray or infrared radiation, which have appeared since World War II, embody a radiation detector at one end of a vacuum tube and an electron lens inside the tube to relay the image on to a phosphor screen at the other end. This arrangement produces a visible picture that may be observed by eye or photographed to make a permanent record.

Television camera tubes detect real images by electronic scanning, the picture on the viewing tube being a replica of the image in the original camera. The combined application of electronics and optics has become common. An extreme example of electro-optics appears in some space cameras, in which the film is exposed, processed, and then scanned by a tiny point of light; the light passing through the film is picked up by a photocell and transmitted to earth by radio, where it is made to control the brightness of another point of light scanning a second piece of film in exact synchronism with the scanning spot in the camera. The whole system thus produces a picture on earth that is an exact replica of the picture photographed in space a few minutes earlier.

Projection screens. The simplest screen for the projection of slides or motion pictures is, of course, a matte white surface, which may be on a hard base as in outdoor theatres or on a stretched cloth indoors. A theatre screen is often perforated to transmit sound from loudspeakers placed behind it.

Improved screen materials have been developed to increase the brightness of the picture to suit the particular shape of the auditorium. A screen covered with tiny beads tends to send the light back in the general direction of the projector, and is suitable for use at one end of a long, narrow auditorium. Another type of screen is covered with fine embossed vertical grooves; this tends to distribute the light in a horizontal band across the audience with little or no vertical spread. A real advantage of these highly reflective screens is that they tend to reflect ambient room light away from the viewer as by a mirror, so that the pictures appear almost as bright and clear by day as in a darkened room.

Reflecting prisms. Reflecting prisms are pieces of glass bounded by plane surfaces set at carefully specified angles. Some of these surfaces transmit light, some reflect light, while some serve both functions in succession. A prism is thus an assembly of plane reflectors at relatively fixed angles, which are traversed in succession by a beam of light.

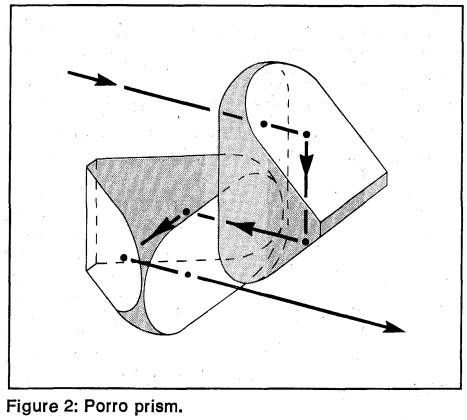


Figure 2: Porro prism.

Prism binoculars

The simplest prism is a triangular block of glass with two faces at right angles and one at an angle of 45°. The face at 45° deflects a beam of light through a right angle. The common Porro prism used in a pair of binoculars contains four 45° reflecting surfaces, two to reverse the beam direction in the vertical plane and two in the horizontal plane (Figure 2). These reflecting faces could be replaced by pieces of mirror mounted on a metal frame, but it is hard to hold mirrors rigidly and harder still to keep them clean. Some microscopes are equipped with a 45° deflection prism behind the eyepiece; this prism may provide two or three reflections depending on the type of image inversion or left-for-right reversal required.

Prisms containing a semireflecting, semitransmitting surface are known as beam splitters and as such have many uses. An important application is found in some colour television cameras, in which the light from the lens is divided by two beam splitters in succession to form red, green, and blue images on the faces of three image tubes in the camera.

Dispersing devices. There are two forms of dispersing element used to spread out the constituent colours of a beam of light into a "spectrum," namely a prism and a grating. The prism, known to Newton, is the older; it separates the colours of the spectrum because the refractive index of the glass is lowest for red light and progressively increases through the yellow and green to the blue, where it is highest. Prism spectroscopes and spectrographs are made in a variety of forms and sizes, but in all cases the blue end of the spectrum is greatly spread out while the red end is relatively compressed.

A diffraction grating is a ruled mirror or transparent plate of glass having many thousands of fine parallel grooves to the inch. It separates the colours of the spectrum by a process of diffraction. Each groove diffracts, or scatters, light in all directions, and in the case of light of one particular wavelength, there will be one direction in which the light wave from one groove lags behind the light wave from the next groove by precisely one or more whole wavelengths. This results in a strong beam of diffracted light in that direction and darkness in all other directions. Since each spectral colour corresponds to a different wavelength, the grating spreads out the spectrum into a fan where it can be observed or photographed. The red rays are bent most and the blue rays least, the opposite of the situation with a prism.

Although a prism or grating is the essential dispersing element in a spectrograph, a fine slit and additional lenses or focussing mirrors must be used to form a sharply defined spectrum. Prism spectroscopes are, of course, limited to those wavelengths for which the prism material is transparent; a reflecting grating can be used for any wavelength that the material will reflect.

Filters and thin films. A colour filter is a sheet of transparent material that modifies a light beam by selective absorption of some colours in relation to others. A neutral filter absorbs all wavelengths equally and merely serves to reduce the intensity of a beam of light without changing its colour.

Filters may be made from sheets of coloured glass, plastic, or dyed gelatin, and in some cases glass cells filled with liquid have been used. Since World War II, another type of filter depending on the interference of light has been developed in which one or more metallic or other types of films of controlled thickness have been deposited on a glass plate, the layers being so thin as to cause selective interference of some wavelengths in relation to others and thus act as a nonabsorbing filter. In this case the rejected colours are reflected instead of being absorbed.

Polarizing filters have the property of transmitting light that vibrates in one direction while absorbing light that vibrates in a perpendicular direction. These filters are used extensively in scientific instruments. In sunglasses and when placed over a camera lens, polarizing filters reduce unwanted reflections from nonmetallic surfaces. Polarizing spectacles have been used to separate the left-eye and right-eye beams in the projection of stereoscopic pictures or movies.

Fibre-optics bundles. A thin rod or fibre of glass or other transparent material transmits light by repeated internal reflections, even when the rod is somewhat curved. An ordered bundle of rods or fibres is thus capable of taking an image projected upon one end of the bundle and reproducing it at the other end. A fibre-optics bundle can be fused together into a rigid channel, or it may be left flexible, only the ends being rigidly fastened together. A flexible fibre bundle has many uses, especially in medical instruments. Because it is exceedingly delicate, however, it must be handled with care; breaking a fibre would cause a black dot to appear in the reproduced image.

Lens design. The preliminary layout of a new optical instrument consists of a number of lenses, mirrors, prisms, and other optical components forming together the basic conception of an optical system that will produce the desired image at the specified location. At this stage lenses are defined merely by their focal length, diameter, and field to be covered. The lens designer must take each item in turn and work out a precise "formula" for manufacture. A lens formula consists of a list of lens elements with their surface curvatures, thicknesses, diameters, and separations, with tolerances, and the types of glass to be used. An indication of the anticipated image quality, expressed in terms of some well-understood graphs or figures, is often included.

The process of lens design begins with the choice of a promising starting point. Company records or patent files are helpful, but if the system is totally novel, the designer relies on his own experience and imagination. Having selected a suitable starting lens, the designer must evaluate it very carefully within the limitations of his particular problem. Thus, for example, the lens may have to cover an angular field of $\pm 20°$ at an aperture of $f/6$ using red light, with a very distant object, or it may have to work at unit (one-to-one) magnification, in white light. Whatever the circumstances, the designer will trace a number of rays from each of several object points by accurate trigonometrical formulas, working to 7 decimal places and expressing angles to less than one second of arc, to see how closely the various rays come together in the focal plane after passing through the lens. The departures of the rays from their ideal meeting point are subdivided for convenience into their specific aberrations, which he will tabulate.

Minimizing aberrations

The next problem is to decide what changes to make in the structure of the lens to reduce the aberrations to a minimum. Any single change in any lens parameter, be it a surface curvature, thickness, airspace, or refractive index, will change all the aberrations and will probably make some of them better and some worse. Therefore it is necessary to make several changes simultaneously if several aberrations are to be reduced together. An experienced designer with a good background in optical theory will often know what changes should be made, but if not, he will make a small change in each parameter separately to determine the effect of this change on each aberration and assemble a table of coefficients, or rate-of-change factors. Knowing what changes are required in each of the various aberrations, the designer can then draw up a set of simultaneous equations, and by solving them he can ascertain what changes should be made in each of the variables in the system. Unfortunately, these changes will not eliminate all the aberrations, because having changed the first curvature, the rays will travel along a different path, and all subsequent coefficients will be in error. For this reason, the designer makes a small fraction of the indicated changes and then recomputes the coefficients and repeats the process.

When working by hand, using logarithms or a desk calculator, a designer becomes remarkably familiar with a particular lens and soon knows what to expect from a given change or series of changes. He therefore does not have to compute all the coefficients every time and can make various shortcuts. The designer keeps watch over many other factors. He may decide, for instance, to maintain the focal length by appropriate changes in the last surface curvature, and he watches the length of the sys-

tem and the diameters and thicknesses of the individual elements to prevent them from becoming unduly large. He also watches how the oblique light beams are trimmed down by the end apertures of the lens and strives to keep the loss of light by this "vignetting" from becoming excessive.

In the early 1950s small electronic computers became available, and lens designers were among the first to make use of them. As a result, the time taken to trace one ray through one refracting surface was reduced from about three minutes to perhaps one second. The computer also brought relief from the deadening labour of hand calculations. Soon lens designers began to think that it might be possible for the computer to do much more than merely trace rays. This hope was realized within less than ten years, and by 1965 it was possible for a very large computer to go through all the steps performed by a human designer, solving the simultaneous equations and changing every lens parameter in such a way as to improve the definition to the greatest possible degree. Furthermore, most computer programs take into consideration such matters as vignetting, focal length, lens size, and the thicknesses of the elements—even for a zoom lens or an enlarging lens that must function properly over a wide range of magnifications. Lenses designed by means of high-speed lens-improvement computer programs are much better than those designed by hand only a few years earlier.

When using a large computer the old well-known "aberrations" are often ignored, and the behaviour of a trial lens is expressed by the shape of the wave front emerging from the rear of the lens. In a perfect lens the emerging wave front would be a sphere centred about the ideal image position in the focal plane, and any departure of the actual wave front from this ideal sphere represents a degree of imperfection in the lens, and hence an "aberration." Aberrations expressed in this way are often easier to handle in a computer than the classical ray aberrations.

THE MANUFACTURE AND TESTING OF OPTICAL PARTS

The traditional processes of lens and prism manufacture are still regularly employed, and indeed the only major innovation since the 1920s is the use of diamond milling, which makes it possible to generate rapidly a close approximation to a required form, be it a lens, mirror, or prism.

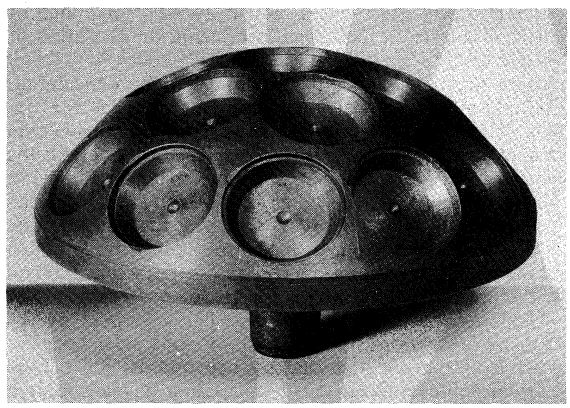


Figure 3: A recessed block to hold several lenses during smoothing and polishing.

Process of lens polishing

Grinding and polishing. Lens surfaces are ground on an iron tool, either flat or suitably curved, using progressively finer grades of wet emery or carborundum powder as abrasives. A number of fine-ground lens blanks are then mounted with pitch on a block (Figure 3) so that they can be polished together. The polishing tool is covered with a thin layer of pitch, wax, or even coarse cloth. Wet rouge or certain other mineral oxides are used as polishing materials. The polishing of glass is a slow process, requiring lenses to be oscillated back and forth, sometimes for hours, against the rotating polisher. Fairly high pressure is used. When both sides of the lens have been polished, it is held between two concentric metal cups by which the lens can be rotated about its optical axis while the edge is ground to size. If a compound objective is being manufactured, several single lenses must be mounted together in a precise coaxial arrangement; the thicknesses, separations, and centring must be kept very close to the prescribed values or the aberration corrections laboriously determined by the lens designer will not be realized.

The lens designer, with his high-speed computer, can be of great assistance to the manufacturer by providing him with a "change table," which indicates numerically how each aberration or other property of a lens changes as a result of small errors in the realization of each constructional parameter. By intelligent use of this table, together with data on the effects of a slight tilting of each surface, it is possible to establish an exact tolerance for every dimension of a lens, thus indicating to the maker just where he must take extra care and where such care is not needed. Refractive index tolerances, for example, indicate whether regular grades of optical glass can be used or whether it is necessary to purchase the glass first and then complete the design using the refractivity data on the particular glasses that will be used. The tilt coefficients indicate which are the more sensitive surfaces requiring support by a rigid ring in the mount and which are the less sensitive surfaces that may be held by other methods.

The standard processes of lens manufacture can be used to generate spherical surfaces with great accuracy, but there are many circumstances in which a nonspherical or aspheric surface is desirable, as for instance the parabolic mirror in a telescope. In spite of the efforts of many inventors, it is still virtually impossible to generate an aspheric surface by machine, and all precise aspheric lenses require a greater or lesser amount of handwork in their manufacture, making them very expensive to produce.

Cylindrical surfaces can be polished by surface contact like spheres, provided relative rotation is prevented between the work and the polisher. Such surfaces are used in lenses for wide-screen motion picture photography and projection and other applications in which the image magnification is required to be different in two perpendicular directions. The surfaces on spectacle lenses are often toric, like the outer face of a curtain ring or doughnut. These are generated by special machines and polished with a soft polisher capable of readily following whatever surface has been generated on the blank. Frequently an antireflective coating is applied to the surfaces of lens elements to reduce the amount of light that is lost by reflections.

Lens testing. A spherical lens surface is checked during manufacture by laying an oppositely polished surface on it—*i.e.*, concave on convex—and observing the interference colours formed in the thin air layer between them. A perfect fit is indicated by a single uniform colour, while any difference between the test-glass surface and the surface being polished is shown by the presence of a pattern of concentric coloured bands known as "Newton's Rings." These may be circular if there is only a difference of surface curvature, but they will be looped or irregular if the lens surface is even slightly non spherical. Aspheric surfaces could be tested in this way also, but the difficulty of making an opposite aspheric surface with sufficient accuracy is almost prohibitive.

After assembly the lens is tested by using it to form an image of a point source or other suitable test object. The image is examined through a microscope. If the lens is intended for photography, test photographs are made of an extended array of critical test charts assembled on a flat wall. A careful determination of the resolving power of the lens, in terms of how many lines per millimetre can be distinctly separated in the image, is often used to give a numerical value to the lens performance.

A more recent numerical test is based on the determination of the "modulation transfer function" of the lens at various points in the field. The test object here is a num-

ber of bar charts at progressively smaller spacing. The lens being tested is used to form an image of these charts, and a fine slit with a photoelectric cell behind it is made to traverse the images and measure the contrast; *i.e.*, the relative brightness of the dark bars and the bright spaces between them. The image contrast so determined is plotted on the vertical axis of a graph against the line spacing as horizontal axis. This graph represents the lens performance in a very complete and comprehensive way. Since about 1965 it has been possible to compute the theoretical modulation transfer function curve of a lens from the design data alone, and this has provided an excellent means for ascertaining whether the design is good enough for the particular purpose, and whether the lens has been made exactly according to formula.

Another more sophisticated test is to place the lens in one arm of an interferometer, with a convex spherical mirror behind the lens to send the light rays back through the lens along their original paths. The emerging wave front is then made to interfere with the incident light waves and produce an interference pattern that is a contour map of the wave front itself. Considerable experience is required to enable the operator to interpret the pattern he sees in this kind of test and to correlate the observations with more direct measures of lens performance.

BIBLIOGRAPHY. W.J. SMITH, *Modern Optical Engineering* (1966), a survey of modern theoretical and practical aspects of optical engineering; R. KINGSLAKE (ed.), *Applied Optics and Optical Engineering*, 5 vol. (1965–69), a definitive work, with the first three volumes treating optical devices and theory and the remaining two volumes giving detailed treatment of the principal types of optical instruments; F. TWYMAN, *Prism and Lens Making*, 2nd ed. (1952), a complete and detailed description of processes for making lenses and prisms of every kind; L.C. MARTIN, *Optical Measuring Instruments* (1924), a basic study of the principles involved in design and use of optical measuring instruments, and *Technical Optics*, 2 vol. (1961), an excellent work on the application of the technical aspects of optics and optical theory to optical systems; A.E. CONRADY, *Applied Optics and Optical Design*, 2 vol. (1957–60), a full description of the theory and practice of lens design, providing detailed instructions for the design of many types of optical systems; F.E. WRIGHT, *Manufacture of Optical Glass and of Optical Systems* (1921), a treatment of problems in optical glass manufacture in the U.S. during World War I.

(R.K.)

Optics, Principles of

Originally, the term optics was used only in relation to the eye and vision. Later, as lenses and other devices for aiding vision began to be developed, these were naturally called optical instruments, and the meaning of the term optics eventually became broadened to cover any application of light, even though the ultimate receiver is not the eye but a physical detector, such as a photographic plate or a television camera. Within the present century optical methods have been applied extensively to regions of the electromagnetic radiation spectrum not visible to the eye, such as X-rays, ultraviolet, infrared, and microwave radio waves, and to this extent these regions are now often included in the general field of optics.

In the present article the image-forming properties of lenses, mirrors, and other devices that make use of light will be considered. For a survey of the design of lenses and optical instruments see OPTICAL ENGINEERING. The wave and quantum nature of light, its velocity, wavelength, polarization, diffraction, and interference may be found in LIGHT. The interrelations between light and electricity, which are involved in light sources and detectors, are treated under ELECTRON TUBE; TELEVISION. The analysis of light into its component colours by prisms and gratings forms the basis of the extensive field of spectroscopy (see SPECTROSCOPY, PRINCIPLES OF); whereas the reception of light by the retina of the eye and the interpretation of images by the brain constitute the subjects of the articles EYE AND VISION, HUMAN; PHOTORECEPTION.

An optical image may be regarded as the apparent reproduction of an object by a lens or mirror system, employing light as a carrier. An entire image is generally produced simultaneously, as by the lens in a camera, but images may also be generated sequentially by point-by-point scanning, as in a television system or in the radio transmission of pictures across long distances in space. Nevertheless, the final detector of all images is invariably the human eye, and, whatever means is used to transmit and control the light, the final image must either be produced simultaneously or scanned so rapidly that the observer's persistence of vision will give him the mental impression of a complete image covering a finite field of view. For this to be effective the image must be repeated (as in motion pictures) or scanned (as in television) at least 40 times a second to eliminate flicker or any appearance of intermittency.

The optical image

The content of this article is represented by the following outline:

I. Geometrical optics
  - Historical background
  - Reflection and refraction
  - Ray-tracing methods
  - Paraxial, or first-order, imagery
  - Optical systems
  - Lens aberrations
  - Image brightness

II. Optics and information theory
  - General considerations
  - Image formation
  - Partially coherent light
  - Optical processing
  - Holography
  - Nonlinear optics

I. Geometrical optics

HISTORICAL BACKGROUND

To the ancients, the processes of image formation were full of mystery. Indeed, for a long time there was a great discussion as to whether, in vision, something moved from the object to the eye or whether something reached out from the eye to the object. By the beginning of the 17th century, however, it was known that rays of light travel in straight lines, and in 1604 Johannes Kepler, a German astronomer, published a book on optics in which he postulated that an extended object could be regarded as a multitude of separate points, each point emitting rays of light in all directions. Some of these rays would enter a lens, by which they would be bent around and made to converge to a point, the "image" of the object point whence the rays originated. The lens of the eye was not different from other lenses, and it formed an image of external objects on the retina, producing the sensation of vision.

There are two main types of image to be considered: real and virtual. A real image is formed outside the system, where the emerging rays actually cross; such an image can be caught on a screen or piece of film and is the kind of image formed by a slide projector or in a camera. A virtual image, on the other hand, is formed inside an instrument at the point where diverging rays would cross if they were extended backward into the instrument. Such an image is formed in a microscope or telescope and can be seen by looking into the eyepiece.

Kepler's concept of an image as being formed by the crossing of rays was limited in that it took no account of possible unsharpness caused by aberrations, diffraction, or even defocussing. In 1957 the Italian physicist Vasco Ronchi went the other way and defined an image as any recognizable nonuniformity in the light distribution over a surface such as a screen or film; the sharper the image, the greater the degree of nonuniformity. Today, the concept of an image often departs from Kepler's idea that an extended object can be regarded as innumerable separate points of light, and it is sometimes more convenient to regard an image as being composed of overlapping patterns of varying frequencies and contrasts; hence, the quality of a lens can be expressed by a graph connecting the spatial frequency of a parallel line object with the contrast in the image. This concept is investigated fully under *Optics and information theory* below.

Optics had progressed rapidly by the early years of the

Lens theories

19th century. Lenses of moderately good quality were being made for telescopes and microscopes, and in 1841 the great mathematician, Karl Friedrich Gauss, published his classical book on geometrical optics. In it he expounded the concept of the focal length and cardinal points of a lens system and developed formulas for calculating the position and size of the image formed by a lens of given focal length. Between 1852 and 1856 Gauss's theory was extended to the calculation of the five principal aberrations of a lens (see below), thus laying the foundation for the formal procedures of lens design that were used for the next 100 years. Since about 1960, however, lens design has been almost entirely computerized, and the old methods of designing lenses by hand on a desk calculator are rapidly disappearing.

By the end of the 19th century numerous other workers had entered the field of geometrical optics, notably an English physicist, Lord Rayleigh (John William Strutt), and a German physicist, Ernst Karl Abbe. It is impossible to list all their accomplishments here. Since 1940 there has been a great resurgence in optics on the basis of information and communication theory, which is treated at length below.

Light rays, waves, and wavelets. A single point of light, which may be a point in an extended object, emits light in the form of a continually expanding train of waves, spherical in shape and centred about the point of light. It is, however, often much more convenient to regard an object point as emitting fans of rays, the rays being straight lines everywhere perpendicular to the waves. When the light beam is refracted by a lens or reflected by a mirror, the curvature of the waves is changed, and the angular divergence of the ray bundle is similarly changed in such a way that the rays remain everywhere perpendicular to the waves. When aberrations are present, a convergent ray bundle does not shrink to a perfect point, and the emerging waves are then not truly spherical.

In 1690 Christiaan Huygens, a Dutch scientist, postulated that a light wave progresses because each point in it becomes the centre of a little wavelet travelling outward in all directions at the speed of light, each new wave being merely the envelope of all these expanding wavelets. When the wavelets reach the region outside the outermost rays of the light beam, they destroy each other by mutual interference wherever a crest of one wavelet falls upon a trough of another wavelet. Hence, in effect, no waves or wavelets are allowed to exist outside the geometrical light beam defined by the rays. The normal destruction of one wavelet by another, which serves to restrict the light energy to the region of the rectilinear ray paths, however, breaks down when the light beam strikes an opaque edge, for the edge then cuts off some of the interfering wavelets, allowing others to exist, which diverge slightly into the shadow area. This phenomenon is called diffraction, and it gives rise to a complicated fine structure at the edges of shadows and in optical images (see LIGHT).

The pinhole camera. An excellent example of the working of the wavelet theory is found in the well-known pinhole camera. If the pinhole is large, the diverging geometrical pencil of rays leads to a blurred image, because each point in the object will be projected as a finite circular patch of light on the film. The spreading of the light at the boundary of a large pinhole by diffraction is slight. If the pinhole is made extremely small, however, the geometrical patch then becomes small, but the diffraction spreading is now great, leading once more to a blurred picture. There are thus two opposing effects present, and at the optimum hole size the two effects are just equal. This occurs when the hole diameter is equal to the square root of twice the wavelength (λ) times the distance (f) between the pinhole and film—*i.e.*, $\sqrt{2\lambda f}$. For $f = 100$ millimetres and $\lambda = 0.0005$ millimetre, the optimum hole size becomes 0.32 millimetre. This is not very exact, and a 0.4-millimetre hole would probably be just as good in practice. A pinhole, like a camera lens, can be regarded as having an f-number, which is the ratio of focal length to aperture. In this example, the f-number is 100/0.32=310, designated $f/310$. Modern camera lenses have much greater apertures, in order to achieve light-gathering power, of around $f/1.2$–$f/5.6$

Resolution and the Airy disk. When a well-corrected lens is used in place of a pinhole, the geometrical ray divergence is eliminated by the focussing action of the lens, and a much larger aperture may be employed; in that case the diffraction spreading becomes small indeed. The image of a point formed by a perfect lens is a minute pattern of concentric and progressively fainter rings of light surrounding a central dot, the whole structure being called the Airy disk after George Biddell Airy, an English astronomer, who first explained the phenomenon in 1834. The Airy disk of a practical lens is small, its diameter being approximately equal to the f-number of the lens expressed in microns (0.001 millimetre). The Airy disk of an $f/4.5$ lens is therefore about 0.0045 millimetre in diameter (ten times the wavelength of blue light). Nevertheless, the Airy disk formed by a telescope or microscope objective can be readily seen with a bright point source of light if a sufficiently high eyepiece magnification is used.

Airy disk

The finite size of the Airy disk sets an inevitable limit to the possible resolving power of a visual instrument. Rayleigh found that two adjacent and equally bright stars can just be resolved if the image of one star falls somewhere near the innermost dark ring in the Airy disk of the other star; the resolving power of a lens can therefore be regarded as about half the f-number of the lens expressed in microns. The angular resolution of a telescope is equal to the angle subtended by the least resolvable image separation at the focal length of the objective, the light-gathering lens. This works out at about four and a half seconds of arc divided by the diameter of the objective in inches.

The Rayleigh limit. As noted above, when a perfect lens forms an image of a point source of light, the emerging wave is a sphere centred about the image point. The optical paths from all points on the wave to the image are therefore equal, so that the expanding wavelets are all in phase (vibrating in unison) when they reach the image. In an imperfect lens, however, because of the presence of aberrations, the emerging wave is not a perfect sphere, and the optical paths from the wave to the image point are then not all equal. In such a case some wavelets will reach the image as a peak, some as a trough, and there will be much destructive interference leading to the formation of a sizable patch of light, much different from the minute Airy disk characteristic of a perfectly corrected lens. Lord Rayleigh in 1879 studied the effects of phase inequalities in a star image and came to the conclusion that an image will not be seriously degraded unless the path differences between one part of the wave and another exceed one-quarter of the wavelength of light. As this difference represents only 0.125 micron (5×10^{-6} inch), it is evident that an optical system must be designed and constructed with almost superhuman care if it is to give the best possible definition.

REFLECTION AND REFRACTION

Reflection. The use of polished mirrors for reflecting light has been known for thousands of years, and concave mirrors have long been used to form real images of distant objects. Indeed, Isaac Newton (died 1727) greatly preferred the use of a mirror as a telescope objective to the poor-quality lenses available in his time. Because there is no limit to the possible size of a mirror, all large telescopes today are of this type.

When a ray of light is reflected at a polished surface, the angle of reflection between ray and normal (the line at right angles to the surface) is exactly equal to the angle of incidence. It can be seen that a convex mirror forms a virtual image of a distant object, whereas a concave mirror forms a real image. A plane mirror forms a virtual image of near objects, as in the familiar looking glass. Plane mirrors are often used in instruments to bend a beam of light into a different direction.

Law of reflection

The law of refraction. When a ray of light meets the surface of separation between two transparent media, it is sharply bent or refracted. Because rays are really only directions and have no physical existence, the passage of

light waves through a surface must be considered if refraction is to be understood. Refraction effects are based on the fact that light travels more slowly in a denser medium. The ratio of the velocity of light in air to its velocity in the medium is called the refractive index of the medium for light of a particular colour or wavelength. The refractive index is higher for blue light than for light at the red end of the spectrum.

In Figure 1, AA' represents a plane wave of light at the instant that A' meets the plane refracting surface $A'B$ separating two media having refractive indices n and n',

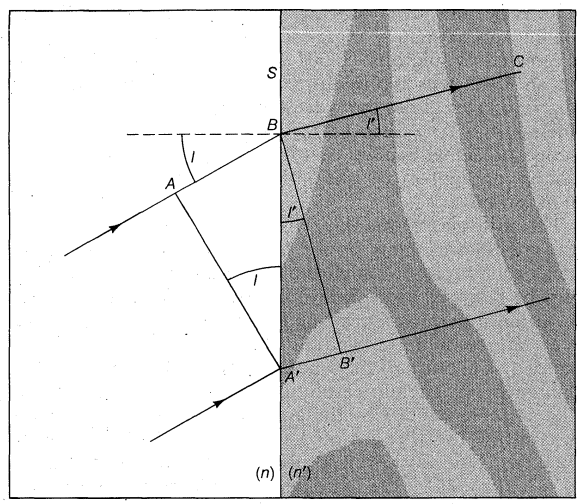


Figure 1: The law of refraction. Plane light wave at position AA' in medium of index n and BB' in medium of index n' (see text).

respectively. During the time taken by the light to travel from A to B in material n, light travels from A' to B' in material of refractive index n', forming the new wave BB' in the second material, proceeding in direction BC. Hence, the relationship $n'/n = AB/A'B'$ can be obtained; and dividing numerator and denominator by BA' gives

$$\frac{n'}{n} = \frac{AB/BA'}{A'B'/BA'} = \frac{\sin I}{\sin I'} \quad . \qquad (1)$$

The angles I and I' are called the angle of incidence and angle of refraction between the refracting surface and the incident and refracted waves, respectively.

Returning now to the convention of considering the movement of light in terms of rays because entering and emerging rays are always perpendicular to the light waves they represent, angles I and I' also denote the angles between the entering and emerging rays and the normal (perpendicular) line to the refracting surface at B.

Equation (1), known as the law of refraction, is generally written: $n' \sin I' = n \sin I$.

Dispersion. The difference between the refractive indices of a transparent material for a specific blue light and a specific red light is known as the dispersion of the material. The usual choices of blue and red lights are the so-called "F" and "C" lines of hydrogen in the solar spectrum, named by Fraunhofer, with wavelengths 4861 and 6563 angstroms (the angstrom unit, abbreviated Å, is 10^{-8} centimetre), respectively. It is generally more significant, however, to compare the dispersion with the mean refractive index of the material for some intermediate colour such as the sodium "D" Fraunhofer line of wavelength 5893 angstroms. The dispersive power (w) of the material is then defined as the ratio of the difference between the "F" and "C" indices and the "D" index reduced by 1, or,

Dispersive power

$$w = \frac{n_F - n_C}{n_D - 1} \quad .$$

Hundreds of different types of optical glass are currently available from manufacturers. These may be represented graphically on a plot of mean refractive index against dispersive power (Figure 2).

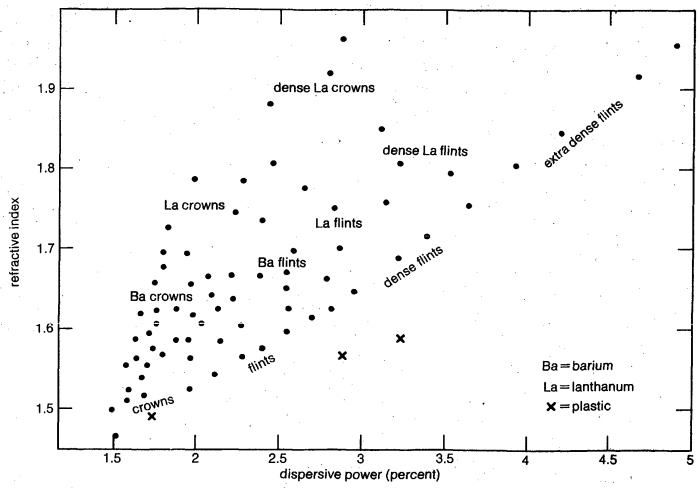


Figure 2: Some representative optical glasses and plastics, showing their refractive indices versus dispersive powers (see text).

Glasses of low dispersive power (less than 0.02) are commonly called crowns, while those of higher dispersion are called flints. These names are a relic of the early types of glass that were manufactured for windows and fine table glassware, respectively. The wide range of types now available is a tremendous help to the lens designer, although his choice is sometimes limited by the high cost of the more extreme types.

Since World War II there has been increasing use of plastic materials in the manufacture of lenses. These substances are especially advantageous for large single lenses in which glass would be excessively heavy, for spectacles when weight and brittleness are important, and since 1957 for small camera lenses in which the demands of high precision combined with low manufacturing cost can be readily met by injection molding.

Total internal reflection. When a ray of light emerges obliquely from glass into air, the angle of refraction between ray and normal is greater than the angle of incidence inside the glass, and at a sufficiently high obliquity the angle of refraction can actually reach 90°. In this case the emerging ray travels along the glass surface, and the sine of the angle of incidence inside the glass, known as the critical angle, is then equal to the reciprocal of the refractive index of the material. At angles of incidence greater than the critical angle, the ray never emerges, and total internal reflection occurs, for there is no measurable loss if the glass surface is perfectly clean. Dirt or dust on the surface can cause a small loss of energy by scattering some light into the air.

Light is totally internally reflected in many types of reflecting prism (see OPTICAL ENGINEERING) and in recently developed fibre optics technology. Long fibres of high-index glass clad with a thin layer of lower index glass are assembled side-by-side in precise order, and light admitted into one end of each fibre is transmitted along it without loss by thousands of successive internal reflections at the interlayer between the glass and the cladding. Hence, an image projected upon one end of the bundle will be dissected and transmitted to the other end, where it can be examined through a magnifier or photographed. Many modern medical instruments, such as cystoscopes and bronchoscopes, depend for their action on this principle. Single thick fibres (actually glass rods) are sometimes used to transmit light around corners to an otherwise inaccessible location.

RAY-TRACING METHODS

Graphical ray-tracing. In 1621 Willebrord Snell, a professor of mathematics at Leiden, discovered a simple graphical procedure for determining the direction of the refracted ray at a surface when the incident ray is given. The mathematical form of the law of refraction, equation (1) above, was announced by the French mathematician René Descartes some 16 years later.

Snell's construction

Snell's construction is as follows: The line AP in Figure 3A represents a ray incident upon a refracting surface at P, the normal at P being PN. If the incident and refracted rays are extended to intersect any line SS parallel to the normal, the lengths PQ and PR along the rays will be

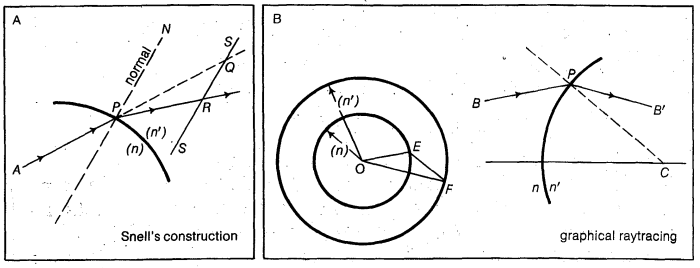


Figure 3: Graphic refraction procedures (see text).

proportional to the refractive indices n and n'. Hence, if PQ and the indices are known, PR can be found and the refracted ray drawn in.

A convenient modification of Snell's construction can readily be used to trace the path of a ray through a complete lens. In Figure 3B, the incident ray BP strikes a refracting surface at P. The normal to the surface is PC. At any convenient place on the page two concentric circles are drawn about a point O with radii proportional to the refractive indices n and n', respectively. A line OE is now drawn parallel to the incident ray BP extending as far as the circle representing the refractive index n of the medium containing the incident ray. From E a line is drawn parallel to the normal PC extending to F on the circle representing the refractive index n'. The line OF then represents the direction of the desired refracted ray, which may be drawn in at PB'. This process is repeated successively for all the surfaces in a lens. If a mirror is involved, the reflected ray may be found by drawing the normal line EF across the circle diagram to the incident-index circle on the other side.

Trigonometrical ray-tracing. No graphical construction can possibly be adequate to determine the aberration residual of a corrected lens, and for this an accurate trigonometrical computation must be made and carried out to six or seven decimal places, the angles being determined to single seconds of arc or less. There are many procedures for calculating the path of a ray through a system of spherical refracting or reflecting surfaces, the following being typical: The diagram in Figure 4 repre-

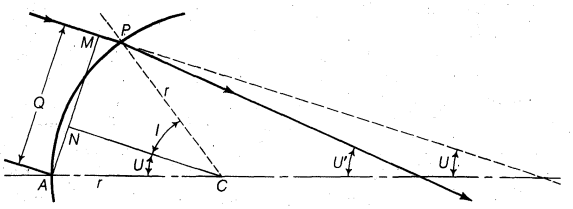


Figure 4: Trigonometrical ray-tracing (see text).

sents a ray lying in the meridian plane, defined as the plane containing the lens axis and the object point. A ray in this plane is defined by its slope angle, U, and by the length of the perpendicular, Q, drawn from the vertex (A) of the surface on to the ray. By drawing a line parallel to the incident ray through the centre of curvature C, to divide Q into two parts at N, the relation is stated as $AN = r \sin U$, and $NM = r \sin I$. Hence

$$Q = r(\sin U + \sin I). \tag{2}$$

From this the first ray-tracing equation can be derived,

$$\sin I = \frac{Q}{r} - \sin U. \tag{3a}$$

Applying the law of refraction, equation (2), gives the second equation

$$\sin I' = \frac{n}{n'} \sin I. \tag{3b}$$

Because the angle $PCA = U + I = U' + I'$, the slope of the refracted ray can be written as

$$U' = U + I - I'; \tag{3c}$$

and, lastly, by adding primes to equation (2),

$$Q' = r(\sin U' + \sin I').$$

Having found the Q' of the refracted ray, transfer to the next surface can be performed by

$$Q_2 = Q_1' - d \sin U_1',$$

in which d is the axial distance from the first to the second refracting surface. After performing this calculation for all the surfaces in succession, the longitudinal distance from the last surface to the intersection point of the emergent ray with the lens axis is found by

$$L' = \frac{Q'}{\sin U'}.$$

Corresponding but much more complicated formulas are available for tracing a skew ray, that is, a ray that does not lie in the meridian plane but travels at an angle to it. After refraction at a surface, a skew ray intersects the meridian plane again at what is called the diapoint. By tracing the paths of a great many (100 or more) meridional and skew rays through a lens, with the help of an electronic computer, and plotting the assemblage of points at which all these rays pierce the focal plane after emerging from the lens, a close approximation to the appearance of a star image can be constructed, and a good idea of the expected performance of a lens can be obtained.

PARAXIAL, OR FIRST-ORDER, IMAGERY

Paraxial rays

In a lens that has spherical aberration, the various rays from an axial object point will in general intersect the lens axis at different points after emerging into the image space. By tracing several rays entering the lens at different heights (*i.e.*, distances from the axis) and extrapolating from a graph connecting ray height with image position, it would be possible to infer where a ray running very close to the axis (a paraxial ray) would intersect the axis, although such a ray could not be traced directly by the ordinary trigonometrical formulas because the angles would be too small for the sine table to be of any use. Because the sine of a small angle is equal to the radian measure of the angle itself, however, a paraxial ray can be traced by reducing the ray-tracing formulas to their limiting case for small angles and thus determining the paraxial intersection point directly. When this is done, writing paraxial-ray data with lowercase letters, it is found that the Q and Q' above both become equal to the height of incidence y, and the formulas (3a), (3b), and (3c) become, in the paraxial limit:

$$i = \frac{y}{r} - u \tag{4a}$$

$$i' = \frac{n}{n'} i \tag{4b}$$

$$u' = u + i - i'. \tag{4c}$$

The longitudinal distance from the last surface to the intersection point of the emerging paraxial ray with the lens axis becomes $l' = y/u'$.

Because all paraxial rays from a given object point unite at the same image point, the resulting longitudinal distance (l') is independent of the particular paraxial ray that is traced. Any nominal value for the height of incidence, y, may therefore be adopted, remembering that it is really an infinitesimal and y is only its relative magnitude. Thus, it is clear that the paraxial angles in equation (4) are really only auxiliaries, and they can be readily eliminated, giving the object–image distances for paraxial rays:

$$n'(l' - r)u' = n(l - r)u \tag{5}$$

and

$$\frac{n'}{l'} = \frac{n}{l} + \frac{n' - n}{r}. \tag{6}$$

Magnification: the optical invariant. It is frequently as important to determine the size of an image as it is to

determine its location. To obtain an expression for the magnification—that is, the ratio of the size of an image to the size of the object—the following process may be used: If an object point B lies to one side of the lens axis at a transverse distance h from it, and the image point B' is at a transverse distance h', then B, B', and the centre of curvature of the surface, C, lie on a straight line called the auxiliary axis. Then, by simple proportion,

$$m=\frac{h'}{h}=\frac{l'-r}{l-r}=\frac{nu}{n'u'}.$$

Hence,

$$h'n'u' = hnu, \qquad (7)$$

Lagrange theorem

and the product (hnu) is invariant for all the spaces between the lens surfaces, including the object and image spaces, for any lens system of any degree of complexity. This theorem has been named after the French scientist Joseph-Louis Lagrange, although it is sometimes called the Smith–Helmholtz theorem, after Robert Smith, an English scientist, and Hermann Helmholtz, a German scientist; the product (hnu) is often known as the optical invariant. As it is easy to determine the quantities h, n, and u for the original object, it is only necessary to calculate u' by tracing a paraxial ray in order to find the image height h' for any lens. If the lens is used in air, as most lenses are, the refractive indices are both unity, and the magnification becomes merely $m = u/u'$.

The Gauss theory of lenses. In 1841 Gauss published a now famous treatise on optics in which he demonstrated that, so far as paraxial rays are concerned, a lens of any degree of complexity can be replaced by two principal, or nodal, points and two focal points, the distances from the principal points to their respective focal points being the focal lengths of the lens, and, furthermore, that the two focal lengths are equal to one another when the refractive indices of object and image spaces are equal, as when a lens is used in air.

The principal and focal points may be defined as follows: Figure 5 shows a lens system of any construction, with a bundle of rays entering from the left in a direction

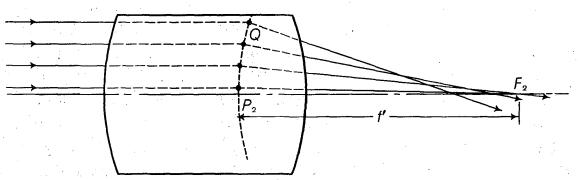


Figure 5: The Gauss theory (see text).

parallel to the lens axis. After refraction by the lens each ray will cross the axis at some point, and the entering and emerging portions of each ray are then extended until they intersect at a point such as Q. The locus of all the points Q is a surface of revolution about the lens axis known as the equivalent refracting locus of the lens. The point where this locus crosses the axis is called the principal point, P_2, and the central portion of the locus in the neighbourhood of the axis, which is virtually a plane perpendicular to the axis, is called the principal plane. The point where the emerging paraxial ray crosses the axis is called the focal point F_2, the distance from P_2 to F_2 being the (posterior) focal length f'. A similar situation exists for a parallel beam of light entering from the right, giving the anterior principal point P_1, the anterior focal point F_1, and the front focal length f. For a lens in air it can be shown that the two focal lengths are equal in magnitude but opposite in direction—*i.e.*, if F_2 is to the right of P_2, then F_1 must lie to the left of P_1, as in the case of an ordinary positive lens (one that gives a real image). In a negative lens (one that gives a virtual image), F_2 lies to the left of P_2, and the posterior focal length f' is negative.

The relation between the distances of object and image from a lens can be easily stated if the positions of the two principal points and the two focal points are known. (In using these expressions, distances are considered positive or negative depending on whether they are measured to the right or to the left from their respective origins.) For a lens in air: (a) If the conjugate distances measured from the respective focal points are x and x', and if m is the image magnification (height of image divided by height of object), then $m = -x'/f' = f'/x$ and $xx' = -f'^2$. (b) If the conjugate distances measured from the respective principal points are p and p' and if m is the image magnification, then $m = p'/p$ and $1/p' = 1/p + 1/f'$. The Lagrange equation (7) requires modification for a distant object because in that case the object height h is infinite, and the slope angle u is zero. If the off-axis distance h is divided by the object distance L, and u is multiplied by L, equation (7) becomes $h' = (n/n')f'\phi$, in which ϕ is the angle in radians subtended by the distant object at the lens. This formula provides a means for defining focal length and for measuring the focal length of an unknown lens.

Determination of focal length

The thin lens. In a thin lens such as a spectacle, the two principal planes coincide within the lens, and then the conjugate distances p and p' in the formula above become the distances of object and image from the lens itself.

The focal length of a thin lens can be computed by applying the surface-conjugate formula (6) to the two surfaces in succession, writing the l of the first surface as infinity and the l of the second surface equal to the l' of the first surface. When this is done, the lens power (P) becomes

$$P=\frac{1}{f'}=(n-1)\left(\frac{1}{r_1}-\frac{1}{r_2}\right).$$

If a number of thin lenses are assembled together in contact, assuming that the combination is still virtually a thin lens, then the power $(1/f')$ of the combination will be equal to the sum of the powers of the separate lenses. Thus, if a thin negative lens is placed in close contact with a thin positive lens of equal power, the combination will have zero power and will no longer act like a lens. This is the basis of a process used by optometrists to ascertain the power of an unknown lens by combining it with one of a series of known lenses from a trial case.

Chromatic aberration. Because the refractive index of glass varies with wavelength, every property of a lens that depends on its refractive index also varies with wavelength, including the focal length, the image distance, and the image magnification. The change of image distance with wavelength is known as chromatic aberration, and the variation of magnification with wavelength is known as chromatic difference of magnification, or lateral colour. Chromatic aberration can be eliminated by combining a strong lens of low-dispersion glass (crown) with a weaker lens made of high-dispersion (flint) glass. Such a combination is said to be achromatic. This method of removing chromatic aberration was discovered in 1729 by Chester Hall, an English inventor, and it was exploited vigorously in the late 18th century in a series of small telescopes. Chromatic variation of magnification can be eliminated by achromatizing all the components of a system or by making the system symmetrical about a central diaphragm. Both chromatic aberration and lateral colour are corrected in every high-grade optical system.

Longitudinal magnification. If an object is moved through a short distance δp along the axis, then the corresponding image shift $\delta p'$ is related to the object movement by the longitudinal magnification $(\overline{m})$. Succinctly,

$$\overline{m} = \delta p'/\delta p = m^2,$$

in which m is the lateral magnification. The fact that the longitudinal magnification is equal to the square of the transverse magnification means that m is always positive; hence, if the object is moved from left to right, the image must also move from left to right. Also, if m is large, then $\overline{m}$ is very large, which explains why the depth of field (δp) of a microscope is extremely small. On the other hand, if m is small, less than one as in a camera, then $\overline{m}$ is very small, and all objects within a considerable range of distances (δp) appear substantially in focus.

Limitation on depth of field

Image of a tilted plane. If a lens is used to form an image of a plane object that is tilted relative to the lens

axis, then the image will also be tilted in such a way that the plane of the object, the plane of the image, and the median plane of the lens all meet. This construction can be derived by the use of the lateral and longitudinal magnification relations just established above. With a tilted object the magnification at any point is given by the ratio of the distances of image and object from the lens at that point in the image, and, consequently, m varies progressively from one end of the image to the other. This arrangement is frequently used in view cameras equipped with "swings" to increase depth of field and in enlargers to rectify the convergence of parallel lines caused by tilting the camera, for example, in photographing tall buildings. The rule finds extensive application in photogrammetry and in the making of maps from aerial photographs.

OPTICAL SYSTEMS

System components. An optical system consists of a succession of elements, which may be lenses, mirrors, prisms, fibre bundles, etc., with, possibly, a light source or some form of detector or electronic image amplifier, depending on the requirements. All optical systems have an aperture stop somewhere in the system to limit the diameter of the beams of light passing through the system from an object point.

Entrance and exit pupils

By analogy with the human eye, this limiting aperture stop is called the iris of the system, its images in the object and image spaces being called the entrance pupil and exit pupil, respectively. In most photographic lenses the iris is inside the objective, and it is often adjustable in diameter to control the image illumination and the depth of field. In telescope and microscope systems the cylindrical mount of the objective lens is generally the limiting aperture or iris of the system; its image, formed behind the eyepiece where the observer's eye must be located to see the whole area being observed, called the field, is then the exit pupil.

The pupils of a lens system can be regarded as the common bases of oblique beams passing through the system from all points in an extended object. In most systems, however, the mounts of some of the lens elements cut into the oblique beams and prevent the beams from being perfectly circular, and the pupils are then not fully filled with light. This effect is known as vignetting and leads to a reduction in illumination in the outer parts of the field of view.

A common feature of many optical systems is a relay lens, which may be introduced to invert an image or to extend the length of the system, as in a military periscope. An example of the use of a relay lens is found in the common rifle sight shown diagrammatically in Figure 6.

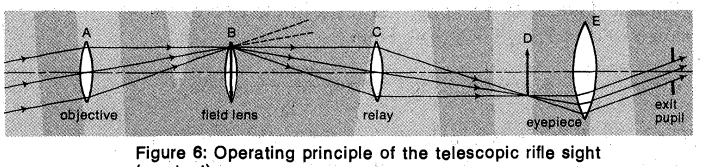


Figure 6: Operating principle of the telescopic rifle sight (see text).

Here the front lens A is the objective, forming an inverted image of the target on the cross wire or reticle at B. The light then proceeds to the relay lens C, which forms a second image, now erect, at D. Beyond this image is the eyepiece E to render the light parallel so that the image may be seen sharply by the observer. Unfortunately, the oblique beam from the objective will usually miss the relay lens, and so a field lens must be inserted at or near the first image B to bend the oblique beams around and redirect them toward the relay lens. The power of the field lens is chosen so that it will form an image of the objective lens aperture on the relay lens aperture. The iris and entrance pupil of this system coincide at the objective; there is an internal pupil at the relay lens, and the exit pupil lies beyond the eyepiece as shown in Figure 6.

Another typical optical system is used in the familiar slide projector, in which a concentrated filament lamp is imaged into the projection lens aperture by a condenser lens. The slide or film transparency to be projected is placed close to the condenser, where there is a broad area of uniform illumination, and it is in turn imaged on the screen by the projection lens.

In principle a concave mirror could be substituted for an image-forming lens in most optical systems. The aberrations of a mirror are generally less than those of a comparable lens, but the fact that a mirror forms its image within the confines of the entering beam leads to many difficulties, and mirrors are therefore used only where the large size or weight of a lens makes the use of a mirror mandatory.

Visual systems. A great many optical instruments—notably spectacles, magnifiers, telescopes, and microscopes—are intended for use in front of an observer's eye to aid vision by providing him with a sharper or larger image. The eye consists of a lens system forming a real image of outside objects on the retina, where the image is converted into nerve impulses that go to the brain and produce the sensation of vision (see EYE AND VISION, HUMAN). In a young person the power of the eye lens can be varied by unconscious muscular effort, enabling him to focus on objects from as close as ten centimetres or so to those that are substantially at infinity. This ability is known as accommodation; the range of accommodation dwindles with age, in humans finally disappearing completely at about age 75.

Accommodation

The larger the retinal image, the larger will an object appear to be, and, as an object is brought closer to the eye, its apparent size increases. The simple magnifier is a positive lens used in front of the eye to enable bringing an object closer than the nearest distance of distinct vision determined by the range of accommodation. If the object is held at the anterior focus of a magnifier, it will appear to be at infinity, and it can be focussed easily. The apparent size of the object will now be equal to the ratio of the object size to the focal length of the lens, whereas without the magnifier the apparent size was equal to the object size divided by the nearest distance of distinct vision, usually assumed to be 25 centimetres (ten inches). The magnifying power of the lens is, therefore, 25 centimetres divided by the focal length. A ten-centimetre lens will give 2.5 times magnification, and so on.

For magnifying powers exceeding about ten, a simple magnifying lens becomes impractical, and a microscope must be used instead. Here the objective lens forms a magnified aerial image of the object, which is further magnified by the eyepiece acting as a simple magnifier. Thus, the objective may magnify an object ten times, and a 2.5-centimetre eyepiece will magnify it a further ten times, giving an overall magnifying power of 100 times.

To magnify a distant object, a telescope is used. The objective lens forms a real image of the object, which is then magnified by the eyepiece. Because the object is inaccessible, the magnifying power of a telescope must be defined in a different way, namely the ratio of the angular subtense of the image to the angular subtense of the object. This turns out to be equal to the ratio of the focal length of the objective lens to that of the eyepiece, multiplied by the magnification of the relay lens, if there is one. Without a relay or other means for erecting the image, a simple telescope will provide an inverted image; this is avoided in the common opera glass by the use of a negative lens for the eyepiece. This so-called Galilean telescope covers only a narrow field and can be given only a low magnifying power. The most familiar small telescope is now the prism binocular in which the image is erected by a pair of prisms mounted between the objective and the eyepiece.

Nonclassical imaging systems. Besides the familiar optical systems so far considered, there are many nonclassical optical elements that are used to a limited extent for special purposes. The most familiar of these is the aspheric (nonspherical) surface. Because plane and spherical surfaces are the easiest to generate accurately on glass, most lenses contain only such surfaces. It is occasionally necessary, however, to use some other axially symmetric surface on a lens or mirror, generally to

Aspheric surfaces

correct a particular aberration. An example is the parabolic surface used for the primary mirror of a large astronomical telescope; another is the elliptic surface molded on the front of the little solid glass reflector units used on highway signs.

Another commonly used optical surface is the side of a cylinder. Such surfaces have power only in the meridian perpendicular to the cylinder axis. Cylindrical lenses are therefore used wherever it is desired to vary the magnification from one meridian to a perpendicular meridian. Cylindrical surfaces are employed in the anamorphic lenses used in some wide-screen motion-picture systems to compress the image horizontally in the camera and stretch it back to its original shape in the projected image.

To correct astigmatism in the eye, many spectacles are made with toric surfaces—*i.e.*, with a stronger curvature in one meridian than in the perpendicular meridian, like the bowl of a teaspoon. These surfaces are generated and polished by special machines and are made by the million every year.

Another nonclassical optical system is the bifocal or trifocal spectacle lens. They are made either by forming two or three separate surfaces on a single piece of glass or obtaining additional power by fusing a piece of high-index glass on to the front of the main lens and then polishing a single spherical surface over both glasses.

Two French scientists, Georges-Louis Buffon and Augustin-Jean Fresnel, in the 18th century suggested forming a lens in concentric rings to save weight, each ring being a portion of what would normally be a continuous spherical surface but flattened out. On a large scale, Fresnel lenses have been used in lighthouses, floodlights, and traffic signals, and as cylindrical ship's lanterns. With fine steps a few thousandths of an inch wide, molded plastic Fresnel lenses are often used as condensers in overhead projectors and in cameras as a field lens in contact with a ground-glass viewing screen.

Lenses have occasionally been made with one surface in the form of a flattened cone. Such lenses produce a long, linear image of a point source, lying along the lens axis; for this reason they are commonly referred to as axicons. They have been used to produce a straight line in space for aligning machines and shafting, but since about 1965 the beam from a gas laser has generally been used instead.

Axicons

LENS ABERRATIONS

Seidel sums. If a lens were perfect and the object were a single point of monochromatic light, then, as noted above, the light wave emerging from the lens would be a portion of a sphere centred about the ideal image point, lying in the paraxial image plane at a height above the axis given by the Lagrange theorem. In practice, however, this condition is most unlikely to occur; it is much more probable that the emerging wave will depart slightly from a perfect sphere, the departure varying from point to point over the lens aperture. This departure is extremely small, being of the order of the wavelength of light that is only half a micron, so it would be impossible to show this departure on a drawing. It can be represented mathematically, however, in the following way: The coordinates of a point in the exit-pupil aperture will be represented by x_0 and y_0, the y_0 coordinate lying in the meridian plane containing the object point and the lens axis. The departure of the wave from the ideal sphere is generally called OPD, meaning optical path difference. It can be shown that OPD is related to x_0 and y_0 by five constants S_1 through S_5, and the quantity h'_0,

$$\text{OPD} = S_1(x_0^2 + y_0^2)^2 + S_2 y_0(x_0^2 + y_0^2)h'_0 + \\ + S_3(x_0^2 + 3y_0^2)h'^2_0 + S_4(x_0^2 + y_0^2)h'^2_0 + S_5 y_0 h'^3_0.$$

Each of these five terms is considered to be a separate "aberration," the coefficients $S_1, \ldots, S_5$, being referred to as Seidel sums. These aberrations are respectively spherical, coma, astigmatism, Petzval field curvature, and distortion. The symbol h'_0 refers to the height of the final image point above the lens axis, and hence it defines the obliquity of the beam.

The five Seidel sums can be calculated by tracing a paraxial ray from object to image through the lens and by tracing also a paraxial principal ray from the centre of the aperture stop outward in both directions toward the object and image, respectively. The angle of incidence i and the ray slope angle u of each of these paraxial rays at each surface are then listed and inserted into the following expressions for the five sums. The angle u_0' represents the final emerging slope of the paraxial ray.

Calculation of Seidel sums

The calculation starts by determining the radius A of the exit pupil by $A = \sqrt{x_0^2 + y_0^2}$ and also the quantity K at each surface by

$$K = \tfrac{1}{2} y n \left(\frac{n}{n'} - 1\right)(i - u').$$

The corresponding K_{pr} for the paraxial principal ray is also determined at each surface. Then, the five aberrations may be written

$$S_1 = \frac{1}{4A^4}\sum Ki^2 \qquad S_2 = \frac{1}{A^3 h'_0}\sum Kii_{pr}$$

$$S_3 = \frac{1}{2A^2 h'^2_0}\sum Ki_{pr}^2 \qquad S_4 = \frac{u'^2_0}{4A^2}\sum \frac{n - n'}{nn'r}$$

$$S_5 = \frac{1}{Ah'^3_0}\sum\left[K_{pr}ii_{pr} + \tfrac{1}{2}h'_0 u'_0(u'^2_{pr} - u_{pr}^2)\right].$$

To interpret these aberrations, the simplest procedure is to find the components x', y' of the displacement of a ray from the Lagrangian image point in the paraxial focal plane, by differentiating the OPD expression given above. The partial derivatives $\partial\text{OPD}/\partial x_0$ and $\partial\text{OPD}/\partial y_0$ represent respectively the components of the slope of the wave relative to the reference sphere at any particular point (x_0, y_0). Hence, because a ray is always perpendicular to the wave, the ray displacements in the focal plane can be found by

$$x' = f\frac{\partial\text{OPD}}{\partial x_0} \quad \text{and} \quad y' = f\frac{\partial\text{OPD}}{\partial y_0},$$

in which f is the focal length of the lens. The aggregation of rays striking the focal plane will indicate the kind of image that is characteristic of each aberration.

This procedure will be applied to each of the five aberration terms separately, assuming that all the other aberrations are absent. Obviously, in a perfect lens x' and y' are zero because OPD is zero. It must be remembered, however, that by using rays instead of waves, all fine-structure effects caused by diffraction will be lost, and only the macroscopic image structure will be retained.

Spherical aberration. The first term in the OPD expression is $\text{OPD} = S_1(x_0^2 + y_0^2)^2$. Hence

$$x' = f\frac{\partial\text{OPD}}{\partial x_0} = 4fA^2 S_1 \cdot x_0 \quad \text{and} \quad y' = f\frac{\partial\text{OPD}}{\partial y_0} = 4fA^2 S_1 \cdot y_0.$$

These displacements can both be eliminated simultaneously by applying a longitudinal shift L to the focal plane. This changes x' by $-Lx_0/f$ and y' by $-Ly_0/f$; hence, if L is made equal to $4f^2A^2S_1$, both ray displacements vanish. The aberration, therefore, represents a condition in which each zone of the lens has a different focus along the axis, the shift of focus from the paraxial image being proportional to A^2. This is known as spherical aberration (see Figure 7).

Coma. The S_2 term in the OPD expression represents the aberration called coma, in which the image of a point has the appearance of a comet. The x' and y' components are as follows:

$$x' = fh'_0 S_2(2x_0 y_0)$$

$$y' = fh'_0 S_2(x_0^2 + 3y_0^2).$$

When this aberration is present, each circular zone of the lens forms a small ringlike image in the focal plane, the rings formed by successive concentric zones of the lens

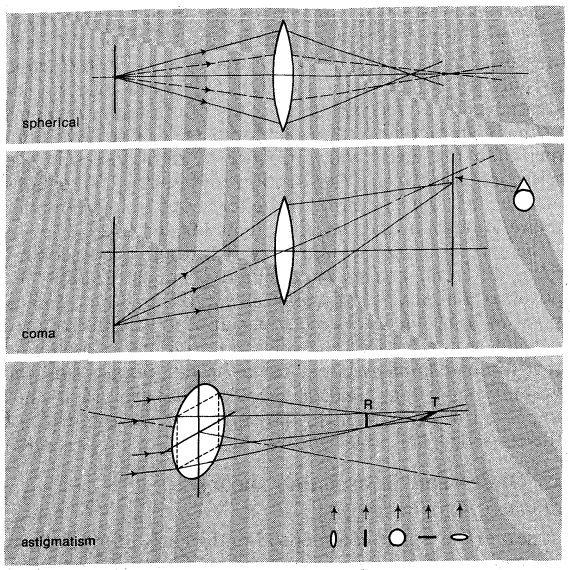


Figure 7: Lens aberrations.

fitting into two straight envelope lines at 60° to each other (Figure 7). Because the brightness of this image is greatest at the tip, coma tends to form a one-sided haze on images in the outer parts of the field.

Astigmatism. If only the S_3 term is present, then

$$x' = 2fh_0'^2 S_3 (x_0)$$

$$y' = 2fh_0'^2 S_3 (3y_0).$$

For any one zone of the lens, x' and y' describe a vertical ellipse with major axis three times the minor axis. The images formed by all the smaller zones of the lens fit into this ellipse and fill it out with a uniform intensity of light. If the image plane is moved along the axis by a distance L, as in focussing a camera, then, at $L = 2f^2h_0'^2S_3$, the ellipse shrinks to a radial focal line (R). Twice this displacement yields a circle; three times this L gives a tangential focal line (T), which is followed by an ellipse with its major axis in the x direction, as in Figure 7, bottom. The usual effect of astigmatism in an image is the appearance of radial or tangential blurring in the outer parts of the field.

Petzval curvature. For the S_4 term taken alone,

$$x' = 2fh_0'^2 S_4 \cdot x_0$$

$$y' = 2fh_0'^2 S_4 \cdot y_0.$$

The image of a point is now a small circle that contracts to a point at a new focus situated at a longitudinal distance $L = 2f^2h_0'^2S_4$ from the paraxial image. As the longitudinal displacement of the focus is proportional to the square of the image height h_0', this aberration represents a pure field curvature without any accompanying loss of definition (all lines remain sharp). It is named after the Hungarian mathematician József Petzval, who studied its properties in the early 1840s. The effect of Petzval curvature can be somewhat offset by the deliberate introduction of sufficient overcorrected astigmatism, as was done in all the pre-anastigmat photographic objectives. This added astigmatism is, of course, undesirable, and in order to design an anastigmat lens having a flat field free from astigmatism, it is necessary to reduce the Petzval sum S_4 drastically.

For a succession of thin lenses (1, 2, 3, . . . etc.) in a system, the Petzval sum becomes simply $1/f_1n_1 + 1/f_2n_2 + 1/f_3n_3 + \ldots$ etc., in which f is the focal length of each element and n is its refractive index. Therefore, to reduce the sum and minimize this aberration, relatively strong negative elements of low-index glass can be combined with positive elements of high-index glass. The positive and negative elements must be axially separated to provide the lens with a useful amount of positive power. The introduction of high-index barium crown glass with a low dispersive power in the 1880s initiated the development of anastigmat lenses.

Distortion. For the S_5 aberration,

$$x' = 0$$

$$y' = fh_0'^3 S_5.$$

When this aberration is present, the entire image point is displaced toward or away from the axis by an amount proportional to the third power of the transverse distance h_0' of the image from the axis. This leads to the formation of an image of a square that is either a barrel-shaped or a cushion-shaped figure.

Other aberrations

It is to be noted that the five Seidel aberrations represent the largest and most conspicuous defects that can arise in an uncorrected optical system. Even in the best lenses in which these five aberrations have been perfectly corrected for one zone of the lens and for one point in the field, however, there will exist small residuals of these aberrations and of many other higher order aberrations, also, which are significantly different from the classical types just described. The typical aberration figures shown in Figure 7 are, of course, grossly exaggerated, and actually it requires some magnification of a star image to render these appearances clearly visible. Nevertheless, they are important enough to require drastic reduction in high-quality lenses intended to make sharp negatives capable of considerable enlargement.

IMAGE BRIGHTNESS

General relations. All photometric concepts are based on the idea of a standard candle, lamps having accurately known candle power being obtainable from the various national standards laboratories. The ratio of the candle power of a source to its area is called the luminance of the source; luminances range from about 2,000 candles per square millimetre at the surface of the Sun down to about 3×10^{-6} candle per square centimetre (3×10^{-6} stilb) for the luminous paint on a watch dial. Ordinary outdoor scenes in daylight have an average luminance of several hundred candles per square foot. The quantity of light flux flowing out from a source is measured in lumens, the lumen being defined as the amount of flux radiated by a small "point" source of one candle power into a cone having a solid angle of one steradian. When light falls upon a surface it produces illumination (*i.e.*, illuminance), the usual measure of illuminance being the foot-candle, which is one lumen falling on each square foot of receiving surface.

It is often important to be able to calculate the brightness of an image formed by an optical system, because photographic emulsions and other light receptors cannot respond satisfactorily if the light level is too low. The problem is to relate the luminance of an object with the illuminance in the image, knowing the transmittance and aperture of the optical system. A small area A of a plane object having a luminance of B candles per square unit will have a normal intensity of AB candles. This source radiates light into a cone of semi-angle U, limited, for example, by the rim of a lens. The light flux (F) entering the cone can be found by integration to be

$$F = \pi AB \sin^2 U \text{ lumens.}$$

If the object luminance is expressed as B_L lamberts, the lambert being an alternative luminance unit equal to $1/\pi$ (*i.e.*, 0.32) candle per unit area, the flux (F) is

$$F = AB_L \sin^2 U \text{ lumens,}$$

because there are π times as many lamberts in a given luminance as there are candles per unit area.

A fraction t of this flux finds its way to the image, t being the lens transmittance, generally about 0.8 or 0.9 but less if a mirror is involved. The area of the image is Am^2, in which m, the magnification, is given by

$$m = \frac{\sin U}{\sin U'}.$$

Hence, the image illuminance (E) is

$$E = (t\pi AB \sin^2 U) \div (A \sin^2 U / \sin^2 U') = t\pi B \sin^2 U' \quad (8)$$

or

$$E = tB_L \sin^2 U' .$$

The image illuminance thus depends *only* on the luminance of the source and the cone angle of the beam proceeding from the lens to the image. This is a basic and most important relation underlying all calculations of image illuminance.

It is often more convenient to convert the angle U' into other better known quantities, such as the f-number of the lens and the image magnification. The relation here is

$$\sin U' = \frac{1}{2(f\text{-number})(1 + m/m_p)} . \quad (9)$$

Definition of f-number

The f-number of the lens is defined as the ratio of the focal length to the diameter of the entrance pupil; m is the image magnification; and m_p is the pupil magnification—*i.e.*, the diameter of the exit pupil divided by the diameter of the entrance pupil. Combining equations (8) and (9) gives

$$\text{Image illuminance} = E = \frac{\pi tB}{4(f\text{-number})^2(1 + m/m_p)^2} .$$

As an example in the use of this relation, if it is supposed that an $f/2$ lens is being used to project an image of a cathode-ray tube at five times magnification, the tube luminance being 5,000 foot-lamberts (1.7 candles per square centimetre), the lens transmittance is 0.8, and the pupil magnification is unity. Then the image illuminance will be

$$E = \frac{0.8 \times 5{,}000}{4 \times 4 \times 36} = 6.9 \text{ foot-candles.}$$

The image is very much less bright than the object, a fact that becomes clear to anyone attempting to provide a bright projected image in a large auditorium.

Distribution of illumination over an image. So far only the illumination at the centre of an image has been considered, but the distribution of illumination over a wide field is often important. In the absence of any lens, the small plane source already considered radiates in a direction inclined at an angle ϕ to the axis with an intensity $AB \cos \phi$. This light has to travel farther than the axial light to reach a screen, and then it strikes the screen at another angle ϕ. The net result is that the oblique illumination on the screen is smaller than the axial illumination by the factor $\cos^4 \phi$.

The same law can be applied to determine the oblique illumination due to a lens, assuming a uniform extended diffusing source of light on the other side of the lens. In this case, however, the exit pupil will not in general be a perfect circle because of possible distortion of the iris by that part of the optical system lying between the iris and the image. Also, any mechanical vignetting in the lens will make the aperture noncircular and reduce still further the oblique illumination. In a camera this reduction in oblique illumination results in darkened corners of the picture, but, if the reduction in brightness is gradual, it is not likely to be detected because the eye adapts quickly to changing brightness as the eyes scan over the picture area. Indeed, a 50 percent drop in brightness between the centre and corners of an ordinary picture is scarcely detectable.

Visual brightness. The apparent brightness of things seen by the eye follows the same laws as any other imaging system, because the apparent brightness is measured by the illuminance in the image on the retina. The angle U' in equation (8) inside the eye is determined by the size of the eye pupil, which varies from about one millimetre to about eight millimetres, depending on the brightness of the environment. Apart from this variation, retinal illuminance is directly proportional to object luminance, and objects having the same luminance appear equally bright, no matter at what distance they are observed.

Retinal illuminance

From this argument, it is clear that no visual instrument, such as a telescope, can possibly make anything appear brighter than when viewed directly. To be sure, a telescope having a large objective lens accepts more light from an object in proportion to the area of the lens aperture, but it magnifies the image area in the same proportion; so the increased light is spread over an increased area of the retina, and the illuminance remains unchanged. Actually, the telescopic view is always dimmer than the direct view because of light losses in the telescope due to glass absorption and surface reflections and because the exit pupil of the telescope may be smaller than the pupil of the eye, thus reducing the angle U'.

The case of a star being observed through a telescope is quite different, because no degree of magnification can possibly make a star appear as anything other than a point of light. Hence, star images appear brighter in proportion to the area of the telescope objective (assuming that the exit pupil is larger than the eye pupil), and the visibility of a star against the sky background is thus improved in proportion to the square of the diameter of the telescope objective lens. (R.K.)

II. Optics and information theory

GENERAL CONSIDERATIONS

A new era in optics commenced in the early 1950s following the impact of certain branches of electrical engineering—most notably communication and information theory. This impetus was sustained by the development of the laser in the 1960s.

The initial tie between optics and communication theory came because of the numerous analogies that exist between the two subjects and because of the similar mathematical techniques employed to formally describe the behaviour of electrical circuits and optical systems. A topic of considerable concern since the invention of the lens as an optical imaging device has always been the description of the optical system that forms the image; information about the object is relayed and presented as an image. Clearly, the optical system can be considered a communication channel and can be analyzed as such. There is a linear relationship (*i.e.*, direct proportionality) between the intensity distribution in the image plane and that existing in the object, when the object is illuminated with incoherent light (*e.g.*, sunlight or light from a large thermal source). Hence, the linear theory developed for the description of electronic systems can be applied to optical image-forming systems. For example, an electronic circuit can be characterized by its impulse response —that is, its output for a brief impulse input of current or voltage. Analogously, an optical system can be characterized by an impulse response that for an incoherent imaging system is the intensity distribution in the image of a point source of light; the optical impulse is a spatial rather than a temporal impulse—otherwise the concept is the same. Once the appropriate impulse response function is known, the output of that system for any object intensity distribution can be determined by a linear superposition of impulse responses suitably weighted by the value of the intensity at each point in the object. For a continuous object intensity distribution this sum becomes an integral. While this example has been given in terms of an optical imaging system, which is certainly the most common use of optical elements, the concept can be used independent of whether the receiving plane is an image plane or not. Hence, for example, an impulse response can be defined for an optical system that is deliberately defocussed or for systems used for the display of Fresnel or Fraunhofer diffraction patterns. (Fraunhofer diffraction occurs when the light source and diffraction patterns are effectively at infinite distances from the diffracting system, and Fresnel diffraction occurs when one or both of the distances are finite.)

The optical system as a communication channel

Temporal frequency response. A fundamentally related but different method of describing the performance of an electronic circuit is by means of its temporal frequency response. A plot is made of the response for a series of input signals of a variety of frequencies. The response is measured as the ratio of the amplitude of the signal obtained out of the system to that put in. If there is no loss in the system, then the frequency response is unity (one) for that frequency; if a particular frequency fails to pass

through the system, then the response is zero. Again, analogously the optical system may also be described by defining a spatial frequency response. The object, then, to be imaged by the optical system consists of a spatial distribution of intensity of a single spatial frequency—an object the intensity of which varies as $(1 + a \cos \omega x)$, in which x is the spatial coordinate, a is a constant called the contrast, and ω is a variable that determines the physical spacing of the peaks in the intensity distribution. The image is recorded for a fixed value of a and ω and the contrast in the image measured. The ratio of this contrast to a is the response for this particular spatial frequency defined by ω. Now if ω is varied and the measurement is repeated, a frequency response is then obtained.

Nonlinear optical systems. The analogies described above go even further. Many optical systems are nonlinear, just as many electronic systems are nonlinear. Photographic film is a nonlinear optical element in that equal increments of light energy reaching the film do not always produce equal increments of density on the film.

A different type of nonlinearity occurs in image formation. When an object such as two stars is imaged, the resultant intensity distribution in the image is determined by first finding the intensity distribution formed by each star. These distributions must then be added together in regions where they overlap to give the final intensity distribution that is the image. This example is typical of an incoherent imaging system—*i.e.*, the light emanating from the two stars is completely uncorrelated. This occurs because there is no fixed phase relationship between the light emanating from the two stars over any finite time interval.

A similar nonlinearity arises in objects illuminated by light from the Sun or other thermal light source. Illumination of this kind, when there is no fixed relationship between the phase of the light at any pair of points in the incident beam, is said to be incoherent illumination. If the illumination of the object is coherent, however, then there is a fixed relationship between the phase of the light at all pairs of points in the incident beam. To determine the resultant image intensity under this condition for a two point object requires that the amplitude and phase of the light in the image of each point be determined. The resultant amplitude and phase is then found by summation in regions of overlap. The square of this resultant amplitude is the intensity distribution in the image. Such a system is nonlinear. The mathematics of nonlinear systems was developed as a branch of communication theory, but many of the results can be used to describe nonlinear optical systems.

This new description of optical systems was extremely important to, but would not alone account for, the resurgence of optical research and development. This new approach resulted in the development of whole new branches of study, including optical processing and holography (see below). It also had an effect, together with the development of digital computers, on the concepts and versatility of lens design and testing. Finally, the invention of the laser, a device that produces coherent radiation, and the development and implementation of the theory of partially coherent light gave the added impetus necessary to change traditional optics into a radically new and exciting subject.

IMAGE FORMATION

Impulse response. An optical system that employs incoherent illumination of the object can usually be regarded as a linear system in intensity. A system is linear if the addition of inputs produces an addition of corresponding outputs. For ease of analysis, systems are often considered stationary (or invariant). This property implies that if the location of the input is changed, then the only effect is to change the location of the output but not its actual distribution. With these concepts it is then only necessary to find an expression for the image of a point input to develop a theory of image formation. The intensity distribution in the image of a point object can be determined by solving the equation relating to the diffraction of light as it propagates from the point object to the lens, through the lens, and then finally to the image plane. The result of this process is that the image intensity is the intensity in the Fraunhofer diffraction pattern of the lens aperture function (that is, the square of the Fourier transform of the lens aperture function; a Fourier transform is an integral equation involving periodic components). This intensity distribution is the intensity impulse response (sometimes called point spread function) of the optical system and fully characterizes that optical system. Figure 8 shows the impulse response for a perfect optical system (no aberration) that is limited only by its aperture.

Intensity distribution

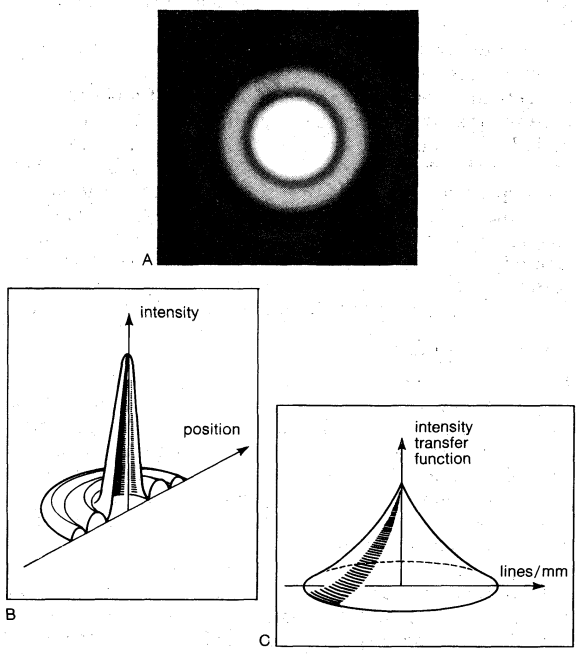


Figure 8: *The intensity impulse response of a perfect (i.e., no aberrations) optical system.*
(A) Photograph; (B) plot; (C) the transfer function of the same aberration-free system.

With the knowledge of the impulse response, the image of a known object intensity distribution can be calculated. If the object consists of two points, then in the image plane the intensity impulse response function must be located at the image points and then a sum of these intensity distributions made. The sum is the final image intensity. If the two points are closer together than the half width of the impulse response, they will not be resolved. For an object consisting of an array of isolated points, a similar procedure is followed—each impulse response is, of course, multiplied by a constant equal to the value of the intensity of the appropriate point object. Normally, an object will consist of a continuous distribution of intensity, and, instead of a simple sum, a convolution integral results.

Transfer function. The concept of the transfer function of an optical system can be approached in several ways. Formally and fundamentally it is the Fourier transform of the intensity impulse response. The transfer function for the optical system for Figure 8B is shown in Figure 8C. Because the impulse response is related to the lens aperture function, so is the transfer function. In particular, the transfer function can be obtained from a knowledge of the aperture function by taking the function and plotting the resultant overlapping areas as the aperture function is slid over itself (*i.e.*, the autocorrelation of the aperture function).

Autocorrelation

Conceptually, however, the transfer function is best understood by considering the object intensity distribution to be a linear sum of cosine functions of the form $(1 + a \cos 2\pi\mu x)$, in which a is the amplitude of each component of spatial frequency μ. The image of a cosine intensity distribution is a cosine of the same frequency; only the contrast and phase of the cosine can be affected by a

linear system. The image of the above object intensity distribution can be represented by $[1 + b \cos (2\pi\mu x + \phi)]$, in which b is the amplitude of the output cosine of frequency μ and ϕ is the phase shift. The transfer function, $\tau(\mu)$, for that frequency is then given by the ratio of the amplitudes:

$$\tau(\mu) = \frac{b}{a} e^{j\phi(\mu)}.$$

If μ is now varied, the spatial frequency response of the system is measured by determining $\tau(\mu)$ for the various values of μ. It should be noted that $\tau(\mu)$ is in general complex (containing a term with $\sqrt{-1}$), although in the example of Figure 8 it is real and positive [$\phi(\mu) = 0$ for all μ].

The transfer function, like the impulse response, fully characterizes the optical system. To make use of the transfer function to determine the image of a given object requires that the object be decomposed into a series of periodic components called its spatial frequency spectrum. Each term in this series must then be multiplied by the appropriate value of the transfer function to determine the individual components of the series that is the spatial frequency spectrum of the image—a transformation of this series will give the image intensity. Thus, any components in the object spectrum that have a frequency for which $\tau(\mu)$ is zero will be eliminated from the image.

PARTIALLY COHERENT LIGHT

Image formation is concerned above with incoherent object illumination, which results in an image formed by the addition of intensities. The study of diffraction and interference, on the other hand, requires coherent illumination of the diffracting object, the resulting diffracted optical field being determined by an addition of complex amplitudes of the wave disturbances. Thus, two different mechanisms exist for the addition of light beams, depending upon whether the beams are coherent or incoherent with respect to each other. Unfortunately, this is not the whole story; it is not sufficient to consider only the two situations of strictly coherent and strictly incoherent light. In fact, strictly incoherent fields are only approximately obtainable in practice. Furthermore, the possibility of intermediate states of coherence cannot be ignored; it is necessary to describe the result of mixing incoherent light with coherent light. It was to answer the question how coherent is a beam of light (or the equivalent one, how incoherent is a beam of light) that the theory of partial coherence was developed. Marcel Verdet, a French physicist, realized in the last century that even sunlight is not completely incoherent, and two objects separated by distances of over approximately $\frac{1}{20}$ millimetre will produce interference effects. The eye, operating unaided in sunlight, does not resolve this separation distance and hence can be considered to be receiving an incoherent field. Two physicists, Armand Fizeau in France and Albert Michelson in the United States, were also aware that the optical field produced by a star is not completely incoherent and hence designed interferometers to measure the star diameter from a measurement of the partial coherence of the starlight. These early workers did not think in terms of partially coherent light but derived their results by an integration over the source. At the other extreme, the output from a laser can produce a highly coherent field.

Partial coherence

The concepts of partially coherent light can best be understood by means of some simple experiments. A circular uniform distant source produces illumination on the front of an opaque screen containing two small circular apertures, the separation of which can be varied. A lens is located behind this screen, and the resultant intensity distribution in its focal plane is obtained. With either aperture open alone, the intensity distribution observed is such that it is readily associated with the diffraction pattern of the aperture, and it may thus be concluded that the field is coherent over the dimensions of the aperture. When the two apertures are opened together and are at their closest separation, two-beam interference fringes are observed that are formed by the division of the incident wave front by the two apertures. As the separation of the apertures increases, the observed interference fringes get weaker and finally disappear, only to reappear faintly as the separation is further increased. As the separation of the apertures is increased, these results show that (1) the fringe spacing decreases; (2) the intensities of the fringe minima are never zero; (3) the relative intensity of the maxima above the minima steadily decreases; (4) the absolute value of the intensity of the maxima decreases and that of the minima increases; (5) eventually, the fringes disappear, at which point the resultant intensity is just twice the intensity observed with one aperture alone (essentially an incoherent addition); (6) the fringes reappear with a further increase in separation of the aperture, but the fringes contain a central minimum, not a central maximum.

If the intensities of the two apertures are equal, then the results (1) through (5) can be summarized by defining a quantity in terms of the maximum intensity (I_{max}) and the minimum intensity (I_{min}), called the visibility (V) of the fringes—*i.e.*, $V = (I_{max} - I_{min})/(I_{max} + I_{min})$. The maximum value of the visibility is unity, for which the light passing through one aperture is coherent with respect to the light passing through the other aperture; when the visibility is zero, the light passing through one aperture is incoherent with respect to the light passing through the other aperture. For intermediate values of V the light is said to be partially coherent. The visibility is not a completely satisfactory description because it is, by definition, a positive quantity and cannot, therefore, include a description of item (6) above. Furthermore, it can be shown by a related experiment that the visibility of the fringes can be varied by adding an extra optical path between the two interfering beams.

Fringe visibility

The key function in the theory of partially coherent light is the mutual coherence function $\Gamma_{12}(\tau) = \Gamma(x_1, x_2, \tau)$, a complex quantity, which is the time averaged value of the cross correlation function of the light at the two aperture points x_1 and x_2 with a time delay τ (relating to a path difference to the point of observation of the interference fringes). The function can be normalized (*i.e.*, its absolute value set equal to unity at $\tau = 0$ and $x_1 = x_2$) by dividing by the square root of the product of the intensities at the points x_1 and x_2 to give the complex degree of coherence, hence

$$\gamma_{12}(\tau) = \frac{\Gamma_{12}(\tau)}{\sqrt{I(x_1)I(x_2)}}.$$

The modulus of $\gamma_{12}(\tau)$ has a maximum value of unity and a minimum value of zero. The visibility defined earlier is identical to the modulus of the complex degree of coherence if $I(x_1) = I(x_2)$.

Often the optical field can be considered to be quasi-monochromatic (approximately monochromatic), and then the time delay can be set equal to zero in the above expression, thus defining the mutual intensity function. It is often convenient to describe an optical field in terms of its spatial and temporal coherence by artificially separating out the space- and time-dependent parts of the coherence function. Temporal coherence effects arise from the finite spectral width of the source radiation; a coherence time Δt can be defined as $1/\Delta\nu$, in which $\Delta\nu$ is the frequency bandwidth. A related coherence length Δl can also be defined as $c/\Delta\nu = \lambda^2/\Delta\lambda^2$, in which c is the velocity of light, λ is the wavelength, and $\Delta\lambda$ the wavelength bandwidth. Providing the path differences in the beams to be added are less than this characteristic length, they will interfere.

The term spatial coherence is used to describe partial coherence arising from the finite size of an incoherent source. Hence, for the equipath position for the addition of two beams, a coherence interval is defined as the separation of two points such that the absolute value $|\gamma_{12}(0)|$ is some prechosen value, usually zero.

The mutual coherence function is an observable quantity that can be related to the intensity of the field. The partially coherent field can be propagated by use of the mutual coherence function in a similar way to the solution of diffraction problems by propagation of the com-

The mutual coherence function

plex amplitude. The effects of partially coherent fields are clearly of importance in the description of normally coherent phenomena, such as diffraction and interference, but also in the analysis of normally incoherent phenomena, such as image formation. It is notable that image formation in coherent light is not linear in intensity but is linear in the complex amplitude of the field, and in partially coherent light the process is linear in the mutual coherence.

OPTICAL PROCESSING

Coherent optical systems. Optical processing, information processing, signal processing, and pattern recognition are all names that relate to the process of spatial frequency filtering in a coherent imaging system—specifically, a method in which the Fraunhofer diffraction pattern (equivalently the spatial frequency spectrum or the Fourier transform) of a given input is produced optically and then operated upon to change the information content of the optical image of that input in a predetermined way.

The idea of using coherent optical systems to allow for the manipulation of the information content of the image is not entirely new. The basic ideas are essentially included in Abbe's theory of vision in a microscope first published in 1873; the subsequent illustrative experiments of this theory, notably by Albert B. Porter in 1906, are certainly simple examples of optical processing.

Abbe's ideas can be interpreted as a realization that image formation in a microscope is more correctly described as a coherent image-forming process than as the more familiar incoherent process. Thus, the coherent light illuminating the object on the microscope stage would be diffracted by that object. To form an image, this diffracted light must be collected by the objective lens of the microscope, and the nature of the image and the resolution would be affected by how much of the diffracted light is collected. As an example, an object may be considered consisting of a periodic variation in amplitude transmittance—the light diffracted by this object will exist in a series of discrete directions (or orders of diffraction). This series of orders contains a zero order propagating along the optical axis and a symmetric set of orders on both sides of this zero order. Abbe correctly speculated upon what would happen as the microscope objective accepted different combinations of these orders. For example, if the zero order and one first order are collected, then the information obtained from the image will be that the object consisted of a periodic distribution, but the spatial location of the periodic structure is not correctly ascertained. If the other first order of diffracted light is included, then the correct spatial location of the periodic structure is also obtained. As more orders are included, the image more closely resembles the object.

Coherent optical data processing became a serious subject for study in the 1950s, partly because of the work of a French physicist, Pierre-Michel Duffieux, on the Fourier integral and its application to optics, and the subsequent use of communication theory in optical research. The work was initiated in France by André Maréchal and Paul Croce, and today a variety of problems can be attempted by the technique. These include removal of raster lines (as in a TV picture) and halftone dots (as in newspaper illustration); contrast enhancement; edge sharpening; enhancement of a periodic or isolated signal in the presence of additive noise; aberration balancing in which a recorded aberrated image can be somewhat improved; spectrum analysis; cross correlation of data; matched and inverse filtering in which a bright spot of light in the image indicates the presence of a particular object.

Examples of data processing

Filtering. The basic system required for coherent optical processing consists of two lenses (Figure 9). A collimated beam of coherent light is used to transilluminate the object. The first lens produces the characteristic Fraunhofer diffraction pattern of the object, which is the spatial frequency distribution associated with the object. (Mathematically, it is the Fourier transform of the object amplitude distribution.) A filter that consists of amplitude (density) or phase (optical path) variations, or both, is placed in the plane of the diffraction pattern. The light passing through this filter is used to form an image, this step being accomplished by the second lens. The filter has the effect of changing the nature of the image by altering the spatial frequency spectrum in a controlled way so as to enhance certain aspects of the object information. Maréchal gave the descriptive title double diffraction to this type of two-lens system.

The filters can be conveniently grouped into a variety of types depending upon their action. Blocking filters have regions of complete transparency and other regions of complete opacity. The opaque areas completely remove certain portions of the spatial frequency spectrum of the object. The removal of raster lines and halftone dots is accomplished with this type of filter. The object can be considered as a periodic function the envelope of which is the scene or picture—or equivalently the periodic function samples the picture. The diffraction pattern consists of a periodic distribution with a periodicity reciprocally related to the raster periodicity. Centred at each of these periodic locations is the diffraction pattern of the scene. Hence, if the filter is an aperture centred at one of these locations so that only one of the periodic elements is allowed to pass, then the raster periodicity is removed, but the scene information is retained (see Figure 9). The

Blocking filters

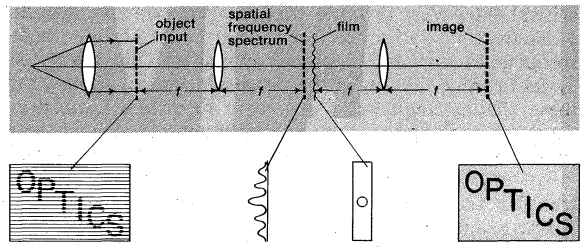


Figure 9: Two-lens coherent optical processing system, showing how the raster periodicity is removed but the scene information is retained (see text).

problem of the removal of halftone dots is the two-dimensional equivalent of the above process. Because the two-dimensional spatial frequency spectrum of an object is displayed in a coherent optical processing system, it is possible to separate out information by means of its orientation. Other applications of blocking filters include band-pass filters, which again have a direct relationship to the band-pass filters in electronic circuits.

A second type of filter is an amplitude filter that will consist of a continuous density variation. These filters can be produced to achieve the enhancement of contrast of the object input or the differentiation of the object. They are often constructed by controlled exposure of photographic film or evaporation of metal onto a transparent substrate.

Amplitude filters

Certain optical processing techniques require that the phase of the optical field be changed, and, hence, a filter with no absorption but varying optical thickness is required. Usually, both the amplitude and the phase have to be modified, however, thus requiring a complex filter. In simple cases the amplitude and phase portions can be made separately, the phase filter being manufactured by using an evaporated layer of transparent material, such as magnesium fluoride. Current practice is to fabricate the complex filter by an interferometric method in which the required complex amplitude function is recorded as a hologram (see below).

The phase-contrast microscope can be considered to be an example of an optical processing system, and the concepts understood by reference to Figure 9. Only the simplest form will be considered here. The spatial frequency spectrum of the phase object is formed and the phase of the central portion of that spectrum changed by $\pi/2$ or $3\pi/2$ to produce positive or negative phase contrast, respectively. To improve the contrast of the image an additional filter covering the same area as the phase filter is used that is partially absorbing (*i.e.*, an amplitude filter). The restriction on this process is that the variations of the

phase $\phi(x)$ are small so that $e^{i\phi(x)} \cong 1 + i\phi(x)$. With incoherent light, phase information is not visible, but many biological samples consist only of variations of refractive index, which results in optical path, and hence phase, differences. The image in the phase-contrast microscope is such that the intensity in that image relates linearly to, and hence is a display of, the phase information in the object—*e.g.*, $I(x) \propto 1 \pm 2\phi(x)$ for positive and negative phase contrast, respectively.

Correction of aberrated images

One of the important motivations for the study of optical processing methods is to achieve some correction of aberrated images. Considerable technological advantage can be gained if photographs taken with an aberrated optical system in incoherent light can be corrected by subsequent processing. Within definable limits this can be accomplished, but the impulse response or the transfer function of the aberrated system must be known. The recorded image intensity distribution is the convolution of the object intensity with the intensity impulse response of the aberrated system. This record is the input to the coherent optical processing system; the diffraction pattern formed in this system is the product of the spatial frequency spectrum of the object and the transfer function of the aberrated system. Conceptually, the filter has to be the inverse of the transfer function in order to balance out its effect. The final image would then ideally be an image of the object intensity distribution. It is critical, however, that the transfer function has a finite value over only a limited frequency range, and only those frequencies that are recorded by the original aberrated system can be present in the processed image. Hence, for these spatial frequencies that were recorded, some processing can be carried out to get a flatter effective transfer function; both the contrast and the phase of the spatial frequency spectrum may have to be changed because the transfer function is, in general, a complex function. Prime examples are for images aberrated by astigmatism, defocussing, or image motion.

HOLOGRAPHY

Theory. Holography (*q.v.*) is a two-step coherent image-forming process in which an intermediate record is made of the complex optical field associated with the object. The invention of the wave-front reconstruction process (now called holography) was first described in 1948 by Dennis Gabor, a Hungarian-born physicist, with a specific application in mind—to attempt to improve the resolution of images formed with electron beams. The technique has, however, had most of its success to date when light beams are employed particularly in the visible part of the spectrum. The first step in the process is to record (often on high-resolution film) the interference pattern produced by the interaction of the light diffracted by the object of interest and a coherent background or reference wave. In the second step, this record, which is the hologram, is illuminated coherently to form an image of the original object. In fact, two images are usually formed—a real image (often called the conjugate image) and a virtual image (often called the primary image). There are two basic concepts that underly this process: first, the addition of a coherent background (or reference) beam. Two optical fields may be considered, the complex amplitudes of which vary as the cosine of an angle proportional to the space coordinate and as the modulus (absolute magnitude) of the cosine of the angle, respectively. From a measurement of the intensity of these fields it is impossible to distinguish them because both vary as the cosine squared of the space coordinate. If a second coherent optical field is added to each of these two fields, however, then the resultant fields become $(1 + \cos x)$ and $(1 + |\cos x|)$, respectively. The measured intensities are now different, and the actual fields can be determined by taking the square root of the intensity. The amplitude transmittance of a photographic record is, in fact, the square root of the original intensity distribution that exposed the film. In a more general sense, an optical field of the form $a(x) \exp [i\phi_1(x)]$, in which $a(x)$ is the amplitude and $\phi_1(x)$ is the phase, can be distinguished from a field $a(x) \exp [i\phi_2(x)]$ by adding a coherent background; the phases $\phi_1(x)$ and $\phi_2(x)$ are then contained as cosine variations of intensity in the resulting pattern. Hence, the problem of recording the phase information of the optical field is circumvented. When the hologram is illuminated, however, the optical field that originally existed in that plane is recreated. To apply the second basic concept—that of an image-forming property—it is necessary to determine what the hologram of a point object is—in actuality it is a sine-wave zone plate or zone lens. If a collimated beam of light is used to illuminate a zone lens, then two beams are produced; the first comes to a real focus, and the other is a divergent beam that appears to have come from a virtual focus. (By comparison, the more classical zone plate has a multitude of real and virtual focuses, and a real lens has but one.) When the object is other than a point, the zone lens is modified by the diffraction pattern of the object; *i.e.*, each point on the object produces its own zone lens, and the resultant hologram is a summation of such zone lenses.

In Gabor's original system the hologram was a record of the interference between the light diffracted by the object and a collinear background. This automatically restricts the process to that class of objects that have considerable areas that are transparent (see Figure 10A). When the

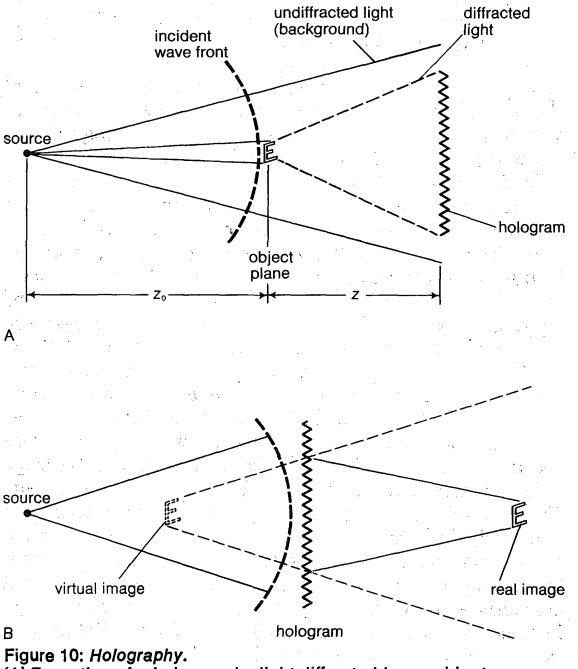


Figure 10: *Holography.*
(A) Formation of a hologram by light diffracted by an object with collinear background (Gabor's original method). (B) Image formation by illumination of the hologram.

hologram is used to form an image, twin images are formed, as illustrated in Figure 10B. The light associated with these images is propagating in the same direction, and hence in the plane of one image light from the other image appears as an out-of-focus component. This type of hologram is usually referred to as an in-line Fresnel hologram because it is the pattern of the object that interferes with the collinear coherent background. The deleterious effects of the second image can be minimized if the hologram is made in the far field of the object so that it is a Fraunhofer diffraction pattern of the object that is involved. This latter technique has found significant application in microscopy, particularly in the measurement of small particles, and in electron microscopy.

A more versatile method of recording the hologram is to add a second beam of light as a reference wave to produce the hologram. The hologram is now the record of the interference pattern produced by the light diffracted by the object and this separate reference wave. The reference wave is usually introduced at an angle to the diffracted beam, hence this method is often called off-axis (or side-band) holography. When the hologram is illumi-

Steps in the holography process

nated, the image-forming beams do not propagate in the same direction but are inclined to each other with an angle twice that between the diffracted beam and the original reference beam. Hence, the light associated with an image is completely separated from the other image.

The generalized hologram

A further technique that has some value and relates to the earlier discussion of optical processing is the production of the so-called generalized or Fourier transform hologram. Here the reference beam is added coherently to a Fraunhofer diffraction pattern of the object or formed by a lens (as in the first stage of Figure 9).

The process described so far has been in terms of transmitted light through the object. The methods involving the separate reference beam can be used in reflected light, and the virtual (primary) image produced from the hologram has all the properties of an ordinary image in terms of three-dimensionality and parallax. Normally, a recorded image is only a two-dimensional representation of the object. Full-colour holograms can be recorded by essentially recording three holograms simultaneously—one in red light, one in blue, and one in green.

Applications. *Image forming.* The applications mentioned here are in three groups: image-forming applications, non-image-forming applications, and the hologram as an optical element. It is notable that all three groups relate to the basic use of the process rather than specific holographic techniques. The first group involves those applications using image formation when, for a variety of reasons, normal incoherent or coherent image formation is not satisfactory. It is not sufficient merely to replace a normal image process by a holographic technique unless there is some significant gain—*i.e.*, the required record can be obtained more easily or more accurately. Applications that fall into this category are holographic microscopy; particle-size analysis; high-speed photography of various types, particularly of gas flows; data storage and retrieval, including displays; image formation through a random medium; and non-optical holography, particularly acoustic holography.

Non-image forming. The second group of interest involves those applications that are not image forming. One of the very real and exciting applications of holography is to the nondestructive testing of fabricated materials. An interesting example of this method is for the testing of tires for the detection of flaws (debonds) that exist between the plies of the tire. The realm of interferometry is thus extended to whole new classes of objects. In a similar but separate development, interference microscopy has been used successfully.

Optical elements. The third and final group involves those applications that use the hologram as an optical element in its own right. This includes the building of accurate, specialized gratings and the application of holographic filters in coherent optical data processing.

Holography has been adapted to the conventional microscope, which is modified by the inclusion of a separate reference beam so that the light diffracted by the object in the microscope is made to interfere with the light from the reference beam. An increase in the depth of field available is achieved by this type of recording process. The image is produced when the hologram is illuminated again by a coherent beam.

The application of holography to particle-size analysis (*e.g.*, to determine the size distribution of dust and liquid droplets) was really the first of the modern-day applications. In a sense, this, too, can be thought of as microscopy. The principles of Fraunhofer holography were developed to solve this particular problem. Because the particles are in motion, a hologram must be made instantaneously. A pulsed-ruby laser technique is therefore used. The hologram is formed between the light diffracted by the particles or droplets and the coherent background light that passes directly through the sample. In reconstruction, a series of stationary images are formed that can be examined at leisure. Hence, a transient event has been transformed into a stationary image for evaluation.

Transient holograms

Data storage and retrieval is perhaps one of the more important applications of holography, which is in the process of development. Because the information about the image is not localized, it cannot be affected by scratches or dust particles. Recent advances in materials have added a further interest to this area, particularly materials that might be erasable and reusable. Hence, holographic optical memories become a distinct possibility.

Among the non-image-forming applications are interferometry, interference microscopy, and optical processing. Holographic interferometry can be done in several ways. The basic technique involves recording a hologram of the object of interest and then interfering the image produced from this hologram with the coherently illuminated object itself. A variation on this technique would be to form two holograms at different times of the same object as it undergoes testing. The two holograms can then be used together to form two images, which would again interfere. The interference fringes seen would be related to the changes in the object between the two exposures. A third technique uses a time-average hologram, which is particularly applicable to the study of vibrating objects.

There are two applications that come under the heading holographic optical elements—the use of holographic gratings and the use of holographic filters for coherent optical data processing.

NONLINEAR OPTICS

Nonlinear effects in optics are now quite readily observable using the highly coherent and highly energetic laser beams. These effects occur when the output of a system is not linearly related to the input (*e.g.*, a nonlinear electronic amplifier can be built with a gain that increases with signal intensity). The most important nonlinear effect is probably frequency doubling. Optical radiation of a given frequency is propagated through a crystalline material and interacts with that material to produce an output of a different frequency that is twice the input frequency. For example, the 10,600 angstroms infrared output of a neodymium laser can, under suitable conditions, be converted into green light at 5300 angstroms in a crystal of barium strontium niobate. (B.J.T.)

BIBLIOGRAPHY. There are many journals and hundreds of books covering the general field of optics, some of the more familiar books include A.C. HARDY and F.H. PERRIN, *The Principles of Optics* (1932); and F.A. JENKINS and H.E. WHITE, *Fundamentals of Optics*, 3rd ed. (1957). At a somewhat more advanced level are R.S. LONGHURST, *Geometrical and Physical Optics*, 2nd ed. (1967); L.C. MARTIN, *Technical Optics*, 2 vol. (1960–61); and MAX BORN and EMIL WOLF, *Principles of Optics*, 2nd ed. (1964). In the purely geometrical field, the following can be recommended: W.T. WELFORD, *Geometrical Optics* (1962), which is fairly elementary and provides an excellent introduction to the subject; and L.C. MARTIN, *Geometrical Optics* (1956), which is somewhat more advanced.

Recommended books on the subject of information theory are EDWARD L. O'NEILL, *Introduction to Statistical Optics* (1963); JOSEPH W. GOODMAN, *Introduction to Fourier Optics* (1968); ARNOLD R. SCHULMAN, *Optical Data Processing* (1970); KENDALL PRESTON, *Coherent Optical Computers* (1972); HENRY LIPSON (ed.), *Optical Transforms* (1972); GEORGE W. STROKE, *An Introduction to Coherent Optics and Holography*, 2nd ed. (1969); JOHN B. DeVELIS and GEORGE O. REYNOLDS, *Theory and Applications of Holography* (1967); and ROBERT J. COLLIER, CHRISTOPH B. BURCKHARDT, and LAWRENCE H. LIN, *Optical Holography* (1971). For particular topics, reference should be made to the *Journal of the Optical Society of America* (monthly); *Applied Optics* (monthly); and EMIL WOLF (ed.), *Progress in Optics* (1961–).

(R.K./B.J.T.)

Optimization, Mathematical Theory of

Optimization is a technique for improving or increasing the value of some numerical quantity that in practice may take the form of temperature, air flow, speed, pay-off in a game, political appeal, destructive power, information, monetary profit, and the like. Techniques of optimization assume such varied forms that no one general description is possible. With the advent of modern technology more and more emphasis has been placed on optimization of various types, and special thinking has developed to the

extent that it is meaningful to speak of a mathematical theory of optimization. Computer technology has been critically important in practical applications, such as in the optimal control of rockets. Further advances in the optimization and control of complex systems will probably depend more on mathematical theory than on technological invention.

An outline of the subject as presented in this article is as follows:

(Ed.)

I. The theory of games

Early work on mathematical optimization was extended by the development of game theory, a branch of mathematics that aims to analyze various problems of conflict by abstracting common strategic features for study in mathematical problems called *games* because they are patterned on such actual games as bridge and poker.

In game theory emphasis is placed on a correct description of the problem. By stressing strategic aspects—aspects controlled by the participants rather than by pure chance—the method goes beyond the classical theory of probability. A theory of games of strategy was broached in 1921 by a French mathematician, Émile Borel, and was established in 1928 by the Hungarian-born U.S. mathematician, John von Neumann. In 1944, von Neumann, then in America, published with the Austrian-born U.S. economist Oskar Morgenstern an account of the theory applied to competitive economic behaviour, based on the presence of several factors that are common to both actual games and economic situations: conflicting interests, incomplete information, and the interplay of rational decision and chance. The account opened a wide range of gamelike problems in sociology, psychology, politics, and war to mathematical attack.

CLASSIFICATION OF GAMES

Singular, dual, and plural games. In classifying games by the number of players, the significant figure is not the number of individuals involved but the number of parties with distinct interests actively represented in the play. It is assumed specifically that these interests are measured in terms of, for example, money or some other numerical scale of utility. This assumption is treated below under *Utility*. The presence of only one such interest, though there may be more than one participant, characterizes a game as singular. Typical are solitaire card games, many types of puzzles, and problems of an isolated economic unit with a single goal (*e.g.*, Robinson Crusoe's plight). Such games are distinguished by the absence of conflict; with no opponent to thwart his plans, the single player need only, in theory, list all his possible courses of action and then select the best, as measured on his scale of utility. Conflict enters with dual games (called by von Neumann and Morgenstern zero-sum two-person games), played by two parties in diametric opposition—what one wins, the other loses. Each party, in seeking an optimal course of action, must reckon with the possible actions of an opponent with contrary aims. This direct conflict of interests is exemplified by two-person board games such as chess, two-handed card games such as cribbage, and two-team card games such as bridge. In contrast, the presence of active interests that are not diametrically opposed such as the sales of two or more sellers in competition with each other stamps a game as plural. Sellers competing for the same market are engaged in a plural game in which the active interests (to be measured in money) are the sales. All games having more than two active interests fall into the plural class; so do two-party bargaining situations in which both parties stand to gain by reaching an agreement.

Extensive and normalized forms. From the rules of any game, two abstract forms can be obtained. The first of these, the extensive form of a game, eliminates all features that refer specifically to the means of playing it—all those features that characterize it as a card game, a dice game, or an economic conflict. Thus, the extensive form amounts to a literal translation of the rules into the technical terms of a formal system designed to describe all games. The second, or normalized form, is a more condensed version of the game, stripped of all features but the choice of over-all strategies.

The raw material for the extensive form comes from the common elements of strategic games, such as the interrelations between the players' choices, the variety and character of information available to them, and the effect of chance on the play. The object of a systematic description of these features is to isolate those junctures at which a player is called upon to make a choice and the states of information on which his choices are based. From this analysis emerges a precise definition of the first key concept on which the theory of games has been built. This is the notion of a pure strategy for a player; namely, a complete set of advance instructions that specifies a definite choice for every conceivable situation in which the player may be required to act. Such a set of instructions represents a total plan that covers all contingencies the player may face during any playing of the game, whether these are attributable to choices of other players or to events governed by chance; its execution could be delegated perfectly well to an agent without discretionary power. The complexity imposed on a single pure strategy by its completeness and the enormous number of pure strategies that may need cataloguing explain the failure of the theory to analyze most common parlour games. The practical difficulties, however, do not prevent the enumeration in theory of all the pure strategies of a game.

The notion of pure strategy

Whereas the extensive form of a game may have a large number of moves, or junctures at which a player must make a choice, in the normalized form each player makes only one move, which consists of a single choice from his set of pure strategies. As a result, the game is replaced by the following prototype: Each player chooses a pure strategy, making his choice in absolute ignorance of the

choices of the other players. After all strategies have been chosen, they are submitted to an umpire who charts the course of the play, making chance moves in accordance with the probabilities dictated by the rules and otherwise using the choices given by the players' pure strategies. In this way there is determined for each player a set of outcomes with associated probabilities, the expected value of which constitutes the player's payoff.

Finite and infinite games. In most games the number of pure strategies available to the players is finite because the rules ordinarily make the game terminate after a finite number of occasions for choice by the players, and because, at each occasion, a player is usually confronted by a finite number of alternatives.

Games with an infinite number of strategies often arise as idealizations of situations that are too complicated to be handled in their original form (such as replacing the full, but finite, range of poker hands by a continuum of numbers). It has proved possible in this way to treat problems of many sorts as normalized dual games in which the pure strategies for each player are represented by the real numbers from zero to one (called games on a square). Infinite games have also provided a fertile source of pathological examples; *i.e*, those offering a contrast to the regular properties found in the finite case. In a different direction, the making of decisions from statistical treatment of a body of data has been used to show that the general problem of statistical inference may be regarded as an infinite dual game in which the statistician is pitted against nature.

UTILITY

The fundamental problem of the theory of games is to find the methods by which a player can obtain what is called a most favourable result. In the first development of the theory, the most favourable result was identified with the greatest expected (monetary) gain, because this, or some similar assumption, is necessary if probabilistic methods are to be applied. There are, however, strong objections to the principle of maximizing expected winnings as a prescription for behaviour, and these led to a re-examination of the concept of utility.

To understand these objections and the modern theory of utility that resulted, the following situation may be considered: A player is required to choose an integer from 1, 2, $\cdots$, n, and is then paid a sum of money P_j that depends upon his choice of j. If the player always prefers more money to less, then rational behaviour is clear in this situation of certainty; namely, the player should choose j so as to make P_j as large as possible. If this situation is altered to introduce uncertainty and risk, then the rule is no longer obvious.

The St. Petersburg paradox

This can be illustrated with a classical example known as the St. Petersburg paradox. A player is confronted with a choice between the status quo and the payment of an entry fee F for an infinite set of lottery tickets numbered 1, 2, $\cdots$, n, $\cdots$. The lottery ticket n is to pay the amount 2^n with associated odds of one chance in 2^n. The rule of maximizing expected winnings counsels the purchase of the lottery tickets, no matter how high the entry fee F. Numerical examples, however, persuade many people of the advantages of the status quo; for example, with $F = \$128$, there is but one chance in 64 that a player who chooses the lottery tickets will break even, and he will otherwise lose at least \$64. If the game is to be played but once, it would mean assuming a high risk for the prospect of highly improbable winnings.

To resolve this paradox, a Swiss mathematician, Daniel Bernoulli, suggested that people do not follow monetary value as an index for preferences but rather are concerned with the moral worth of money. Further, in a situation involving risk, they seek to maximize the expected value of moral worth (what has been called moral expectation). Finally, he proposed a quite serviceable function to measure the moral worth of an amount of money, namely, its logarithm.

Whatever the defects of this function as a universal measure of preferences, there is a need for a numerical index that will reflect accurately the choices of an individual in situations of risk. Interest in this problem was first shown by the British logician F.P. Ramsey who defined utility operationally in terms of individual behaviour. Independently, von Neumann and Morgenstern confronted this problem while laying the foundations for the theory of games. Informally, they showed that if a player can consistently express his preferences between every possible pair of gambles on outcomes, then there exists a utility function defined on the possible outcomes such that its expected value is an accurate guide to the player's choices in situations of risk. This numerical measure of utility permits the definition of the most favourable result as that with the greatest utility.

Utility function

NORMALIZED DUAL GAMES

Matrix games; saddlepoints. A finite dual game, when normalized, has a simple conceptual scheme. Each of the players, A and B, chooses a pure strategy from a finite list, without knowing the other's choice. The outcome thus determined can be specified by a single number P, the payoff A stands to receive: positive, negative, or zero according as A wins from B, loses, or draws. By the diametric opposition characteristic of a dual game, B then stands to win $-P$. In this way the game is described by a rectangular array of numbers, in which each horizontal row corresponds to a pure strategy for A and each vertical column to a pure strategy for B; the entry P common to the row and column is the payoff to A from B that these pure strategies produce. The array of numbers is called the payoff matrix; normalized finite dual games presented in such form are called matrix games. (Examples appear in Tables 1 and 2; in each case the payoff matrix is the central box of arabic numbers.) Because the matrix entries represent gains to A and losses to B, A's aim is to maximize and B's to minimize the result that comes out of the matrix.

Table 1: Games of Odd and Even

wt.		1	1	B
no.		I	II	
1	I	−1	1	0
1	II	1	−1	0
A		0	0	÷2

(1) minimax = 0

wt.		7	5	B
no.		I	II	
7	I	−2	3	1
5	II	3	−4	1
A		1	1	÷12

(2) minimax = $\frac{1}{12}$

wt.		1	0	B
no.		I	III	
0	I	−2	−4	−2
1	II	3	5	3
A		3	5	÷1

(3) minimax = 3

wt.		1	2	1	B
no.		I	II	III	
1	I	−2	3	−4	0
2	II	3	−4	5	0
1	III	−4	5	−6	0
A		0	0	0	÷4

(4) minimax = 0

wt.		0	17	0	0	11	B
no.		I	II	III	IV	V	
0	I	−2	3	−4	5	−6	−15
15	III	−4	5	−6	7	−8	−3
13	IV	5	−6	7	−8	9	−3
A		5	−3	1	1	−3	÷28

(5) minimax = $-\frac{3}{28}$

Table 2: Two Games of Morra

wt.			0	1	1	0	
show			I		II		B
call			II	III	III	IV	
1	I	II	0	1	−1	0	0
0		III	−1	0	0	1	0
0	II	III	1	0	0	−1	0
1		IV	0	−1	1	0	0
A			0	0	0	0	÷2

wt.			0	20	15	0	
show			I		II		B
call			II	III	III	IV	
0	I	II	0	2	−3	0	−5
21		III	−2	0	0	3	0
14	II	III	3	0	0	−4	0
0		IV	0	−3	4	0	0
A			0	0	0	7	÷35

As A weighs the consequences of various courses of action open to him in a dual game, the theory of games counsels him to gauge each by the gain it assures him, regardless of what B does. In effect, this is equivalent to the pessimistic assumption by A that B knows his plan and will counter it to limit his gain to a bare minimum. So, to maximize his assured gain, A is led to seek a maximum of minima, abbreviated max-min; he should act so as to make as great as possible the least gain to which B can limit him. The antithetical aim of B is to choose a course of action that will hold to a minimum the greatest loss A may inflict. Thus, B's goal is a minimum of maxima, abbreviated min-max. The clear fact that A cannot establish a floor under his possible gains that is higher than the ceiling B succeeds in placing over his possible losses is expressed by a simple formula expressed in terms of a "less than or equal to" relationship (see Box, formula 1).

Max-min and min-max

With evaluation limited to pure strategies only, this is the most precise statement that can be made for the entire class of dual games. (The effect of further restrictions is considered below under *Extensive games—a poker example.*) The highest gain-floor for A, using a single pure strategy, is obtained by choosing the strategy row in which the least entry is greatest; the lowest loss-ceiling for B, using a single pure strategy, is obtained by choosing the strategy column in which the greatest entry is least. These two values are equal only when there is an entry in the payoff matrix that is, at the same time, the least in its row and the greatest in its column. Such an entry is called a saddlepoint (*e.g.*, see Table 1, case (3), in which the entry 3 is a saddlepoint).

Mixed strategies; minimax theorem. To close the gap between max-min and min-max in the general case in which the payoff matrix contains no saddlepoint, it is necessary to broaden the concept of strategy. Although the decisive innovation of weighted averages of pure strategies occurred in 1713, it was not exploited fully until it was rediscovered by von Neumann. In the broadened concept the frequencies with which a player uses his pure strategies in the long run are specified by weights (some but not all of which may be zero). Operationally, the selection of the pure strategy to be used in a particular playing of the game is left to a suitable chance mechanism, which selects in a random manner from among the pure strategies in accordance with the weights the player has chosen (*e.g.*, in matching pennies, as ordinarily played, the pure strategy of showing heads uniformly and that of showing tails uniformly are randomized with equal weights by tossing the coin before showing it). This randomization by a chance mechanism avoids a mixture pattern from which the opponent might profit in the long run (*e.g.*, to alternate uniformly between heads and tails in matching pennies would weight them equally but be quite unwise in repeated play). Such a mixture of pure strategies, randomized in fixed proportions, is called a mixed strategy. It is clear that neither player's position is worsened by using mixed strategies, because any pure strategy is still tenable as a mixed strategy by assigning nonzero weight only to it.

Definition of mixed strategy

(1) A's gain-floor = max-min ≤ min-max = B's loss-ceiling

(2) max-min = min-max

When A's gain-floor and B's loss-ceiling are re-evaluated in terms of mixed strategies, it turns out that in any matrix game the maximum of minima and the minimum of maxima are equal (see 2). This minimax theorem of von Neumann (1928) is the main theorem of normalized finite dual games and the keystone of the whole theory. It shows that any such game has a solution consisting of (1) a minimax value (the common max-min and min-max), (2) an optimal mixed strategy for A that assures him a gain of at least the minimax value on the average, and (3) an optimal mixed strategy for B that insures him against a loss on the average of more than the minimax value.

Examples of matrix games. A and B choose whole numbers independently; A wins if the sum is odd, B wins if the sum is even. Roman numerals are used throughout in the tables to denote the numbers from which A and B make their choices. The first case shown in Table 1 is a rudimentary dual game; it is the same as matching pennies, with B winning if head (I) matches head (I) or tail (II) matches tail (II), and A winning otherwise. The other four cases are variants that suggest the range of possibilities that arise even in small-scale dual games. Directly to the right of an A choice and below a B choice there appears the resulting payoff, positive if A wins (and B loses), negative if A loses (and B wins). In the rudimentary first case the amount of the payoff is simply one unit, but throughout the other four cases in Table 1 the amount of the payoff is the sum of the numbers chosen by A and B; *e.g.*, if A chooses I while B chooses III, as in cases (3), (4), and (5), the payoff entry is −4 because 1 + 3 = 4, an even number, lost by A, won by B. The vertical column of weights listed for A in each case combines with a parallel column of payoffs to form the weighted payoff sum shown at the foot of that column; *e.g.*, in case (5), column I, 0(−2) + 15(−4) + 13(5) = 5 is the weighted payoff sum. The horizontal row of weights listed for B combines with a parallel row of payoffs to form the weighted payoff sum shown at the right end of that row. These weighted sums become weighted averages when divided by the sum of the weights, written (with ÷) at the lower right. By randomizing their choices with the particular weights listed, A and B achieve optimal strategies; on the average, A is assured of winning at least the minimax value shown and B against losing more than this value. Each solution shown in Table 1 may be verified by observing that the minimax value is both the minimum of the A-weighted average payoffs and the maximum of the B-weighted average payoffs. In case (3), the payoff matrix has a saddlepoint, viz., the entry 3.

How optimal strategies are achieved

Symmetric games. A and B each show one or two fingers, at the same time calling aloud the numbers two, three, or four in an attempt to guess the total number of fingers shown; a right call wins from a wrong call—otherwise it is a draw. This is a two-finger form of an ancient game known as morra played today in Italy. In the rudi-

mentary first version shown in Table 2, the amount of the payoff is simply one unit for a win and zero for a draw; in the somewhat more sophisticated second version the amount of a winning payoff equals the total number of fingers shown by A and B (*e.g.*, if A shows I and calls II, whereas B shows II and calls III, the payoff entry in the second version is −3, because the total number of fingers shown by A and B is $1 + 2 = 3$ and hence A's call is wrong and B's right). In either version the game is fully symmetric: both players face identical choices and rewards. This is partly disguised by the tabulation, which presents the payoffs from the viewpoint of A. To interchange A with B, not only must horizontal payoff rows be transposed into vertical columns but also the sign of each payoff must be changed (to B's viewpoint). The net result of such a twofold shift leaves the payoff arrays in Table 2 just as they are—which is the characteristic property of the tabulation of a fully symmetric dual game. As a result of the symmetry, any optimal strategy for one player must also be optimal for the other and the minimax value must be zero. Consequently, in each of the above versions of morra, the listings of weights for A and B that achieve the zero minimax could be reversed, and this shows that both players have more than one optimal strategy. Actually, in the first version, any weights p, q, q, p yield an optimal strategy; and in the second version, any weights $0, p, q, 0$ yield an optimal strategy, provided $p/(p + q)$ is a fraction between 20/35 and 21/35.

Computation of solutions. The solution of a matrix game with m columns and n rows of payoff entries amounts to the solution of a system of linear inequalities in $m + n + 1$ unknowns, A's m weights, B's n weights, and the minimax value. The system comprises n inequalities that express the A-weighted average payoffs as equal to or greater than the minimax value and m inequalities that express the B-weighted average payoffs as equal to or less than the minimax value; in addition, the $m + n$ weights must be non-negative, reduced usually to fractional form with unit sum for each set, so that the weighted payoff sums become weighted averages without division. It has been shown that any basic solution of such a system is characterized as the unique solution of a suitably chosen subsystem of linear equations. As a result, the solution (if unique) or the full set of solutions (if more than one exist) can be determined by at most a finite number of arithmetic operations, but this number increases so rapidly, as the numbers m and n of rows and columns increase, that the solution of a large-scale matrix game by this method is formidable or impracticable.

The solution of a pair of dual linear programs, which call for maximizing and minimizing linear functions of several variables subject to constraints given by linear inequalities or equations, is equivalent to the solution of a matrix game, and vice versa. Hence, the simplex method devised for linear programming can be applied to solve reasonably sized matrix games. Iterative procedures for approximating game solutions have been formulated to exploit high-speed computers.

Games on a square. In a matrix game, a finite array of payoffs may be replaced by an infinite array of payoffs,

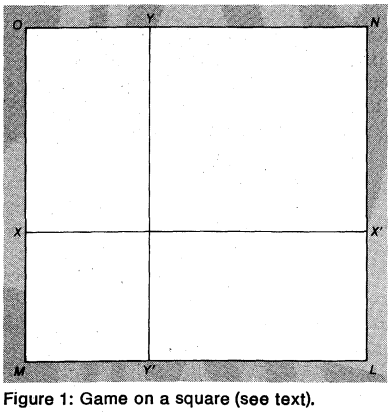


Figure 1: Game on a square (see text).

(3) $P = 4x + 2y + y^2 - x^2 - 4xy - 2$

(4) $P = 0.2 + (y - 0.6)^2$ if $x = 0.8$ and
$P = 0.2 - (x - 0.8)^2$ if $y = 0.6$

one for each point of a square $OMLN$ (see Figure 1). A may choose a point X in the edge OM and, independently, B may choose a point Y in the edge ON; then A gets from B the payoff attached to the point of intersection of the lines XX' and YY' perpendicular to OM and ON. This is an infinite dual game on a square. Figure 2 presents an example: A and B divide the area $OMJLKN$, formed by the square $OMLN$ and two isosceles right triangles MJL and LKN, by means of the lines XX'' and YY'' based on independent choices by A and B of X in OM and Y in ON; A gets the unshaded area, B the shaded area. By direct calculation, taking $OM = ON =$ one unit and $OX = x$, $OY = y$, the net payoff (A's area minus B's area) is found to be a quadratic in x and y (see 3); that is, an algebraic expression in which x and y are at most squared. A wishes to maximize the payoff P; B to minimize P (that is, to maximize $-P$).

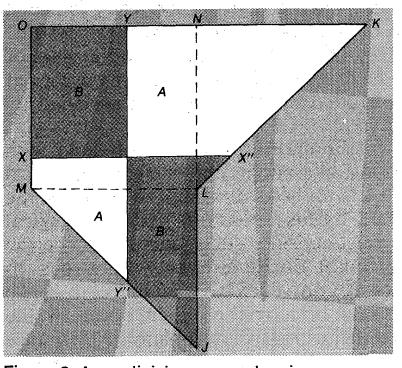


Figure 2: Area-division game showing saddlepoint division (see text).

By analogy with a matrix game, a saddlepoint for a game on a square is at the intersection of a line XX' and a line YY', for which the payoff at the point of intersection is equal to or less than every other payoff on XX', and is equal to or greater than every other payoff on YY' (see Figure 1). (A saddlepoint is so called because it appears as an actual surface saddlepoint in the three-dimensional graph of the payoff over the square.) In the area-division game (Figure 2) $x = 0.8$, $y = 0.6$ give a saddlepoint having payoff 0.2. The net payoff can be rearranged (see 4) to show that by choosing $x = 0.8$, A assures himself at least 0.2, whereas, by choosing $y = 0.6$, B guards against losing more than 0.2. The claim that this point is a saddlepoint can also be verified by geometric reasoning, based on the fact that it is the critical position at which XX'' and YY'' bisect one another (in Figure 2). The saddlepoint

The solution in pure strategies possessed by the area-division game is exceptional, because games on a square, like matrix games, do not usually possess saddlepoints. In general, each player must weight the points in his interval, OM or ON, if he is to do the best he can for himself. Such distributions of weights over an interval are known as probability distributions; they play a role analogous to that of mixed strategies in matrix games. Yet, even when both players employ probability distributions, the minimax theorem does not hold without some mild restriction on how the payoffs vary over the square. For the continuous case of payoffs without abrupt jumps, the equality of max-min and min-max was established in 1938.

EXTENSIVE GAMES—A POKER EXAMPLE

Pure strategies. The primary objective of the study of a game in extensive form is the analysis of the combinatorial structure defined by the rules and its influence on the subsequent definition of strategies. This structure can

Table 3: Simplified Poker in Extensive Form

deal → A to {see →±4,0; raise →B to {fold → 4; see →±6,0; raise →A to {fold →−6; see →±7,0}}}

1st stage	2nd stage	3rd stage	4th stage
equally likely: H4?4 H → A → {H4H4 → 0*; ?6H4 H → B → {X6X4 → 4*; H6H6 → 0*; H6?7 H → A → {X6X7 → −6*; H7H7 → 0}}			
H4?4 L → A → {H4L4 → 4*; ?6L4 H → B → {X6X4 → 4; H6L6 → 6; H6?7 L → A → {X6X7 → −6*; H7L7 → 7}}			
L4?4 H → A → {L4H4 → −4; ?6H4 L → B → {X6X4 → 4*; L6H6 → −6*; L6?7 H → A → {X6X7 → −6; L7H7 → −7}}			
L4?4 L → A → {L4L4 → 0; ?6L4 L → B → {X6X4 → 4; L6L6 → 0; L6?7 L → A → {X6X7 → −6; L7L7 → 0}}			

*Indicates irrational terminal situation.

be presented in the form of a branching diagram (of the sort exhibited in Table 3) in which the junctures represent occasions at which a player or a chance mechanism (specified by the rules, such as cards, dice, etc.) is called upon to select one of the several branches that continue the course of the play. These branches contain the state of information of the next player. If the branching ceases at any stage, the play is over and the payments specified by the rules are listed.

Several junctures in the diagram may be prefixed with the same information for a single player (*e.g.*, in a card game, these may be situations that differ only in the unrevealed cards held by his opponents). For strategical purposes, a player must consider such junctures to be the same, because he cannot distinguish between two such junctures and hence cannot decide in advance to make distinct choices should one or the other occur in the course of play. This means that the problem of formulating a pure strategy for a player is solved by listing all of his junctures with distinct states of information and then making a definite choice of a branch for each. Such a plan fulfills the requirements for a pure strategy set forth previously.

Historically, the first result on games in extensive form was stated (for the game of chess) by Ernst Zermelo and given a complete proof by von Neumann. It applies to games with perfect information; that is, games in which a player is always informed of the complete previous history of the play. The theorem asserts that finite dual games with perfect information can always be solved by pure strategies without randomization. In other words, there is always a saddlepoint in the payoff matrix of such games. As regards chess, in particular, this means that exactly one of the following three alternatives is valid: (1) White has a pure strategy that wins, no matter what Black does; (2) both players have pure strategies that ensure at least a draw, no matter what the other does; (3) Black has a pure strategy that wins, no matter what White does. The theory gives no practical method, however, for deciding which assertion is true.

The three alternatives in chess

Behaviour strategies. The lack of a solution in pure strategies for a general finite dual game means that a player must randomize his actions in some manner to obtain the amount due him (as measured by the minimax value). In the extensive form of a game, it is possible to introduce a method, alternative to that of mixed strategies, for varying his choices from contest to contest of the game in an undecipherable pattern. This method specifies by weights the long-run frequencies with which a player chooses the various alternatives presented to him in a given state of information. Rather than using a chance mechanism to select a complete plan before the playing of a game, a player employs such a device throughout the play, at each occasion for choice, to select an alternative in accordance with weights he has chosen. Such a decentralized system of play, randomizing the choices for each distinct state of information in fixed proportions, is called a behaviour strategy, because these weights are the statistics of a player's behaviour that an external observer could gather from a long series of contests.

Because the number of weights needed to specify a behaviour strategy is, in general, much smaller than the number of weights in a mixed strategy, it is useful to know when they produce the same results. It has been shown that a player can use either with equal success in any game, with perfect recall; *i.e.*, any game in which a player's state of information always includes everything he has done or known previously. Poker is a game with perfect recall and hence can be played effectively with behaviour strategies. Bridge, on the other hand, does not have perfect recall because each player is a pair of persons, neither of whom is informed of the other's hand. Because there can be no correlation between the selections made by the chance devices for different states of information when a player employs a behaviour strategy, it is possible that bridge players must use mixed strategies to achieve the minimax value.

Recall in poker and bridge

Simplified poker. A simplified version of poker is played by A and B with a deck containing a large number of cards marked high (*H*) and low (*L*). First, each player provides ("antes") four chips and is dealt a single card; the four possible deals are assumed to be equally likely. Then A has two options: to "see," or to "raise" by adding two chips to the pot. If A decides to see, the higher hand wins the pot and equal hands split the pot equally. If A decides to raise, B has three options: to "fold," to see by adding two chips to the pot, or to raise by adding three chips to the pot. If B folds, A wins the pot (with no hands revealed). If B sees, the higher hand wins the pot and equal hands split the pot equally. If B raises, A has two options: to fold, or to see by adding one chip to the pot. If A folds, B wins the pot (with no hands revealed). If A sees, the higher hand wins the pot and equal hands split the pot equally.

The possible sequences of choices by the players, independent of the actual cards held, are shown at the top of Table 3. The player to act precedes, and his alternatives follow, each juncture. The possible payoffs (listed as A's gain, because this is a dual game) are given when the play is over. The precise payoff that occurs is determined by the cards held. In the body of Table 3 appears the complete branching diagram, taking account of the deal. Four-place symbols present the states of information. The first two symbols give the hand and current bet of A, the last two symbols give the hand and current bet of B. All are given as known to the player who is next to act; *e.g.*, the symbol ?6*L*4 preceding B means that B is to act, not knowing A's card, but knowing that A has bet six chips, and B holds a low card and has bet four chips. For ready reference, the actual card held is listed below each question mark. Whenever the play ends, the payoff to A is listed; if either player has folded, the symbol *X* is used to indicate that neither hand need be revealed.

Payoffs in poker

Certain terminal situations (starred in the Table) will never occur in rational play; *e.g.*, player A will never see in the second stage while holding a high card because he stands to win at least as much (and sometimes more) by raising. Similar arguments advise both players to bet the limit at every juncture in which they hold a high card. Reduced in this manner, a pure strategy for either player consists merely of how high he should carry his bet on a low card. These pure strategies are listed as two-place symbols to the left and above the payoff matrix in the left half of Table 4. The first number (7) gives the limit a player will bet on a high card, the second (4, 6, or 7) gives the limit on a low card. To compute the payoff to A when two of these strategies are played, an average is taken over the four terminal situations that can follow the four possible deals; *e.g.*, when A plays 76 against 74 for B, the play ends in *H*7*H*7, *X*6*X*4, *X*6*X*7, or *X*6*X*4 and the corresponding entry is $\frac{1}{4}(0) + \frac{1}{4}(4) + \frac{1}{4}(-6) + \frac{1}{4}(4) = 2/4$. (The unreduced payoff matrix would consist of nine rows and nine columns, labelled 44, 46, 47, 64, 66, 67, 74, 76, 77.)

Computing payoffs

The weights listed are the only weights that achieve the

Table 4: Simplified Poker
(optimal mixed strategies and optimal behaviour strategies)

wt.		9	1	2	
	HL	74	76	77	B
6	74	0	$\frac{2}{4}$	$\frac{3}{4}$	2
2	76	$\frac{2}{4}$	0	$-\frac{5}{4}$	2
4	77	$\frac{1}{4}$	$-\frac{1}{4}$	0	2
	A	2	2	2	÷12
minimax = $\frac{1}{6}$					

player	information	option	weight
A	*L*4?4	see	6
		raise	6
	*L*6?7	fold	2
		see	4
B	?6*L*4	fold	9
		see	1
		raise	2

minimax value. Player A must play bluffing strategies (either 76 or 77) one-half of the time, whereas B must decide to bluff one-fourth of the time.

The average behaviour imposed by these optimal mixed strategies upon a player with a given state of information constitutes an optimal behaviour strategy. The optimal weights, so defined, are given in the right half of Table 4 (*e.g.*, with information *L*4?4, player A assigns the seeing pure strategy 74 the weight 6 and assigns the raising pure strategies 76 and 77 the weights 2 + 4 = 6; in common parlance this means that, in the long run, A sees as often as he raises upon being dealt a low card).

PLURAL GAMES

Economic example: noncooperative solutions. Two basically different approaches to the behaviour of parties to plural games have been proposed. The solutions that they suggest are contrasted in the following (highly simplified) problem from economics: Two sellers, A and B, in competition with each other, are contemplating a price cut of the same amount. Each knows that if they both sell at the same price they will divide the market evenly, but if one cuts his price and the other does not, then the former will capture the entire market. If the total profits at the higher and lower prices are known, this situation gives rise to a symmetric two-party plural game. The top part of Table 5 lists the payoffs for four of these games; in all cases, six units of profit are available at the higher price, whereas the lower price, in the four cases, yields four, two, zero, and minus two units of profit, respectively (negative profit is positive loss). The profits of A and B in each of the four games are shown to the right of each pair of prices (*i.e.*, pure strategies).

Table 5: Price Cut Example

prices		profits							
		(1)		(2)		(3)		(4)	
A	B	A	B	A	B	A	B	A	B
high	high	3	3	3*	3*	3*	3*	3*	3*
high	low	0	4	0	2	0	0	0	−2
low	high	4	0	2	0	0	0	−2	0
low	low	2*	2*	1*	1*	0*	0*	−1	−1
coalition		maximum assured profit							
A with B		6		6		6		6	
A alone		2		1		0		0	
B alone		2		1		0		0	

*Indicates solution.

As a solution to the sellers' problem, the noncooperative theory of John Nash suggests and proves the existence of a pattern of independent action in which there is no incentive for deviation by any player alone. As such, it is an extension of the minimax theory for dual games. A price cut by each seller is an equilibrium point in each of the first three games, and if both keep the higher price, the result is an equilibrium point in the last three games. These solutions (starred in the top part of Table 5) can be checked by verifying that unilateral changes in strategy are not rewarded. In the first game, the stability of the solution derives from the fact that a lone decision by either party to sell high would decrease his profit from two units to nothing. Although it is clear in this case that both sellers would benefit from an agreement binding each other to the higher price, this pattern is considered unstable because either seller could increase his profit by unilateral action.

In like manner, a cut by a single seller can be considered a (fictitious) coalition between that seller and the buyers of the market but can be upset by lone action by the other seller to his advantage. Indeed, the dilemma posed by the first game consists of the instability, with respect to unilateral defections, of each of the three possible two-party coalitions that may form among the two sellers and the market.

Cooperative solutions. The cooperative theory of von Neumann and Morgenstern for plural games (their general *n*-person games) is built upon these coalitions and introduces the possibility that the parties to an agreement may distribute their payments so as to maintain the coalition. The bottom part of Table 5 continues the analysis of the sellers' problem from this point of view. It lists for each of the four games the maximum assured profit for A and B acting together, for A alone and for B alone (*e.g.*, the greatest profit A can assure himself in the fourth game is zero; to do this he holds to the higher price). The cooperative theory suggests, for all four games, that (1) the sellers agree to hold to the higher price; (2) that each receive at least as much from the coalition as by independent action; (3) that the remainder be distributed in some manner between them (*e.g.*, in the second game, each seller surely receives one unit and the four additional units are divided in an unspecified manner). No mode of distribution is singled out, because even though one seller may prefer one division to another, he cannot enforce it by a unilateral change of his strategy or his allegiance. On the other hand, every other distribution of the profits is less advantageous to the collective interest of the two active parties, the sellers. In general, the stability of a cooperative solution derives from these two sources: internal consistency and external domination.

Stability of a cooperative solution

Simple games; a power index. The cooperative theory has a political aspect when the only prospects of a coalition are to win or lose; this is the case in voting systems, such as legislatures, committees, and voting stockholders. Such games are called simple, and the only two possible outcomes, win or lose, for the coalitions are indicated numerically by one and zero. An a priori numerical indicator of the relative equities of the players of any cooperative game has been proposed that is easily calculated for such a situation. This power index measures the average contribution of a player to the coalitions to which he might belong, taking account of the order in which the members join the coalition; *e.g.*, a participant in a simple game contributes to a coalition only if his entrance changes it from a losing to a winning coalition.

A representative example of this class is a four-man committee composed of a chairman, with three votes, and three other members, with one vote each. A simple majority of votes win, whereas ties are considered as a loss for both sides. The 24 possible sequences in which the members might vote are listed in Table 6; the chairman is denoted by A and the three ordinary members by b, c, and d. The pivotal player, the first to complete a majority in each voting order, is starred in each column. Because the chairman is pivotal three-fourths of the time, his power index is 75 percent. The ordinary members share the remaining fourth equally and have as indices 8⅓ percent. Thus the voting ratio of 3:1 results in a power ratio of 9:1 in this simple game. (A.W.T./H.W.K.)

GAME PLAYING PROGRAMS

Many attempts have been made at writing programs for playing games on a computer. Although, in general, the programs are not based purely on the theory of games, they use many of the techniques developed in this field.

Table 6: Committee Voting Example

voting	24 possible orders																							
First	A	A	A	A	A	A	b	b	c	c	d	d	b	b	c	c	d	d	b	b	c	c	d	d
Second	b*	b*	c*	c*	d*	d*	A*	A*	A*	A*	A*	A*	c	d	b	d	b	c	c	d	b	d	b	c
Third	c	d	b	d	b	c	c	d	b	d	b	c	A*	A*	A*	A*	A*	A*	d	c	d	b	c	b
Fourth	d	c	d	b	c	b	d	c	d	b	c	b	d	c	d	b	c	b	A*	A*	A*	A*	A*	A*

*Indicates pivotal player.

Invoking the minimax principle

Board games. The usual approach is to construct a search tree of possible moves to a given depth and to evaluate the board positions so obtained using an evaluation function based on the programmer's own knowledge of the game. The minimax principle is used to choose the next best move. For example, in Figure 3 the search tree is a depth three. Each node represents a board position and the board positions at level three are all evaluated. A position deemed beneficial to the computer is denoted by a positive score, the larger the score the better the position. If it is the computer's move next, then the move taken is that shown in the Figure. For most board games, the search space is extremely large. Various methods of reducing the search space of possible moves are used in order to keep the problem to manageable proportions. The technique of pruning was introduced in one of the early famous chess playing programs, and this technique was shown to reduce search space considerably. The program attains an acceptable level, about the standard of an average-to-good player. A successful checker (British, draughts) playing program has been written, one that is of world championship standard. In 1969, an article was published concerning playing the game of *go* on a computer. A *go* board has a grid of 19 × 19 lines and thus there is a search space of about 10^{761} nodes. The author suggests a new approach based on applying a battery of basic principles of play. Together with a simple "look-ahead," the recommendations give good positional play and avoid tactical capture traps that could lead to the loss of a stone or two.

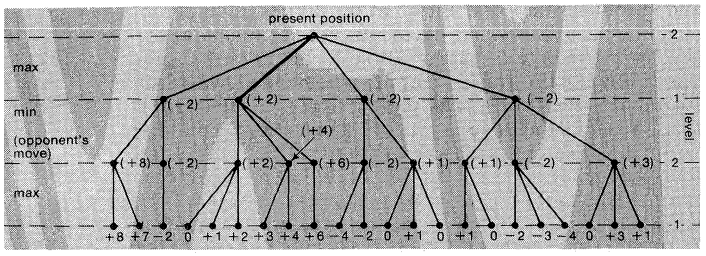


Figure 3: Search tree of depth three, showing minimax principle. The best next move after the present position is indicated by the heavy line (see text).

Other games. Straightforward programs have also been written to play card games (bridge, poker, etc.), three dimensional tic-tac-toe (British, noughts and crosses) and *kalah*. (Ed.)

II. Linear and nonlinear programming (mathematical programming)

GENERAL OBSERVATIONS

Mathematical programming may be described in terms of its mathematical structure and computational procedures or in terms of the broad class of important decision problems which can be formulated as the minimization (maximization) of a function of several variables that are subject to a system of side constraints. For example, a linear program is defined as the minimization of a linear "objective" function whose variables satisfy a system of linear inequalities.

The scope of mathematical programming

In practice, mathematical programming usually refers to such linear programs, the general study of nonlinear programs (those in which either the objective function or at least one of the constraint functions is nonlinear), integer programs (that is, linear programs with the additional restriction that some or all of the variables must be integer valued), stochastic programs (those programs involving random variables), and network flow theory (transportation or flow through networks). As such, mathematical programming overlaps, has contributed to, and has been influenced by operations research, mathematical economics, control theory, dynamic programming, and combinatorial theory.

The term programming had its origin in programming (*i.e.*, planning, scheduling) the quantity and timing of the various activities of an organization such as a factory, an airline, the defense establishment, the national economy, or world trade. (It is not to be confused with "programming" as used for the task of preparing a sequence of instructions for a computer.) The goal is to find an optimum schedule.

A simple example, "the assignment problem," illustrates the essential difficulty. A factory has 70 men with different qualifications and it is desirable to assign them to 70 jobs. If a "value" can be attached to assigning a particular man to a particular job, then the problem becomes one of selecting out of 70! (which is the product of integers from 1 to 70) possible ways of permuting the assignments the one that yields the maximum total value to the factory. Because 70! is approximately 10^{100}, it would take an electronic computer executing 1,000,000 operations per second over 10^{87} years (or many times the projected life of the universe) to examine all the permutations. Such decision problems are common and have resulted in the development of clever formulation of the mathematical models, powerful mathematical methods of solution, and efficient computer algorithms (step-by-step procedures). The solution of linear programs is said to constitute 25 percent of all the time used by computers to solve scientific problems.

ORIGINS AND INFLUENCES

Although widely used now to solve everyday decision problems, linear programming was comparatively unknown before 1947. No work of any significance was carried out before this date, even though the French mathematician Jean-Baptiste-Joseph Fourier seemed to be aware of the subject's potential as early as 1823. A Russian mathematician, Leonid Vitalevich Kantorovich, who published an extensive monograph in 1939 *Matematicheskie metody organizatsi i planirovaniya proizvodstva* ("Mathematical Methods for Organization and Planning of Production"), is credited with being the first to recognize that certain important broad classes of scheduling problems had well-defined mathematical structures. Unfortunately his proposals remained unknown both in the Soviet Union and elsewhere for nearly two decades. Meanwhile, linear programming had developed considerably in the United States and Western Europe. In the period following World War II, officials in the United States government felt that efficient coordination of the energies of a whole nation in the advent of atomic war would require the use of scientific planning techniques. The advent of the computer made such an approach feasible.

The first linear programming model

Intensive work began in 1947 in the United States Air Force. The linear programming model was proposed because it is simple, practical, and yet provided a sufficiently general framework for representing interdependent activities that must share scarce resources. The system (*e.g.*, national economy) is viewed as made up of various activities (*e.g.*, production, training, shipping, disposal, and storing). Each activity is assumed to require a flow of inputs and outputs of various types of items (*e.g.*, men, materials) proportional to the level of the activity. Activ-

ity levels are assumed to be representable in terms of positive numbers or zero. The revolutionary feature of the approach, however, was the expressing of the goal, as that of minimizing (maximizing) an objective function (*e.g.*, maximizing sorties in the case of the air force or maximizing profits in industry). Before 1947 all practical planning was characterized by a series of authoritatively imposed rules of procedure and priorities. General objectives were never stated, probably because of the impossibility of performing the calculations necessary to minimize an objective function under constraints. In 1947 a method composed of successive tests for optimality at extreme points and intervening linear movement along polygonal edges called the simplex computational method was introduced which turned out to be indeed efficient. Interest in linear programming grew rapidly and by 1951 its use spread to industry. It is almost impossible to name an industry in the early 1970s that is not using mathematical programming in some form, although its use varies greatly, even within the same industry.

The current interest in linear programming by economists appears to be an anachronism. The French economist François Quesnay's attempt in his *Tableau économique* (1758) to interrelate the role of the landlord, the peasant, and the artisan was a crude example of a linear programming model. Léon Walras, another French economist, in 1874 proposed a sophisticated approach that had as part of its structure fixed technological coefficients (as assumed in linear programs). Oddly enough, until the 1930s, the linear-type model was little exploited.

Von Neumann in a 1937 paper analyzed a steadily expanding economy based on alternative methods of production and fixed technological coefficients. As far as mathematical history is concerned, the study of linear inequality systems excited virtually no interest before 1936. In 1911 a vertex-to-vertex movement along edges of a polyhedron (as is done in the simplex method) was suggested as a way to solve a problem that involved optimization, and in 1941 movement along edges was proposed for a problem involving transportation. Credit for laying much of the mathematical foundations should probably go to von Neumann. In 1928 he published his famous paper on *Game theory*. In 1947 he conjectured the equivalence of linear programs and matrix games, introduced the important concept of duality, and made several proposals for the numerical solution of linear programming and game problems. Serious interest by other mathematicians began in 1948 with the rigorous development of duality and related matters.

The general simplex method, already mentioned (see above ***Origins and influences***) and to be discussed in detail (see below *The simplex method*), was first programmed in 1951 for the United States Bureau of Standards SEAC computer. Starting in 1952, the simplex method was programmed for use on various models of IBM (International Business Machines) computers and later for those of other companies. These programs turned out to be practical. As a result, commercial applications of linear programs in industry and government grew rapidly. New computational techniques and variations of older techniques continued to be developed.

At the present time there is much interest in solving large linear programs with special structures, for example, corporate models and national planning models that are multistaged, dynamic, and exhibit a hierarchical structure. It is estimated that certain underdeveloped countries have the potential of increasing their GNP (gross national product) anywhere from 10 percent to 15 percent per year if detailed growth models of the economy could be constructed, optimized, and implemented.

LINEAR PROGRAMMING THEORY

Linear programming in factory production

Basic ideas. A simple problem in linear programming is one in which it is necessary to find the maximum (or minimum) value of a simple function, such as $x_1 + 2x_2$, subject to certain constraints. An example of its application might be that of a factory producing two commodities. In any production run, it produces x_1 of the first type and x_2 of the second. If the profit on the second type is twice that on the first, then $x_1 + 2x_2$ represents the total profit. The function $x_1 + 2x_2$ is known as the objective function.

Clearly the profit will be highest if the factory devotes its entire production capacity to making the second type of commodity. In a practical situation, however, this may not be possible; a set of constraints is introduced by such factors as availability of machine time, manpower, or raw materials. For example, if the second type of commodity requires a raw material that is limited so that no more than five can be made in any batch, then x_2 must be less than or equal to five; *i.e.*, $x_2 \leq 5$. If the first commodity requires another type of material limiting it to eight per batch then $x_1 \leq 8$. If 1 and 2 take equal times to make and the machine time available allows a maximum of ten to be made in a batch, then $x_1 + x_2$ must be less than or equal to 10; *i.e.*, $x_1 + x_2 \leq 10$.

Two other constraints are that x_1 and x_2 must each be greater than or equal to zero, because it is impossible to make a negative number of either; *i.e.*, $x_1 \geq 0$ and $x_2 \geq 0$. These conditions result in a simple linear program that can be stated in concise algebraic form. The problem is to find the values of x_1 and x_2 for which the profit is a maximum. Any solution can be denoted by a pair of numbers (x_1, x_2); for example, if $x_1 = 3$ and $x_2 = 6$, the solution is (3, 6). These numbers can be represented by points plotted on two axes; in Figure 4

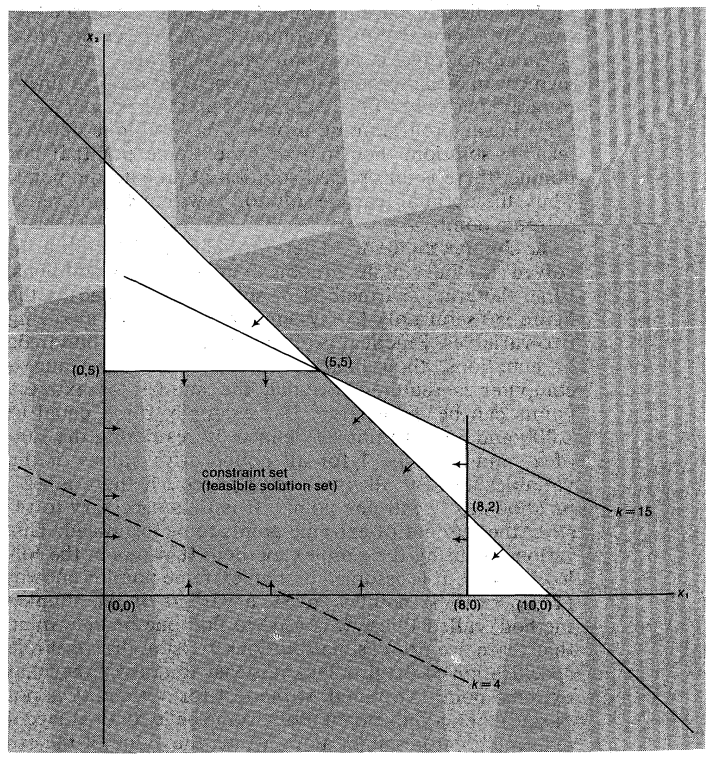


Figure 4: Constraint set bounded by the five lines $x_1 = 0$, $x_2 = 0$, $x_1 = 8$, $x_2 = 5$, and $x_1 + x_2 = 10$. These enclose an infinite number of points that represent feasible solutions.

the distance along the horizontal axis represents x_1 and that along the vertical represents x_2. An infinite number of points exist corresponding to an infinite number of solutions, but because of the constraints the feasible solutions must lie within a certain region. For example, the constraint $x_1 \geq 0$ means that points representing feasible solutions lie on or to the right of the x_2 axis. The constraint $x_2 \geq 0$ means that they also lie on or above the x_1 axis. Application of all the constraints leads to the result that points representing solutions lie within the shaded area bounded by a polygon formed by intersection of the lines $x_1 = 0$, $x_2 = 0$, $x_1 = 8$, $x_2 = 5$, and $x_1 + x_2 = 10$. These form the constraint set. For example, to make three items of commodity 1 and four of 2 is a feasible solution since the point (3, 4) lies in this region.

To find the best solution, the objective function $x_1 + x_2 = k$ is plotted on the graph for some value of k, say $k = 4$. This value is indicated by the broken line in Figure 4. As k is increased, a family of parallel lines are produced and the line for $k = 15$ just touches the constraint set at the point (5, 5). If k is increased further, the values of x_1 and x_2 lie outside the set of feasible solutions. Thus the best solution is that in which equal quantities of each commodity are made. The points (0, 0), (0, 5), (5, 5), (8, 2), and (8, 0) are the extreme points of the constraint set, and the problem involves finding the extreme point the coordinates of which yield the largest value for k.

The standard form of a linear program is expressed as an instruction to maximize (or minimize) a linear expression in n nonnegative variables, $\{x_i\}$, subject to m linear equations (see 5). In applications, the equations correspond to material balance of various items and the variables to levels of various activities (*e.g.*, production processes). Any system of linear inequalities can be reduced to this standard form by simple substitutions.

Definition of a feasible solution

A set of x_j values that is nonnegative and satisfies the equations is a "feasible" solution and corresponds to a feasible but not necessarily optimal program. Such a solution is viewed as a point in n-dimensional space.

The constraint set—*i.e.*, the set of feasible solutions—has a property known as convexity. In general, a set S is convex if and only if it contains the entire line segment between any two of its elements. An element of a convex set S is called an extreme point if it does not lie on the line segment joining two other points of S. A line segment in S is called an edge if no point on it lies also on a line segment in S that crosses it. The constraint set of a linear program has only a finite number of extreme points.

If a linear programming problem has a unique optimal feasible solution, then it is at an extreme point; if not unique, there is an optimal solution at an extreme point. Thus, the optimal feasible solution is within a finite set of extreme points.

The simplex method. There are certain difficulties involved in the simple graphical approach to solutions. First, the graphical method of solution illustrated by the figure is useful only for systems of inequalities involving two variables. Practical problems often involve hundreds of equations, thousands of variables, however, and a computer is required. Second, the number of extreme points can be quite large. For example, for n equal to 2,000 and a constraint set defined by restricting the sum of x_j and y_j to be 1 for all nonnegative values of the variables and for all n values of j (see 6), there are 2^n, or about 10^{600}, extreme points. It is thus necessary to restrict the number of extreme points to be examined. This is done by using the simplex method. It works in the following way. It is assumed that an extreme point is known. (If no extreme point is given, a variant of the simplex method, called Phase I, is used to find one or determine that there are no feasible solutions.) Using the algebraic specification of the problem, it is easy to test whether that extreme point is optimal. If the test for optimality is not passed, then a movement along some edge to an adjacent extreme point is sought along which the value of the objective function increases at the fastest rate. Sometimes one can move along an edge and make the objective function value increase without bound. If this occurs, the procedure terminates with a prescription of the edge along which the objective goes to positive infinity. If not, a new extreme point is reached having at least as high an objective function value as its predecessor. The sequence described is then repeated. Termination occurs when an optimal extreme point is found or the unbounded case occurs.

The simplex method solves the numerical example given above in the following way: The problem is put into canonical form by converting the linear inequalities into equations by introducing slack variables $x_3 \geq 0$, $x_4 \geq 0$, $x_5 \geq 0$, and the variable x_0 for the value of the objective function; the problem may then be restated as that of finding nonnegative quantities $x_1, \cdots, x_5$ and the largest possible x_0 satisfying the resulting equations (see 7). One

Slack variables

(5)
$$\begin{cases} \text{maximize } p_1x_1 + \cdots + p_nx_n \\ \text{subject to } a_{11}x_1 + \cdots + a_{1n}x_n = b_1 \geq 0 \\ \qquad a_{21}x_1 + \cdots + a_{2n}x_n = b_2 \geq 0 \\ \qquad \vdots \qquad\qquad \vdots \\ \qquad a_{m1}x_1 + \cdots + a_{mn}x_m = b_m \geq 0 \\ \qquad x_1 \geq 0, \cdots, x_n \geq 0 \end{cases}$$

(6) $x_j + y_j = 1, \quad x_j \geq 0, \quad y_j \geq 0 \quad \text{for } j = 1, \cdots, n$

(7)
$$\begin{cases} -x_0 + x_1 + 2x_2 = 0 \\ x_1 + x_3 = 8 \\ x_2 + x_4 = 5 \\ x_1 + x_2 + x_5 = 10 \end{cases}$$

(8)
$$\begin{cases} x_0 = 0, \ x_1 = 0, \ x_2 = 0, \ x_3 = 8 \\ x_4 = 5, \ x_5 = 10 \end{cases}$$

(9) $(x_0, x_1, x_2, x_3, x_4, x_5) = (10, 0, 5, 8, 0, 5)$

(10)
$$\begin{cases} -x_0 + x_1 - 2x_4 = -10 \\ x_1 + x_3 = 8 \\ x_2 + x_4 = 5 \\ x_1 - x_4 + x_5 = 5 \end{cases}$$

(11) $(x_0, x_1, x_2, x_3, x_4, x_5) = (15, 5, 5, 3, 0, 0)$

(12)
$$\begin{cases} -x_0 - x_4 - x_5 = -15 \\ x_3 + x_4 - x_5 = 3 \\ x_2 + x_4 = 5 \\ x_1 - x_4 + x_5 = 5 \end{cases}$$

(13) $x_4 \geq 0, \ x_5 \geq 0, \ x_0 = 15 - x_4 - x_5 \leq 15$

(14) $x_1 = 5, \ x_2 = 5, \ x_3 = 3, \ x_4 = x_5 = 0$

obvious solution of the resulting system of equations is found by setting $x_1 = 0$, $x_2 = 0$ (see 8), which corresponds to the extreme point at the origin in the figure. If x_1 (or x_2) were increased from zero while the other one x_2 (or x_1) is fixed at zero, the values of x_0, x_3, x_4, x_5 could be made to satisfy the equations, and the objective value x_0 would increase as desired. The variable x_2 produces the largest increase of x_0 per unit change; so it is used first. Its increase is limited by the nonnegativity requirement on the variables. In particular, if x_2 were increased beyond 5, x_4 would become negative.

At $x_2 = 5$, this situation produces a new solution (see 9) and corresponds to the extreme point (0, 5) in the figure. The system of equations is put into equivalent form (see 10) by solving for the variables x_0, x_2, x_3, x_5, which are nonzero in the above solution in terms of those variables now at zero; *i.e.*, x_1 and x_4. It is now apparent that an increase of x_1 while holding x_4 equal to zero will produce a further increase in x_0. The nonnegativity restriction on x_3 prevents x_1 from going beyond 5. The new solution (see 11) corresponds to the extreme point (5, 5) in the figure. Solving for x_0, x_1, x_2, x_3 in terms of the variables x_4, x_5 (which are currently at zero value) yields a final equivalent system of equations (see 12). It is to be noted that no variable can be changed from its present value and yield a feasible solution with a higher value than $x_0 = 15$ (see 13). Hence, an optimal solution is determined (see 14).

The dual problem. Associated with each linear programming problem is a second linear programming problem known as its dual. In this association, the original problem is referred to as the primal. Von Neumann provided a form for the primal problem (see 15) that yields an elegant statement of the dual problem (see 16).

An intimate relationship exists between a linear program and its dual. For any vector x^0 (see ANALYSIS, VECTOR AND TENSOR) satisfying the constraints of the primal and any vector y^0 satisfying the constraints of its dual, the primal objective value is never greater than the dual objective value, namely, $px \leq yb$. One implication of this relationship is that when the dual problem possesses at least one feasible vector, the primal objective function is bounded from above.

If the objective values satisfy $px^0 = y^0b$, then it is easy to show that x^0 and y^0 if feasible are optimal solutions of the primal and dual programs. If a vector x^0 is an optimal solution of the primal problem, there exists an optimal solution of the dual problem such that px^0 equals y^0b. Moreover, if $v^0 = b - Ax^0$ and $u^0 = p - y^0A$, then for optimal primal and dual feasible vectors x^0 and y^0, the relations $y_i{}^0v_i{}^0 = 0$ for $i = 1, \cdots, m$ and $x_j{}^0u_j{}^0 = 0$ for $j = 1, \cdots, n$ hold and are referred to as complementary slackness conditions. Thus, if $x_j{}^0$ is positive in an optimal solution, then the jth inequality constraint of the dual holds as an equality; *i.e.*, its slack, measured by $u_j{}^0$, satisfies $u_j{}^0 = 0$.

Complementary slackness conditions

NONLINEAR PROGRAMMING THEORY

A mathematical programming problem is called nonlinear if the objective function $f(x)$ to be minimized, in which x is a vector, or any of the constraint functions are nonlinear.

Classification of the problems. One broad class of nonlinear programming problems is that concerning minimizing $f(x)$ subject to no constraints (the unconstrained problem); another is the linearly-constrained nonlinear programs that include as a special subclass quadratic programs concerned with the minimization of a quadratic function (see 17) subject to linear constraints. Another is the chemical equilibrium problem in which the Gibbs function, a measure of the free energy of a chemical system (see 18), is to be minimized subject to (linear) mass-balance equations (see 19) and nonnegativity conditions (see 20). The x_j in the problem represent the unknown number of molecules of different types in a system under constant temperature and pressure. In the mass-balance equations the b_i are the given number of atoms of various types; the a_{ij} are the number of atoms of type i in a single molecule of type j. The c_j in the Gibbs function are related to the constants in the law of mass action and are considered as constants only for dilute solutions.

Nonlinear programs are also classified according to whether defining functions have the appropriate convexity property. A function $F(x)$ defined on a convex set S is convex if (and only if) the set of points lying on or above its graph is convex. In analytic terms, convexity is equivalent to an inequality relation (see 21). The function $G(x)$ on S is concave if and only if $F(x) = -G(x)$ is convex. A useful property of any convex function $F(x)$ defined on the convex set S is that the set of all points in S such that $F(x) \leq 0$ is a convex set. (But the set of vectors x, such that $F(x) = 0$, is not necessarily convex.)

When a quadratic function has the property that its quadratic part (see 22) is nonnegative for all choice of values for x_i, x_j, then $Q(x)$ is a convex function. Linear functions are both convex and concave.

If $f(x)$ and $h_1(x), \cdots, h_l(x)$ are convex functions, defined for each vector x appearing as the argument of the functions, then the problem of minimizing $f(x)$ subject to $h_j(x) \leq 0$ $(j = 1, \cdots, l)$ is called the convex programming problem. One important property of convex functions is the sufficiency of local conditions for idenitfying a minimum. If f is a convex function on the convex set S and x^0 is a point of S such that $f(x^0) \leq f(x)$ for all x in S that are sufficiently close to x^0, then

Convex programming problem

(15) PRIMAL: Maximize px subject to $x \geq 0, Ax \leq b$ in which $A = [a_{ij}]$ is an $m \times n$ matrix, $x = (x_1, x_2, \cdots, x_n)^T$, $p = (p_1, p_2, \cdots, p_n)$ and $b = (b_1, b_2, \cdots, b_m)^T$.

(16) DUAL: Minimize yb subject to $y \geq 0$, $yA \geq p$, in which $y = (y_1, y_2, \cdots, y_m)$.

(17) $$Q(x) = \sum_{j=1}^{n} c_jx_j + \tfrac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} d_{ij}x_ix_j$$

(18) $$F(x) = \sum_{j=1}^{n} c_jx_j + \sum_{j=1}^{n} x_j \log\left(x_j \Big/ \sum_{j=1}^{n} x_j\right)$$

(19) $$\sum_{j=1}^{n} a_{ij}x_j = b_i, \qquad i = 1, \cdots, m$$

(20) $$x_j \geq 0, \qquad j = 1, \cdots, n$$

(21) $F(\lambda x + \mu y) \leq \lambda F(x) + \mu F(y)$ for all $x \in S$, $y \in S$ and $\lambda \geq 0$, $\mu \geq 0$ such that $\lambda + \mu = 1$.

(22) $$\sum_{i=1}^{n}\sum_{j=1}^{n} d_{ij}x_ix_j$$

(23) $$\nabla f(x^0) = \left[\frac{\partial f(x^0)}{\partial x_1}, \cdots, \frac{\partial f(x^0)}{\partial x_n}\right] = (0, \cdots, 0)$$

(24) If f and g_i are differentiable and the m gradient vectors $\nabla g_1(x^0), \cdots, \nabla g_m(x^0)$ are independent, then there exist numbers $\lambda_1, \cdots, \lambda_m$ such that the vector relation $\nabla f(x^0) = \lambda_1\nabla g_1(x^0) + \cdots + \lambda_m\nabla g_m(x^0)$ holds. The numbers $\lambda_1, \cdots, \lambda_m$ are known as Lagrange multipliers.

(25) $$L(x, \lambda) = f(x) + \lambda_1 g_1(x) + \cdots + \lambda_m g_m(x)$$

x^0 is also a (global) minimum of f on S; that is, $f(x^0) \leq f(x)$ for all x in S. This property can be exploited in computational procedures aimed at calculating optimal solutions of convex programming problems.

Methods of solution. The question of identifying optimal solutions of nonlinear programs is often discussed assuming that the constituent functions are differentiable. The simplest such case is the unconstrained minimization of a differentiable function $f(x)$. If $x^0 = (x_1, \cdots, x_n)$ is an optimal solution, then the gradient vector of f at x^0 vanishes (see 23). Geometrically, this means that the tangent plane to the graph of f at the point $[x^0, f(x^0)]$ is horizontal. For a convex function f, the vanishing of the gradient is enough to guarantee that the point x^0 is a (global) minimum of f. But in general, the vanishing of the gradient can occur at points that are not even local minima, namely, at saddle points and local maxima.

The equality-constrained minimization problem is of the form: minimize $f(x)$ subject to $g_i(x) = 0$ for $i = 1, \cdots, m$ and $m < n$. If the functions $f, g_1, \cdots, g_m$ are differentiable, the attempt to isolate a (local) minimum point x^0 for this problem by finding numbers $\lambda_1, \cdots, \lambda_m$ such that $\nabla f(x^0) = \sum_{i=1}^{m} \lambda_i \nabla g_i(x^0)$ is called the method of Lagrange multipliers (see 24). In this approach, one looks for unconstrained (local) minima of an associated Lagrangian function (see 25).

When all the constituent functions are differentiable,

there is an analogous result on the necessary conditions of optimality for inequality-constrained minimization problems (see 26). These necessary conditions are closely related to the Kuhn-Tucker conditions. The mathematicians H.W. Kuhn and A.W. Tucker of the United States showed in a now classic, nonlinear programming paper that if a certain regularity condition holds, a relation is true between the gradient of f and the gradients of subsidiary functions h_j at an optimal solution x^0. In particular, if f and the functions h_j are convex, then a solution x^0 that satisfies stated conditions for some vector λ of multipliers is a global minimum.

Convex quadratic programming problems

Some methods for solving convex quadratic programming problems are closely related to the simplex method for linear programming and share with it the feature of terminating after only a finite number of iterations. Most of these methods make use of the Kuhn-Tucker conditions because the differentiability and regularity requirements are automatically satisfied. These characteristics are seldom found in nonlinear programming procedures in general. Instead of leading in a finite number of steps to an exact solution, they (at best) yield a point close to a point at which a local minimum is attained.

One of the historically important ancestors of nonlinear programming methods is the method of steepest descent due to the 19th-century French mathematician Augustin-Louis Cauchy. Its original application was to solving systems of nonlinear equations. For example, to solve the system of simultaneous equations $g_i(x) = 0$, $(i = 1, \cdots, m)$ a sum of squares of the functions appearing in the equations can be formed (see 27). An attempt can be made to obtain its unconstrained minimum. For every vector x, the functional value $f(x)$ is nonnegative because it is the sum of squares. Consequently, x^0 is a solution if and only if $f(x^0) = 0$.

The method is easily stated: From a starting point (trial solution) $x^{(1)}$, compute the direction of steepest descent, $-\nabla f(x^{(1)})$. Using a fixed step size, symbolized by the Greek letter sigma, σ, minimize the function f along the line segment of length σ issuing from $x^{(1)}$ in the direction of steepest descent. The point $x^{(2)}$ that minimizes $f(x)$ along this line segment becomes the new trial solution. In general, the preceding step is repeated from at each trial solution $x^{(k)}$ to obtain the next one $x^{(k+1)}$.

The method of steepest descent is valid under the assumption that the function f is continuously differentiable, and all points satisfying $f(x) \leq f(x^{(1)})$ lie within a fixed finite distance from the origin. The aim of the procedure is actually to locate a point x^* satisfying the necessary local conditions of optimality: $\nabla f(x^*) = 0$. Because such a point might be reached only in a limiting sense, it is customary to terminate the computational process when a point $x^{(k)}$ is reached at which the absolute value $|\partial f(x^{(k)})/\partial x_j|$ is less than a prescribed value for $j = 1, \cdots, n$. If the function f being minimized is convex, a point x^* at which $\nabla f(x^*)$ is zero must be a global minimum.

Another method deals with the linearly-constrained nonlinear programming problem: Minimize $f(x)$ subject to $Ax = b$, $x \geq 0$ in a manner that uses both linear programming and the Kuhn–Tucker conditions. It is assumed that the objective function $f(x)$ is continuously differentiable. It is then natural to replace f by its linear approximation at a feasible point and solve the associated linear program. For example, if $x^{(k)}$ is a feasible point, f can be replaced by its linear approximation (see 28). Assuming that $y^{(k)}$ is the solution of this linear program, a search can then be performed to minimize f along the line joining $x^{(k)}$ to $y^{(k)}$. The point $x^{(k+1)}$ at which the latter minimum is attained becomes the new trial solution at which the linear approximation to f can be formed for the repetition of the process. The computation can be interrupted if a point x^0 is reached at which the solution point y^0 of the associated linear programming problem satisfies $\nabla f(x^0)(y^0 - x^0) \geq 0$. It can be shown that the point x^0 must be optimal when the objective function f is convex.

Of the many algorithms for the solution of nonlinear programming problems, none is known to be superior to the generalized reduced-gradient method. Its roots lie in the reduced-gradient method (for linearly-constrained nonlinear programs) and ultimately in the simplex method itself. (G.B.D./R.W.Co.)

(26)
$$\begin{cases} \text{Minimize } f(x) \text{ subject to } h_j(x) \leq 0 \text{ for } j = 1, \cdots, l. \\ \text{Then } x^0 \text{ is a locally minimizing solution only if there} \\ \text{exist multipliers } \lambda_0, \lambda_1, \cdots, \lambda_l \text{ not all zero such that} \\ \lambda_0 \nabla f(x^0) = \lambda_1 \nabla h_1(x^0) + \cdots + \lambda_l \nabla h_l(x^0) \text{ and} \\ h_j(x^0) \leq 0, \quad \lambda_j \geq 0, \quad \lambda_j \cdot h_j(x^0) = 0 \text{ for } j = 1, \cdots, l. \end{cases}$$

(27)
$$f(x) = [g_1(x)]^2 + \cdots + [g_m(x)]^2$$

(28)
$$F(y) = f(x^k) + \nabla f(x^k)(y - x^k)$$

III. Cybernetics

The presentation to follow deals only with the principles of cybernetics and includes reference to mathematical aspects such as information theory, automata theory, and cybernetic systems. The technological aspects of cybernetics are multifold and can be found in COMPUTERS; CONTROL SYSTEMS; and ROBOT DEVICES.

DEFINITIONS OF THE TERM CYBERNETICS

The term cybernetics comes from the ancient Greek word *kybernētikos* ("good at steering") referring to the art of the helmsman. In the first half of the 19th-century, the French physicist André-Marie Ampère, in his classification of sciences, suggested that the still nonexistent science of the control of governments be called cybernetics. This term was soon forgotten until used by the United States mathematician, Norbert Wiener, as the title for his book published in 1948. In that book Wiener made reference to an 1868 article by the British physicist James Clerk Maxwell (*q.v.*) on governors and pointed out that "governor" is derived, via Latin, from the same Greek word that gives rise to the term "cybernetics." The date of Wiener's publication is generally accepted as the date of the birth of cybernetics as an independent science. Wiener defined cybernetics as "the science of control and communications in the animal and machine." This definition relates cybernetics closely first of all with the theory of automatic control and with physiology, particularly the physiology of the nervous system. Subsequently, the computer and the areas of mathematics related to it (*e.g.*, mathematical logic) had a great influence on the development of cybernetics. The reason is that the computer can be used not only for automatic calculation, but also for all conversions of information, including various types of information processing used in control systems. This ability made two different views of cybernetics possible. The narrower view, common in the Western countries, defines cybernetics as the science of control of complex systems of various types (technical, biological, and social). In many countries (for example, in the United States), particular emphasis is given to those aspects of cybernetics that are used in the generation of control systems in technology and in living organisms. In addition to cybernetics, the science of computers and the general rules of information processing is being developed in the Western countries (English—computer science; French—*informatique*).

The broader interpretation of the subject of cybernetics common in the Soviet Union includes not only control but all forms of information processing. This definition includes therefore the Western computer science as one of the component parts of cybernetics.

PRINCIPLES

Information theory. It should be emphasized that the concept of information and its conversion is an important first concept of cybernetics no matter how its subject is defined (see INFORMATION THEORY). Information arises with any act of selection or limitation of possibilities of

selection. For example, if it is known that a room can contain 100 men, then the message that the room at a given moment contains 50 men or that the number of people in the room is less than 40 carries information. The message that the number is less than 200 carries no information, because the first, fixed selection is in no way thus limited. In developing the principle of limitation of selection, Claude E. Shannon, an electrical engineer in the United States, introduced the method of changing the quantity of information.

The nature and types of signals

Information is transmitted by signals. In pure cybernetics, the physical nature of the signals is completely disregarded. It is important only that the signals can be differentiated from each other. The form of the set of possible signals and the nature of their changes are also significant. For example, a signal carrying information on the number of persons in a room cannot take on fractional values. The values of these signals change in jumps. On the other hand, a signal giving information of the temperature of the air in the room cannot change its value, say from 19° C to 20° C, without passing through all intermediate values. Signals of the first type are called discrete and of the second type, continuous. The same terms are used in relation to the information represented by these signals.

In describing continuous information within a certain accuracy, it can always be reduced to discrete information. The usual method of representation of discrete information is as a finite sequence of signals selected from a certain fixed finite set of signals called the (abstract) alphabet, for example the set of letters in the Latin alphabet, the set of decimal numbers, etc. One important but simple fact is the possibility of representing any discrete information in the form of sequences of signals of only two different types, as is done, for example, in the dots and dashes of the well-known Morse telegraph code. Problems of various forms of representation of discrete information make up the subject of a special division of theoretical cybernetics called encoding theory.

Automata as information converters. Encoding is the simplest form of information conversion. In the general case, these conversions are performed by so-called information converters, or data processors. An information converter, referred to in cybernetics as a machine or automaton (see AUTOMATA THEORY), is a device that converts a certain set of signals called the input signals to another set, the set of output signals. The input signals arrive at the converter through the input channels, the output signals leave through the output channels. Furthermore, in the general case, the processor can store signals in the form of a certain set of parameters, the different values of which determine the internal state of the processor, that is, the condition of its memory.

A processor is called continuous if all of the parameters that define it (memory, input, and output signals) are continuous. In the case of discrete processors, all of these signals must be discrete. Processors dealing both with continuous and with discrete signals are called hybrid processors.

The changes in output signals and in the internal state of any processor depend generally both on the input signals and on the internal state of the processor at a given time. For a processor without memory, the values of the output signals at any moment in time depend only on the values of the input signal at the same moment in time. If all of these dependences are fixed by fully defined, unambiguous functions (*e.g.*, single-valued), the processor is called deterministic. If there are random dependences, the processor is called probabilistic.

Information processors in which no functional components and no internal structural parts other than those described above can be distinguished are called elementary. Nonelementary processors, also called cybernetic systems, consist of networks composed of elementary (or in any case simpler) processors. A network is constructed by connecting some (or all) of the output channels of the processors composing the network to some (or all) of their input channels. A portion of the input and output channels of the elementary processors may be appropriately connected to the input and output signals of the entire system.

The concept of elementariness or nonelementariness of a processor is relative and depends on the depth of penetration into the subject. For example, if interest is confined to the functions of the brain, not in its structure, it can be analyzed as an elementary processor (although it is characterized by a tremendous number of parameters). It is known, however, that the brain is an extremely complex system, composed of more elementary processors, the neurons. In turn, upon transition to the molecular level, the neurons themselves can be regarded as extremely complex cybernetic systems.

The cybernetic approach to systems study

Study of systems. The study of systems and particularly of the complex probabilistic systems is one of the most important tasks in cybernetics. The specific feature of the cybernetic approach to the study of systems is that it abstracts itself from the actual nature of the elements of the system and the signals circulating in it. This allows a general mathematical apparatus and general methods of investigation to be developed, suitable for systems of varying nature and purpose. In the study of continuous cybernetic systems, a system of ordinary differential equations is particularly significant.

Theory of algorithms. In the discrete case, the modus operandi is the so-called theory of algorithms. An algorithm is an arbitrary, finite system of rules of any nature, allowing expressions in any alphabet to be converted to new expressions in the same (or any other) alphabet. So-called algorithmic languages are used for precise descriptions of the rules for these conversions (see AUTOMATA THEORY: *Classification of automata;* and COMPUTERS: *Language categories*).

It is extremely important that the presence in an algorithmic language of a relatively small number of means of expression can give it the property of universality. Universality of an algorithmic language means the possibility of expressing in this language any conversion of discrete information that can be defined using a finite number of rules, that is, expressed as a certain algorithm in any other algorithmic language.

The fact of existence of universal algorithmic languages means that it is possible to construct universal discrete information processors. To do this, it is sufficient to fix upon some concrete alphabet and to construct a memory device capable of storing any word (expression) in the alphabet and a device capable of performing all of the elementary rules of a certain universal algorithmic language.

Universal digital computers are such universal information processors if the limitations resulting from their finite memory volume are not considered. Until computers appeared, the only universal data processor was the natural information processor, the human brain.

The availability of technical universal data processors is highly significant for the development of cybernetics in at least two ways. First of all, this fact opens unlimited possibilities for the automation not only of the physical, but also of the mental activity of man, and this automation does not require that new technical devices be invented for each case. It is sufficient to study and describe the rules defining the type of activity to be automated and to program them; *i.e.*, express them in one of the algorithmic languages used in existing computers.

Mathematical modelling. Second, the use of universal data processors—computers—gives cybernetics a method of scientific investigation of systems that is new in principle—so-called mathematical modelling. Until this method appeared, scientists actually had only two different methods of study: experimental and theoretical. In the former, experiments were performed only with the system itself, or with an actual, physical model of the system. In the second, it was necessary to be able to solve equations describing the system.

Mathematical modelling occupies an intermediate position between these two methods: there is no necessity to construct an actual physical model of the system. It is replaced by a mathematical model; *i.e.*, a description of the system in some algorithmic language. There is also no

need to solve complex mathematical problems related to this description (*i.e.*, to solve systems of differential equations). The description of the system is simply entered into a computer that models the behaviour of the system (*i.e.*, provides precise descriptions of the system) under various conditions defined according to the purpose of the research assignment.

This method allows the scientist to produce a full description of complex systems, the individual parts of which are studied by different people or even in different sciences. One example is the human organism. Its individual parts (circulatory system, digestive system, nervous system, glandular system, etc.) are studied by different specialists, even though the parts are closely related to each other.

Modelling the human organism

To produce a mathematical model of the organism, first of all it must be divided into individual parts and the state of each part described by some system of parameters. Among these parameters might be continuous quantities (for example, the percent content of sugar in the blood) or discrete quantities (for example, qualitatively differing levels of secretory function of the liver). The next step is description of the relationships between the separate parts. One of the most typical forms of relationship for this example is expressed by a sentence such as: When organ A shifts from state m_1 to state m_2, and organ B shifts from state n_1 to state n_2 after k days (weeks, months), organ C will change its state from l_1 to l_2 with probability p. The specialists must describe all of the relations of this type with which they are familiar (called simple relationships) to form the required mathematical model. Operations with this model in the computer permit the establishment of the complex relationships (*i.e.*, the influence of organs on other organs not directly, but through still other organs). This type of model can be used to study versions of the development of various diseases, various methods of treatment, etc.

Mathematical modelling can in principle be performed in any universal data processor, including the human brain. The brain, however, being a very complex system itself, is comparatively poorly suited for such routine but laborious work as the modelling of complex systems. Consequently, only the appearance of the computer provided a qualitative jump and made it possible to perform effective study of really complex systems in various areas of knowledge. Actually, the concern is with coarse qualitative criteria, differentiating simple systems from complex; if the structure and behaviour of a system can be studied by a single man in a reasonable time, the system is called simple. If the efforts of many persons and the use of special technical equipment (computers) are required to draw the whole picture, the system is called complex.

Cybernetic systems. *Self-teaching mechanisms.* Complex tasks arise in so-called self-teaching systems, in which attempts to achieve a certain final goal lead to changes in the methods of its attainment, and the setting of various intermediate goals. One of the simplest examples of this type of system is the so-called Shannon's mouse. It is a moving automaton placed in a maze. The final purpose of the automaton is to find the "food" placed at some point in the maze.

At first, the mouse uses a simple trial and error algorithm, bumping into the walls until an exit is found. During the process of this type of search, it "learns" and memorizes the plan of the maze. After once reaching the goal (by chance), the mouse will use an entirely different, much more economical, and seemingly "intelligent" algorithm for reaching the goal in a subsequent experiment, based on "knowledge" of the layout of the maze.

Self-teaching is one type of self-improvement of control systems. In this case, the improvement occurs without changing the structure of the system. It is also possible, however, that the upper levels of a system may change the structure of lower levels in order to improve their functioning. This type of self-improvement is naturally called self-development.

By the 1970s, methods had been designed for quantitative definition of self-improvement. In this case, in pure theory, the boundary between self-teaching and self-development disappears. In practice, however, the difference between these types of self-improvement is usually rather clear.

Biocybernetics. Many control systems in biology are constructed according to the principles of self-teaching and self-development. For example, in all probability, only a small portion of the organization of the human brain is determined genetically. Everything else is produced as a result of the effects of many, rather effective, mechanisms of self-improvement.

Highly complex examples of self-development can be found in the processes of biological evolution, as well as various social processes.

For systems of the high degree of complexity characteristic of living organisms and human society, an ever greater role is played not just by control processes, but by cognition processes. These systems have highly perfected systems of sensors and effectors (for example, the eye and the hand), capable of recognizing complex patterns and performing widely varied actions. The central portion of the system has a multilevel structure; the upper links of the structure develop abstract concepts and recognize the deeper regularities existing both in the system itself and in its surroundings.

By the 1970s, cybernetics had achieved significant success in the solution of the problem of pattern recognition, automation (by computer) of the processes of logical conclusion, formation of new concepts, and other problems. This success has allowed the creation of universal robots controlled by computers and imitating rather complex forms of conscious behaviour. (Ed.)

IV. Control theory

GENERAL BACKGROUND

As long as human culture has existed, control has always meant some kind of power over man's environment. Cuneiform fragments suggest that the control of irrigation systems in Mesopotamia was a well-developed art at least by the 20th century BC. There were some ingenious control devices in the Greco-Roman culture, the details of which have been preserved. Methods for the automatic operation of windmills go back at least to the Middle Ages. Large-scale implementation of the idea of control, however, was impossible without a high-level of technological sophistication, and it is probably no accident that the principles of modern control started evolving only in the 19th century, concurrently with the Industrial Revolution. A serious scientific study of this field began only after World War II and is now a major aspect of what has come to be called the second industrial revolution.

Examples of feedback control

Although control is sometimes equated with the notion of feedback control (which involves the transmission and return of information)—an isolated engineering invention, not a scientific discipline—modern usage tends to favour a rather wide meaning for the term; for instance, control and regulation of machines, muscular coordination and metabolism in biological organisms, prosthetic devices; also, broad aspects of coordinated activity in the social sphere such as optimization of business operations, control of economic activity by government policies, and even control of political decisions by democratic processes. Scientifically speaking, modern control should be viewed as that branch of system theory concerned with changing the behaviour of a given complex system by external actions. (For aspects of system theory related to information, see below.) If physics is the science of understanding the physical environment, then control should be viewed as the science of modifying that environment, in the physical, biological, or even social sense.

Much more than even physics, control is a mathematically-oriented science. Control principles are always expressed in mathematical form and are potentially applicable to any concrete situation. At the same time, it must be emphasized that success in the use of the abstract principles of control depends in roughly equal measure on the status of basic scientific knowledge in the specific field of application, be it engineering, physics, astronomy, biology, medicine, econometrics, or any of the social sciences. This fact should be kept in mind to avoid confu-

sion between the basic ideas of control (for instance, controllability) and certain spectacular applications of the moment in a narrow area (for instance, manned lunar travel).

EXAMPLES OF MODERN CONTROL SYSTEMS

To clarify the critical distinction between control principles and their embodiment in a real machine or system, the following common examples of control may be helpful. There are several broad classes of control systems, of which some are mentioned below.

Machines that cannot function without (feedback) control. Many of the basic devices of contemporary technology must be manufactured in such a way that they cannot be used for the intended task without modification by means of control external to the device. In other words, control is introduced after the device has been built; the same effect cannot be brought about (in practice and sometimes even in theory) by an intrinsic modification of the characteristics of the device. The best-known examples are the vacuum-tube or transistor amplifiers for high-fidelity sound systems. Vacuum tubes or transistors, when used alone, introduce intolerable distortion, but when they are placed inside a feedback control system any desired degree of fidelity can be achieved. A famous classical case is that of powered flight. Early pioneers failed, not because of their ignorance of the laws of aerodynamics, but because they did not realize the need for control and were unaware of the basic principles of stabilizing an inherently unstable device by means of control. Jet aircraft cannot be operated without automatic control to aid the pilot, and control is equally critical for helicopters. The accuracy of inertial navigation equipment (the modern space compass) cannot be improved indefinitely because of basic mechanical limitations, but these limitations can be reduced by several orders of magnitude by computer-directed statistical filtering, which is a variant of feedback control.

The role of computers in feedback control

Control of machines. In many cases, the operation of a machine to perform a task can be directed by a human (manual control), but it may be much more convenient to connect the machine directly to the measuring instrument (automatic control); *e.g.*, a thermostat (temperature-operated switch) may be used to turn on or off a refrigerator, oven, air-conditioning unit, or heating system. The dimming of automobile headlights, the setting of the diaphragm of a camera, the correct exposure for colour prints, may be accomplished automatically by connecting a photocell or other light-responsive device directly to the machine in question. Related examples are the remote control of position (servomechanisms), speed control of motors (governors). It is emphasized that in such case a machine could function by itself, but a more useful system is obtained by letting the measuring device communicate with the machine in either a feedforward or feedback fashion.

Control of large systems. More advanced and more critical applications of control concern large and complex systems the very existence of which depends on coordinated operation using numerous individual control devices (usually directed by a computer). The launch of a spaceship, the 24-hour operation of a power plant, oil refinery, or chemical factory, the control of air traffic near a large airport, are well-known manifestations of this technological trend. An essential aspect of these systems is the fact that human participation in the control task, although theoretically possible, would be wholly impractical; it is the feasibility of applying automatic control that has given birth to these systems.

Biocontrol. The advancement of technology (artificial biology) and the deeper understanding of the processes of biology (natural technology) has given reason to hope that the two can be combined; man-made devices should be substituted for some natural functions. Examples are the artificial heart or kidney, nerve-controlled prosthetics, and control of brain functions by external electrical stimuli. Although definitely no longer in the science-fiction stage, progress in solving such problems has been slow not only because of the need for highly advanced technology but also because of the lack of fundamental knowledge about the details of control principles employed in the biological world.

Robots. On the most advanced level, the future task of control science is the creation of robots. This is a collective term for devices exhibiting animal-like purposeful behaviour under the general command of (but without direct help from) man. Industrial manufacturing robots are already fairly common, but real breakthroughs in this field cannot be anticipated until there are fundamental scientific advances with regard to problems related to pattern recognition and the mathematical structuring of brain processes.

Obstacles to the creation of robots

PRINCIPLES OF CONTROL

The scientific formulation of a control problem must be based on two kinds of information: (A) the behaviour of the system (*e.g.*, industrial plant) must be described in a mathematically precise way; (B) the purpose of control (criterion) and the environment (disturbances) must be specified, again in a mathematically precise way.

Information of type A means that the effect of any potential control action applied to the system is precisely known under all possible environmental circumstances. The choice of one or a few appropriate control actions, among the many possibilities that may be available, is then based on information of type B; and this choice, as stated before, is called optimization.

The task of control theory is to study the mathematical quantification of these two basic problems and then to deduce applied-mathematical methods whereby a concrete answer to optimization can be obtained. Control theory does not deal with physical reality but only with its mathematical description (mathematical models). The knowledge embodied in control theory is always expressed with respect to certain classes of models, for instance, linear systems with constant coefficients, which will be treated in detail below. Thus control theory is applicable to any concrete situation (*e.g.*, physics, biology, economics) whenever that situation can be described, with high precision, by a model that belongs to a class for which the theory has already been developed. The limitations of the theory are not logical but depend only on the agreement between available models and the actual behaviour of the system to be controlled. Similar comments can be made about the mathematical representation of the criteria and disturbances.

Once the appropriate control action has been deduced by mathematical methods from the information mentioned above, the implementation of control becomes a technological task, which is best treated under the various specialized fields of engineering. The detailed manner in which a chemical plant is controlled may be quite different from that of an automobile factory, but the essential principles will be the same. Hence further discussion of the solution of the control problem will be limited here to the mathematical level.

To obtain a solution in this sense, it is convenient (but not absolutely necessary) to describe the system to be controlled, which is called the plant, in terms of its internal dynamical state. By this is meant a list of numbers (called the state vector) that expresses in quantitative form the effect of all external influences on the plant before the present moment, so that the future evolution of the plant can be exactly given from the knowledge of the present state and the future inputs. This situation implies, in an intuitively obvious way, that the control action at a given time can be specified as some function of the state at that time. Such a function of the state, which determines the control action that is to be taken at any instant, is called a control law. This is a more general concept than the earlier idea of feedback; in fact, a control law can incorporate both the feedback and feedforward methods of control.

In developing models to represent the control problem, it is unrealistic to assume that every component of the state vector can be measured exactly and instantaneously. Consequently in most cases the control problem has to be broadened to include the further problem of state deter-

mination, which may be viewed as the central task in statistical prediction and filtering theory. Thus, in principle, any control problem can be solved in two steps: (1) Building an optimal filter (so-called Kalman filter) to determine the best estimate of the present state vector; (2) determining an optimal control law and mechanizing it by substituting into it the estimate of the state vector obtained in step 1.

In practice, the two steps are implemented by a single unit of hardware, called the controller, which may be viewed as a special-purpose computer. The theoretical formulation given here can be shown to include all other previous methods as a special case; the only difference is in the engineering details of the controller.

The mathematical solution of a control problem may not always exist. The determination of rigorous existence conditions, beginning in the late 1950s, has had an important effect on the evolution of modern control, equally from the theoretical and the applied point of view. Most important is controllability; it expresses the fact that some kind of control is possible. If this condition is satisfied, methods of optimization can pick out the right kind of control using information of type B.

Controllability

CONTROL OF LINEAR SYSTEMS

Systems with constant coefficients. The preceding considerations may be illustrated much more directly by taking the special case of linear systems with constant coefficients. The point of view and most of the results described below originated after 1950, but the beginnings of the theory are more than 100 years old dating from Maxwell's work published in 1868. In fact, this class of models is the only one for which a reasonably complete mathematical theory of control exists so far.

One may denote by $y(t)$ the number giving the value of the output of the plant at time t; similarly, the number $u(t)$ denotes the value of the input to the plant at the same time. A classical way to describe the behaviour of the plant (information of type A) is to assume that the input and output are related by a linear differential equation with constant coefficients $\alpha_1, \cdots, \alpha_n$ (see 29), which is called the open-loop equation of the plant. In this case the state vector is the list $(y(0), \cdots, y^{(n-1)}(0))$ consisting of the output at time $t = 0$ and the first $n - 1$ derivatives with respect to time of the function $y(t)$ at $t = 0$. Under these circumstances, the control law is a linear combination of the state variables, with coefficients $\beta_1, \cdots, \beta_n$ (see 30).

The fact that this control law is linear in the state is an assumption needed to assure the manageability of mathematical machinery for the problem. It is desirable to assume also that the coefficients $\beta_1, \cdots, \beta_n$ are all constants so that the same formula can be used at any time. [A minus sign in the control law (30) expresses the classical idea of negative feedback; in the general theory, this has become merely a notational convention.]

Substitution of the control law into the linear differential equation of the plant (29) leads to the closed-loop differential equation for the controlled output of the plant (see 31). Here the coefficients γ_i are the differences $\alpha_i - \beta_i$. So the effect of control is to replace the coefficients α_i of the open-loop equation with the γ_i of the closed-loop equation. Since the β_i in the control law are arbitrary, it follows that arbitrary changes can be effected in the dynamical behaviour of the plant by means of control. In particular, it is possible to make a plant stable by means of the control law because the criterion for stability is given by the positiveness of the Routh–Hurwitz determinants (see 32). These conditions can of course always be met since the γ_i can be arranged to have any values.

The criterion for stability

It is necessary to comment on this seemingly overly general (but indeed correct) result from two points of view. First, the result assumes that the plant description (see 29) is exactly true and that the state list $(y(t), \cdots, y^{(n-1)}(t))$ is known at every time t. These assumptions are never exactly correct, but they can be approximated in many cases with sufficient accuracy by restricting the output $y(t)$ to small deviations about a given steady-state value and by constructing a statistical filter to estimate the derivatives of the function $y(t)$ with respect to t. Second, the model (see 29) does not represent the most general linear system with constant coefficients; in fact, this model, because of its special form, implicitly possesses the property of controllability that may not hold for the general model in the linear, constant coefficient class.

(29) $$\frac{d^n y}{dt^n} + \alpha_1 \frac{d^{n-1} y}{dt^{n-1}} + \cdots + \alpha_n y = u(t)$$

(30) $$-u(0) = \beta_n y(0) + \beta_{n-1} \dot{y}(0) + \cdots + \beta_1 y^{(n-1)}(0)$$

(31) $$\frac{d^n y}{dt^n} + \gamma_1 \frac{d^{n-1} y}{dt^{n-1}} + \cdots + \gamma_n y = 0$$

(32) $$\gamma_i > 0, \quad \begin{vmatrix} 1 & \gamma_2 \\ \gamma_1 & \gamma_3 \end{vmatrix} > 0, \ldots$$

(33) $$\frac{dx}{dt} = Fx + gu(t), \qquad y(t) = \langle h, x(t) \rangle$$

(34) $$F = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ 0 & 0 & 0 & \cdots & 1 \\ -\alpha_n & -\alpha_{n-1} & -\alpha_{n-2} & \cdots & -\alpha_1 \end{bmatrix}, \qquad g = \begin{bmatrix} 0 \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \\ 1 \end{bmatrix}$$

(35) $$\det[g \quad Fg \quad F^2 g \quad \cdots \quad F^{n-1} g] \neq 0$$

The explicit form of the controllability condition is of great practical and philosophical importance. It is a generic condition: it is satisfied by almost all systems of the linear, constant-coefficient class. Because the state-vector equations (33) accurately represent most physical systems making only small deviations about their steady-state behaviour, it follows further that in the natural world control in the small is, in principle, almost always possible. This fact of nature is the theoretical basis of practically all the presently existing control technology.

The most general system of this class (with one input and one output) is given by a pair of state-vector equations (see 33); the symbol x denotes the state vector (a list of n arbitrary numbers), F is an $n \times n$ matrix, g is an n-vector, h is an n-vector, and $< , >$ is the inner product; the symbols $u(t)$ and $y(t)$ have the same meaning as above. The equations of the plant discussed above (see 29) can always be transformed into the general case (see 33), but not necessarily vice versa.

Transforming the plant equations discussed above to the general form gives the canonical matrices F, g (shown in 34). If these expressions are substituted into the controllability criterion (see 35) for the general system (33), it is seen that the special model for the plant that was used above (see 29) is always controllable. Thus, it is clear that the true significance of the celebrated controllability concept can be grasped only after the control problem has been formulated in the state-vector style. In the classical development of control theory, too much emphasis was placed on special models for the plant (see 29); the advances since 1950 to a large extent resulted from a recognition that the most general model corresponding to the basic restrictive hypotheses of linearity and constant coefficients is given by the state-vector equations (see 33); this recognition eventually led to the discovery that for the general class there is a nontrivial controllability condition (see 35).

Control limitations in nonlinear modelling

On the other hand, little is known at present, in the scientific sense, about the ultimate limitations of control when the models in question are not linear. In particular, it is not known under what conditions control is possible in the large; that is, for arbitrary deviations from existing conditions. This lack of scientific knowledge should be kept in mind in assessing often exaggerated claims by economists and sociologists in regard to a possible improvement in human society by governmental control.

Optimal control. To reduce the arbitrariness of the control law (see 30), it is necessary to use information of type B. This is done, conceptually, in the following way. It is assumed (after a change of variables, if necessary) that the state $x = 0$ represents the ideal condition of the plant to be controlled. Any given state $x(0)$ not equal to zero at time $t = 0$ is regarded as an undesirable deviation, attributed to unavoidable disturbances acting on the system. The task of control is to restore the ideal condition $x(t) = 0$ in the best possible way. The specification of a suitable optimality criterion is frequently dictated by mathematical expediency.

The long mathematical history of the calculus of variations has suggested that the most natural criterion to consider is the minimization of an integral from zero to infinity, which measures the deviation of $x(t)$ from zero as well as the cost of control. For example, the coefficients in the control law may be explicitly determined by minimizing a quadratic integral (see 36) by the choice of the function $u(t)$ with respect to solutions of the open-loop differential equation (see 29). The applied-mathematical technique for solution is well developed and requires the solution of a nonlinear differential equation of the Riccati type, which can be conveniently carried out by digital computer. It can be shown also that this general procedure includes older techniques based on semi-quantitative engineering analyses, for instance, the idea of maximizing the loop gain.

In many respects, optimization theory represents a satisfactory solution of the control problem (in some cases even outside the linear case), but basic difficulties remain in regard to the appropriate choice of the criterion represented by the quadratic forms in the performance index (see 36).

Optimal filtering and state estimation. At least in the linear case, the solution of the state-estimation problem can be accomplished by methods similar to that for optimal control; this follows from the well-known Duality Principle which states that the dual equations (matrices and vectors replaced by their transposes) of control optimization are the equations for state estimation. The optimal filter is then a dynamical system similar to (29), in which the principal task is to determine a dual control law similar to (30).

In practice, the state estimation problem is much more important than the control problem; approximately 90–95 percent of practical controllers are devoted to state estimation, only about five percent being required for the implementation of the control law.

NONLINEAR CONTROL SYSTEMS

Whenever the controlled variable $y(t)$ is allowed to have large deviations from the steady state, the linear constant-coefficient model will cease to represent the plant accurately because of intrinsic nonlinearities involved in the description of most natural dynamical phenomena. Equally important are intentionally introduced nonlinearities that result from reasons of economy, simplicity and reliability of engineering, or from ignorance of the fact that the savings achieved by nonlinear control devices may be negated by factors resulting from the greater intrinsic difficulty of control. A good example of unavoidable nonlinearities in the object to be controlled is a space vehicle whose rockets, at the current state of technology, can only be controlled by turning them on or off. Continuously variable control action (which is desirable) in this case is technologically impossible; the high cost of rocket engines, however, justifies extremely sophisticated computer-based technology to achieve optimal control. The purpose of the computer is then simply one of switching the rockets on or off; the difficulty of the control lies in the extreme precision with which the time must be determined when this takes place.

Example of a nonlinearity

$$(36) \qquad \int_0^\infty [Q(x(t)) + R(u(t))]\,dt \qquad (Q, R \text{ are quadratic forms})$$

When the basic mathematical data can be stated in the conventional optimization framework, effective methods are available for solving the optimal control problem even in the nonlinear case. The methods for doing this are an extension of the classical methods of the calculus of variations. These methods, however, have yielded little theoretical insight and their straightforward application becomes prohibitively expensive for large-scale systems.

Even less satisfactory is the status of the nonlinear optimal filtering and state estimation problem, which, as has been noted, is a critical part of the general solution of the control problem. Because of nonlinearity, the Duality Principle no longer applies; even the formulation of the problem is controversial.

Nonlinear systems do not represent a special case, but simply everything that is not subjected to the special assumption of linearity; in a sense "nonlinear" is synonymous with "unknown." Scientific progress will undoubtedly occur by singling out special classes of systems subject to restrictive structural assumptions other than linearity.

(R.E.K.)

BIBLIOGRAPHY. The following works treat game theory: J. VON NEUMANN and O. MORGENSTERN, *Theory of Games and Economic Behavior*, 3rd ed. (1953); M.D. DAVIS, *Game Theory, A Nontechnical Introduction* (1970); JOHN D. WILLIAMS, *The Compleat Strategyst*, rev. ed. (1966); MARTIN SHUBIK (ed.), *Readings in Game Theory and Political Behavior* (1954); ROBERT D. LUCE and HOWARD RAIFFA, *Games and Decisions* (1957); G. OWEN, *Game Theory* (1968); A. RAPOPORT, *N-Person Game Theory, Concepts and Applications* (1970); and RICHARD B. BRAITHWAITE, *The Theory of Games As a Tool for the Moral Philosopher* (1955).

Technical references relating game theory to decision theory and econometrics include: ABRAHAM WALD, *Statistical Decision Functions* (1950, reprinted 1971); TIALLING C. KOOPMANS *et al.* (eds.), *Activity Analysis of Production and Allocation* (1951); DAVID BLACKWELL and M.A. GIRSHICK, *Theory of Games and Statistical Decisions* (1954); and R.M. THRALL, C.H. COOMBS, and R.L. DAVIS (eds.), *Decision Processes* (1954).

The history, theory, and applications of linear programming may be found in GEORGE B. DANTZIG, *Linear Programming and Extensions* (1963); see also GEORGE HADLEY, *Linear Programming* (1962). For the classic work on the subject, see LESTER R. FORD and D.R. FULKERSON, *Flows in Networks* (1962). An alternate source is T.C. HU, *Integer Programming and Network Flows* (1967). LEON S. LASDON, *Optimization Theory for Large Systems* (1970), deals with linear and nonlinear programming problems. One of the pathbreaking books on linear and nonlinear programming is G. ZOUTENDIJK, *Methods of Feasible Directions* (1960). OLVI L. MANGASARIAN, *Nonlinear Programming* (1969), deals exclusively with theory; while WILLARD I. ZANGWILL, *Nonlinear Programming* (1969), is concerned primarily with algorithms. From various conferences have come J. ABADIE (ed.), *Nonlinear Programming* (1967) and *Integer and Nonlinear Programming* (1970); and GEORGE B. DANTZIG and ARTHUR F. VEINOTT, JR. (eds.), *Mathematics of the Decision Sciences*, 2 vol. (1968)—all feature papers on a wide range of subjects.

The reader interested in cybernetics may wish to consult NORBERT WIENER, *Cybernetics*, 2nd rev. ed. (1961), a very general discussion; and VIKTOR GLUSHKOV, *Introduction to Cybernetics* (1966). See also STAFFORD BEER, *Cybernetics and Management* (1959); and JIRI KLIR and MIROSLAV VALACH, *Kybernetické modelování* (1965; Eng. trans., *Cybernetic Modelling*, 1967). NORBERT WIENER (*op. cit.*). A historical overview of feedback devices may be found in OTTO MAYR, *The Origins of Feedback Control* (1970). Among good books on modern control and system theory, see RICHARD E. BELLMAN, *Dynamic Programming* (1957). More mathematical treatments are L.S. PONTRYAGIN *et al.*, *The Mathematical Theory of Optimal Processes* (1962); and E.B. LEE and L. MARKUS, *Foundations of Optimal Control Theory* (1967). Control theory in the wider context of system theory is treated in R.E. KALMAN *et al.*, *Topics in Mathematical System Theory* (1969), see especially Chapter 2. ROGER W. BROCKETT, *Finite-*

Dimensional Linear Systems (1970), surveys the fundamental problems of description and optimization of linear, constant-coefficient systems. For information on biocontrol, see DOUGLAS and K. STANLEY-JONES, *The Kybernetics of Natural Systems* (1960).

(A.W.T./H.W.K./G.B.D./R.W.Co./R.E.K./Ed.)

Orange Free State

The Orange Free State—Oranje Vrystaat in the Afrikaans language—is the second smallest of the four provinces of the Republic of South Africa. It has an area of 49,866 square miles (129,152 square kilometres) and a population of over 1,600,000 of which about 1,300,000 are black Africans and about 300,000 are white. Landlocked, the province is bordered to the north by Transvaal Province, to the east by Natal Province and by the independent state of Lesotho, and to the south and west by Cape Province. The administrative capital is Bloemfontein, which is also the judicial capital of the Republic of South Africa.

More than three-quarters of the white population is Afrikaner, and the state is a stronghold of Afrikaner culture. While its undulating plains produce up to 40 percent of South Africa's maize, a quarter of its wheat, and a quarter of its wool, the Orange Free State also contains the most productive of South Africa's seven principal goldfields, from which uranium is also obtained. As South Africa's central province, the state is the focus of its transport network. (For associated physical features, see ORANGE RIVER; VELD; for general historical background, see SOUTHERN AFRICA, HISTORY OF.)

History. Before the arrival of Europeans, the region was the home of seminomadic Bantu tribes. In the 18th century, Europeans first crossed the Orange River, which forms part of the state's southern border, to enter the area. In the early 19th century, trekboers (seminomadic, pastoral farmers of Dutch descent) began to settle the area. After 1836 came the Great Trek, a migratory movement in which Boer farmers seeking freedom from British rule also moved north across the Orange River. From 1848 to 1854 the British administered the territory as the Orange River Sovereignty, after which the British withdrew, and the Boer settlers formed the independent Orange Free State. The constitution of the new state combined traditional Boer institutions with Dutch and United States constitutional theory. During the first few years of the state's existence, it was much harassed by raids from Basuto (Sotho) tribesmen from the east; the Basutos were, however, at length conquered, and part of their territory was annexed. During the South African War (1899–1902), the Orange Free State fought against the United Kingdom by the side of its sister state the South African Republic (now the Transvaal), with which it had a defensive alliance. The two Boer republics won some victories against the British army but could not finally prevail. In 1900 the Orange Free State was annexed by the United Kingdom as the Orange River Colony. Self-government was restored in 1907, after which, in 1910, the colony became the Orange Free State Province of the Union of South Africa. In 1961, when the Union of South Africa became the Republic of South Africa, the province remained unchanged in form and administration.

The Great Trek

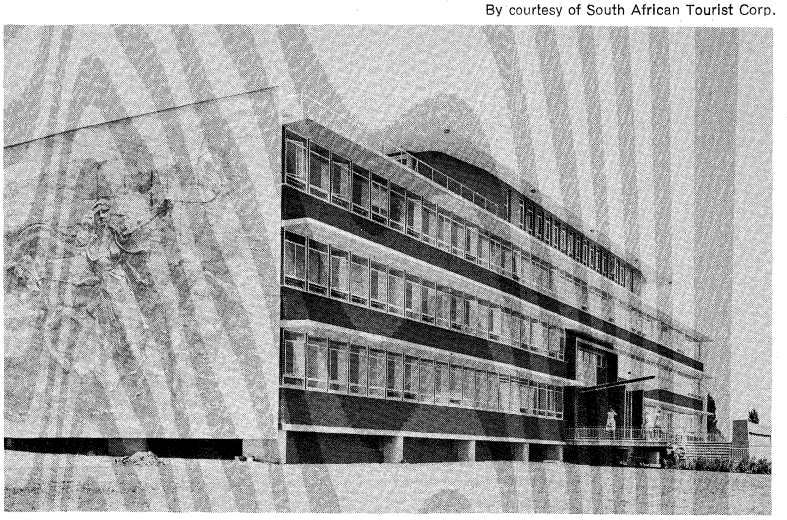
The landscape. *Relief.* The Orange Free State is situated on the Highveld, a high plateau that rises to 6,000 feet on its eastern boundary, sloping down to the west to almost 4,000 feet. Its surface is formed by beds of sandstones and shales (laminated rock form of clays) and is extremely even, except where it is broken by the occurrence of dolerite (coarse basalt) intrusions (inflows of formerly molten rock) or by the occurrence of mesas. This general evenness results in low gradients for river courses; coupled with the semiarid climate, it also produces a peculiar type of wind-eroded pan (depression) in the west that is filled with water during the rainy season. The province is entirely drained by two rivers—the upper Orange River, which, as mentioned, forms the southern boundary of the province, and the Vaal River, which forms part of the northern boundary, together with their tributaries. The largest of the tributaries is the Caledon River, which flows into the Orange from the Drakenberg mountains in neighbouring Lesotho. The Wilge River, which drains the northeast, is the principal tributary of the Vaal. An interesting geologic feature is the Vredefort Dome, a series of incomplete circular ranges enclosing a mass of old granite partly eroded by the Vaal River.

Soils. The eastern part of the province is covered by claylike, heavy, and somewhat acidic soils. Farther to the west are alkaline soils that can be impervious to water. The northwestern region forms the so-called Sandveld, on which good crops are obtained only in years of high rainfall. The west is covered by sandy soil resembling that of the Kalahari, where successful cultivation is possible only under irrigation.

Climate. The climate varies from a warm and temperate type with an annual rainfall of 40 inches in the east to a semiarid type with a rainfall of 15 inches in the far west. Mean annual surface temperatures gradually increase from about 58° F (14° C) in the east to 62° F (17° C) in the west. Frost is common over the entire province from May to September, while dust storms normally occur during drought-stricken summer periods. Whirlwinds are common during warmer days; and hailstorms, sometimes destructive, occur on an average about seven times a year in the east but only three times a year in the west. Because rainfall is unreliable, long periods of drought are frequent each year. Thunderstorms, almost exclusively a summer phenomenon, are the normal sources of precipitation, occurring between 60 and 100 days a year. Sunshine is abundant; there are only about six days a year on which Bloemfontein has no sunshine.

Rainfall and temperatures

By courtesy of South African Tourist Corp.



University of the Orange Free State, Bloemfontein.

The rate of evaporation is consequently high. Because rainfall is sporadic, much reliance is placed on underground supplies of water and on storage dams.

Vegetation and animal life. From west to east the vegetation consists of four successive zones—desert shrub, sweet grassveld, mixed grassveld, and sour grassveld. In most of the province, animal life, like the vegetation, has given way to crop cultivation and human settlement. Many animal and plant species are now protected by legislation. At Willem Pretorius Game Reserve, Highveld game animals, including herds of such species of antelope as blesbok and springbok, are plentiful; zebra and giraffe are also to be found. Golden Gate Highlands National Park was developed to restock the area with antelope species—such as black wildebeest and blesbok—and other animals and to establish a habitat for a wide variety of birds. Several of the smaller reptiles and snakes, such as adders, ringhals (venomous spitting snakes), and tree snakes, are still to be found over the entire province. The rarely observed freshwater fishes of the Orange and Vaal river systems include yellowfish, carp, and barbel.

Population. *Ethnic composition.* The 1970 population, totalling 1,649,000, included 296,000 whites, 36,000 Coloureds (racially mixed; an official designation of the Republic of South Africa), and 1,317,000 Bantu. The density averages about 33 persons per square mile. The ethnic pattern is revealed by the home language spoken. Among the white population, 84 percent are Afrikaans speaking; 13 percent speak English; and the remainder speak German, Dutch, or other European languages. Historically, the Afrikaners, whose ancestors came from the preindustrial Netherlands, tended to be farmers; while the English, who came from an industrialized Britain, tended to be city dwellers. Today, about 74 percent of the Afrikaners and 90 percent of the English live in towns and cities. In general, the rural white population increases from south to north, with Bloemfontein and the northwestern goldfields being the most densely populated regions. The Bantu population was originally largely confined to the northern and eastern districts, adjoining Lesotho, but today their numbers are rapidly increasing in the goldfield area. About 60 percent belong to the South Sotho (also known as Sotho proper or Basuto) tribe. Various other tribal groups include the Zulu to the northeast (who represent about 10 percent), the city-dwelling Xosa (Xhosa; 13 percent), and the Tswana (8 percent). Bantu densities also increase in a northerly direction. The Coloureds, almost without exception, speak Afrikaans; slightly more than 57 percent of them live in towns and cities; the remainder increase in number toward the southwest.

Bantu tribal groups

Religious affinities. Throughout the province, church buildings are dominant in towns. The Provincial Council begins the day's work with solemn prayer, and, in the distant farm house, divine worship is part of the family's evening activities. Less than 1 percent of the white population adheres to no religion. The three main Afrikaans churches (Nederduits Gereformeerde, Nederduits Hervormde, and Gereformeerde) account for 77 percent of white church membership, and the three main English churches (Anglican, Methodist, and Presbyterian) for 11 percent. White Christians feel a responsibility toward nonwhite peoples and engage in missionary activity. The Coloureds are almost entirely Christian: 68 percent are members of the above-mentioned churches, and 13 percent are Lutherans. The Bantu of the Orange Free State are more Christianized than those of the neighbouring Transvaal. Only 9 percent have no church affiliation; 19 percent are attached to Bantu Separatist Churches; and 46 percent are members of the six main churches.

Rural and urban settlement. The white population is dispersed over the country on individually owned farms and single homesteads. New discoveries of gold since World War II in the northwest have resulted in a maize and cattle country being turned into a mining and industrial belt. New towns, including Welkom (1970 population 132,000), Virginia (population 46,000), and Odendaalsrus (population 31,000), have sprung up. Welkom, the second largest town, has grown rapidly since 1950 and provides an example of modern planning, in which residential suburbs radiate from the town centre with green wedges of open spaces between them. Bloemfontein, the largest city, had about 148,000 inhabitants in 1970 and was the province's administrative, transport, educational, and cultural centre. It is also developing industrially. Kroonstad, with about 51,000 people, is the third largest town, while Sasolburg (29,000) is important for its burgeoning chemical industry. Most of the remaining towns are small, with some centres supporting a population of less than 10,000.

Administration. The Orange Free State is administered by its Provincial Council, the members elected every five years. The Council has jurisdiction on such matters as taxation to provide revenue for white education, hospitals, roads, and nature conservation. Local authorities consist of about 70 city or town councils, 10 village management boards, and 40 committees for the management of small-holding areas, and they derive their powers from the Provincial Council. Forty-nine magisterial districts control the administration of justice and, to a lesser extent, some functions of other governmental departments that do not justify offices in all districts. Of South Africa's four provinces, the Orange Free State has the smallest Bantu areas; located at Thaba 'Nchu and Basotho-Qwaqwa (formerly Basotho ba Borwa), they amount to 158,000 acres (64,000 hectares) or 0.5 percent of the total area of the province.

Social conditions. There are 13 provincial hospitals, two of them in the two Bantu homelands. Apart from the leper, mental, and tuberculosis institutions, the provincial and private hospitals and clinics have more than 1,500 and 2,000 beds for white and nonwhite patients, respectively. The province provides a wide range of health services, including a field service. Increased attention to health has led to a decline in infant mortality over the last decades.

Schooling, when available, is free for all and is compulsory for whites between 7 and 16 years of age. Where demand justifies the establishment of facilities, Coloureds between seven and 14 years are also compelled to attend schools; but Bantu education is not yet enforced, although Bantu literacy, which amounts to about 85 percent of the age group between seven and 20 years old, already has risen to a level considerably higher than in African territories outside the Republic. The University of the Orange Free State, located in Bloemfontein, has an enrollment of almost 3,000 white students. Administration of primary and secondary education and the training of teachers for whites are under provincial authority, while administration of the university and higher technical training are under the central government's Department of Higher Education. The central government is also responsible for nonwhite education.

Education

Provincial authorities, in cooperation with private organizations, undertake a wide variety of welfare services, including care for the aged, workshops and housing for the blind, and the provision of housing for the Bantu in larger urban centres.

Economy. The Orange Free State is not as rich in mineral resources as the neighbouring Transvaal. The gold resources of the Welkom district are the most important. Diamond production totalled 220,000 metric carats in 1968 alone. Estimated coal reserves of 3,400,000,000 tons are worked in the north and are mainly used in the coal-to-oil project at Sasolburg that supplies about 10 percent of the nation's gasoline requirements. Salt is derived from the pans of the semiarid region, and bentonite (a moisture-absorbing clay) is mined from a large deposit near Parys on the Vaal River.

The province forms part of the country's grain belt. The wetter area of the northeast grew about 30 percent of South Africa's annual maize output in 1965. In the northeastern and eastern districts, wheat is also grown, producing about 30 percent of the national crop for 1965. The undulating plains afford excellent grazing, and stock farming is important. The province produces almost one-quarter of South Africa's wool output, and its

Agricultural products

1,500,000 head of cattle are mainly located in the wetter northeastern districts of Harrismith, Vrede, and Bethlehem.

Water is the province's most essential resource. Apart from several storage dams—like the large Allenmanskraal, the Erfenis, and the Kalkfontein—the region will benefit enormously from the vast Hendrik Verwoerd and Van der Kloof dams, part of the great Orange River Project. The first was completed in 1972 and will provide water for Bloemfontein as well as for the gold-mining complex at Welkom; the second is due for completion in about 1974.

Transportation and communications. The state-owned railways provide a total of 1,678 track miles (231 of them electrified). Railways are supplemented by state, as well as privately owned, road motor service. There are over 2,000 miles of bitumen-surfaced national and provincial roads, exclusive of streets under the control of local authorities. The J.B.M. Hertzog Airport at Bloemfontein is the largest in the province and is of growing importance. Post office and telecommunication services expand from year to year to meet an increasing demand. The South African Broadcasting Corporation broadcasts to the province in English, Afrikaans, and several Bantu languages. There is a radio station at Bloemfontein. Television is to be installed in 1974.

Cultural life. The cultural life of the whites resembles that in Western countries. As the Orange Free State is predominantly a rural province, the more austere and moralistic aspects of the Afrikaner character are probably more in evidence here than elsewhere. The well-known hospitality, a general trait of white South African life, is a result of the stimulus of a mutual dependence that developed in an earlier era when life on the trekker's frontier was hard and sometimes dangerous. Bantu culture is still strongly influenced by tribal life; the supremacy of the chiefs is recognized, and the heritage of traditional animist religion is still in evidence, despite the influence of Christianity. The sunny climate influences all aspects of life, from clothing to food and drink and from literature to architecture. Cultural institutions such as museums and theatres are to be found in the principal towns.

BIBLIOGRAPHY. W. ALBERTYN (ed.), *Official South African Municipal Yearbook* (annual); BUREAU OF STATISTICS, REPUBLIC OF SOUTH AFRICA, *Population Census 1960*; M.M. COLE, *South Africa*, 2nd ed. (1966), physical and human geography; DEPARTMENT OF PLANNING, REPUBLIC OF SOUTH AFRICA, *Development Atlas* (1966), an extensive work containing maps and detailed descriptive information; A.K. HAAGNER, *South African Mammals* (1920); and with R.H. IVY, *Sketches of South African Bird Life*, 2nd ed. (1914), two standard works, illustrated; A.C. HARRISON *et al.*, *Fresh-Water Fish and Fishing in South Africa*, ch. 4 (1963), on species in the Vaal and Orange river systems; D.H. HOUGHTON, *The South African Economy*, 2nd ed. (1967), a general survey; L.C. KING, *South African Scenery*, 3rd ed. rev. (1963), geomorphology and topography; E. PALMER and N. PITMAN, *Trees of South Africa* (1961); N.C. POLLOCK and S. AGNEW, *An Historical Geography of South Africa* (1963); *State of South Africa Yearbook*, a general descriptive and statistical annual; J. VISSER, *Poisonous Snakes of Southern Africa* (1966), a descriptive classification with colour photographs; J.H. WELLINGTON, *Southern Africa*, vol. 1 (1955).

(J.N.S.)

Orange River

The Orange River in southern Africa, the sixth longest river on the continent and the longest south of the Tropic of Capricorn, rises as the Sinqu River in the Lesotho Highlands, flows westward across the veld as the Orange River, after which, also known as the Gariep River, it forms the southern border of the Kalahari before draining into the Atlantic Ocean at Alexander Bay, South Africa. Although the direct distance from its source to its mouth is less than 800 miles (1,300 kilometres), in its twisting, turning course it travels for 1,300 miles (2,100 kilometres). The river forms the boundary between two of South Africa's provinces, the Orange Free State and Cape Province, as well as that between South West Africa and the Republic of South Africa.

The drainage basin has an area of at least 329,000 square miles (852,000 square kilometres). The western part of the basin is generally dry, flat, and uncultivable without irrigation. The river itself is of vital economic importance to the region through which it flows. Two projects—the Orange River Project and the Oxbow Scheme—are planned to provide needed water for South Africa's arid plains and electrical power for its mines and industrial cities. (For an associated physical feature, see VELD; for historical background, see SOUTHERN AFRICA, HISTORY OF.)

The river's course. *The upper Orange.* The headwaters of the Orange River rise on the plateau formed by the Lesotho Highlands, where horizontal basaltic lavas have been cut into by numerous streams from the Drakenberg Escarpment in the east and the Maluti Mountains in the west. The main source is officially recognized as the Sinqu River, which rises near the plateau's eastern edge. The Khubedu headwater, which rises near Mont-aux-Sources to the north, is, however, better known and more accessible. Still farther north is the lesser known Madibamatso headwater, site of the Oxbow Scheme. Headwaters

The Lesotho headwaters flow over the turf soil that covers the lava, dropping 70 feet to 90 feet per mile. Below 29°40′ south, they have cut through the lava to expose sandstones; purple, red, and blue shales; and mudstones—rocks that contribute to heavy silt deposits farther down the river's course. To the west of the Lesotho boundary, the river flows south and west through more open country, where the sandstones, shales, and mudstones appear on the surface and where hard dolerite (basalt) outcrops form small hills and flat-topped mountains. Near Aliwal North, the river has eroded a broad valley with a width of some 30 miles and a depth of more than 1,000 feet. The river's channel, however, varies greatly in both width and depth because of dolerite outcrops that sometimes narrow its channel to 3,000 or 4,000 feet. Here, the average gradient is 3.5 feet per mile. The river receives the Caledon as a tributary near the town of Bethulie and then swings to the northwest.

The middle Orange. At its confluence with the Vaal, the river turns southwest and flows over limestone shales and tillite (glacial clayey deposit). At Prieska it makes another sharp bend—this time to the northwest—that marks the beginning of its middle course. Quartzites and ironstones form a "barrier zone" through which the river has cut deep *poorts*, or gorges, such as the Groot and Klein Noute (Great and Little Narrows). The gradient in this tract is about 2.5 feet per mile. At Upington the river spreads out over a granite surface to a width of about a mile. In this area the river splits up into innumerable channels, between which are islands varying in length from a few hundred yards to six or seven miles. In this stretch the river attains its greatest width—nearly four miles in places. About 40 miles downstream from Upington, however, the river bed is suddenly narrowed to about 700 yards by a bed of quartzite that forms the Neus Berg Ridge.

The lower Orange. Some 20 miles below Kakamas the river—again flowing in several channels—forms the Augrabies Falls. There, after descending about 80 feet in a series of rapids, the river plunges some 400 feet into a pool more than 140 feet deep. The river flows through an almost vertical-sided gorge, with some right-angled bends, for about 11 miles, emerging again into more open country. The lower course of the river, from the Augrabies Falls to the sea, is sometimes called the Gorge Tract. Where the rock surface is soft, the river valley is generally open. Where the river traverses harder quartzite granodiorite (a granular igneous rock) and diabase (a form of basalt), however, it is confined between almost vertical cliffs more than 1,000 feet high in places. Some of the most rugged passages are found in the last great bend of the river, as it flows north along the Richtersveld before turning west to the coastal Namib Desert. The Gorge Tract

The river reaches the sea a few miles north of the little inlet known as Alexander Bay. The mouth is about 2.5 miles wide and is nearly closed by a sandbar, which is widely breached, however, during high floods. The

gap in the southern end of the bar is maintained by the outflow of water from the river mouth during low tides and by the tidal inflow at high tides.

The riverine population. The high valleys of the Orange River's headwaters are uninhabited and are used by the Sotho (Basuto) people for grazing land. Between the Lesotho border and the town of Aliwal North, maize (corn) is cultivated on the valley's grassland, which is also used as pasturage for cattle and sheep by the South African white population. To the west, the river passes through dry shrub country that is in general suitable only for grazing. Some irrigated sections, however, occur along the river's course; the largest such areas are at the Boegoebergdam, where one-half of South Africa's cotton is grown, and between Augrabies Falls and Upington. While scattered, white-owned farms stand within reach of the river's freshwater supply, there are no large towns along the riverbank. This situation will remain little changed by the Orange River Project, as most of the water accumulated by its three large dams will be diverted to the valleys of the Fish and Sundays rivers and to the cities of Port Elizabeth, Bloemfontein, and Kimberley.

History. The first white man known to cross the river to the north bank was an Afrikaner elephant hunter, Jacobus Coetsee, who forded the Groot River, as it was then called, near the river mouth in 1760. Later expeditions across the river in the 18th century were led by Capt. Hendrik Hop; Capt. Robert Jacob Gordon, a Dutch officer; Lieut. William Paterson, an English traveller; and François Le Vaillant. They explored the river from its middle course to its mouth, and Gordon named it in honour of the Dutch House of Orange.

Throughout the 19th century, the Orange River marked the northern limit of British power in southern Africa. Beginning in the 1830s, the Boers crossed it in search of land and freedom from British rule; they named their first republic—the Orange Free State—after the river.

Climate

Hydrology. There are two distinct rainfall patterns in the river basin; each is directly related to the rate of flow. Above the Vaal River confluence, the river receives the heavy summer rainfall of 60 to 80 inches and the melted winter snows of the Lesotho Highlands and the western Maluti Mountains, as well as an annual rainfall of about 18 inches that occurs along the eastern basin. Although this section of the river comprises only 7 percent of the area of the entire catchment basin, it contributes 58 percent of the river's total flow. Downstream at Prieska, the Orange has an annual runoff of 5,500,000 acre-feet.

Below the Vaal confluence, rainfall decreases from nine inches a year to less than two inches in the coastal Namib Desert. Temperatures increase in a westerly direction and result in a higher rate of evaporation. The total annual runoff of water from the remaining 215,000 square miles of the catchment basin is reduced to 165,000 acre-feet.

Development and projects. *Navigation and bridges.* Navigation is impossible throughout the river's course because of its irregular flow, its constant interruption by falls and rapids, and the silting that occurs in its channels and at the river mouth. Many bridges cross the river along its course between Aliwal North and Oranjemund, the largest being at Upington.

Dams and future projects. Most damming is accomplished by privately owned weirs constructed for irrigation. Larger irrigation and hydroelectric projects are hampered by the enormous amount of water-borne silt that clogs up reservoirs, such as that of the concrete Boegoeberg Dam, and reduces storage capacity. Completed in 1931, it is located midway between Prieska and Upington where the river cuts through a hard, quartzite outcrop. With a storage capacity of 43,500 acre-feet, it irrigates about 42,000 acres over a distance of 150 miles.

The Orange River Project

To obtain comprehensive control, as in the Orange River Project, however, it is necessary to locate projects farther upstream. Below the Vaal River confluence the flow is too sluggish. The Orange River Project, therefore, is located between the Caledon and Vaal confluences. The plan consists of six consecutive projects that are expected to take 30 years to complete; work began in 1962. The projects include the Hendrik Verwoerd Dam (already completed), which will form the main reservoir, located just west of the Caledon confluence; the Van der Kloof Dam, with irrigation canals to both banks, 70 miles downstream; the Torquay Dam, between Hopetown and Douglas; a 52-mile tunnel to carry water from the Verwoerd Dam to the Great Fish River; an irrigation canal between the Fish and Sundays rivers; and another irrigation canal from Torquay Dam to the Kalkonfontein Dam, on the Vaal River. The three major dams will be able to store 37,000,000 acre-feet, and total storage will reach 105,000,000 acre-feet.

The new dams will be equipped to deal with heavy silt accumulation. They will be provided with deep dead-storage areas, and their high walls will be raised if necessary to allow for a higher water level. The impounded water is to irrigate some 762,000 acres of land along the Orange River and in tributary valleys extending as far west as the Sak River in northwestern Cape Province, as well as the Witsands area in the southeastern Kalahari. The project also will supply an area of 54,000 square miles in northern Cape Province and southern Orange Free State with an annual power output of about 6,000,000,000 kilowatt-hours.

Near the river's headwaters in Lesotho, another project, the Oxbow Scheme, is still in the planning stage. It is to be Lesotho's major irrigation project; South Africa will have the opportunity of purchasing large quantities of the excess water and power. The plan calls for the damming of the Madibamatso headwater about 11 miles west of Mont-aux-Sources in the Maluti Mountains. It will produce 350,000,000 kilowatt-hours of electric power a year and will supply the Orange Free State with 40,000,000 gallons of water daily.

BIBLIOGRAPHY. The earliest, and in many ways the most interesting, account of the Orange River is *The Journal of Hendrik Jacob Wikar* (1779), written in Dutch, with an English translation by A.W. VAN DER HORST (1935). There were other travellers' accounts of the river in the 19th century, but it was not until the early 20th century that a concerted attempt was made to explore economic potentialities. A.D. LEWIS traversed the river's course on foot from Pella Drift to the river mouth, recording his impressions in his "Report on Flying Reconnaissance of the Lower Orange River," *Report of the Director of Irrigation 1912–1913* (1914). Later publications include some geographical studies, one of the most comprehensive being that of J.H. MOOLMAN, "The Orange River, South Africa," *Geogrl. Rev.*, 36:653–674 (1946), illustrated. Other geographical studies include those of J.H. WELLINGTON: "The Middle Course of the Orange River," *S. Afr. Geogr. J.*, 16:58–68 (1933); "The Evolution of the Orange River Basin: Some Outstanding Problems," *S. Afr. Geogr. J.*, 40:3–30 (1958); and his book *Southern Africa* (1955), in which ch. 12 of vol. 1 is concerned with many aspects of the river and its tributaries. Less formal accounts include LAWRENCE GREEN's entertaining *To the River's End* (1948), which contains episodes and legends relating to riverine settlements. Recent literature has been connected mainly with the Orange River Project; this includes a number of papers given at the symposium arranged by the South African Association for the Advancement of Science and are contained in the Orange River number of the *S. Afr. J. Sci.*, vol. 61 (1965), where the subjects include the geological and geomorphological character of the upper part of the basin, the hydroelectrical potential, plant ecology, the freshwater fishes, and other aspects. Government publications include the *Report on the Proposed Orange Development Project* (1963), and a comprehensive account issued in an 80-page supplement, "Dykes Against Drought," *Financial Mail* (Johannesburg, April 25, 1969).

(J.H.We.)

Oratory

Oratory is the rationale and practice of persuasive public address. As a form of oral communication, it is immediate in its audience relationships and reactions, but it may also have broad historical repercussions. The orator may become the voice of political or social history. Thus Rufus Choate, an American legal advocate of the early 19th century, said, in a tribute to Daniel Webster,

> It is a peculiarity of some schools of eloquence that they embody and utter, not merely the individual genius and charac-

ter of the speaker, but a national consciousness,—a national era, a mood, a hope, a dread, a despair,—in which you listen to the spoken history of the time.

A vivid instance of the way a speech can focus the concerns of a nation was Martin Luther King's address to a massive civil rights demonstration in Washington, D.C., in 1963. Repeating the phrase "I have a dream," King applied the oratorical skill he had mastered as a preacher to heighten his appeal for further rights for U.S. Negroes to an intensity that galvanized millions.

Principles of oratory. The essentials of a given oration are a speaker; an audience; a background of time, place, and other conditions; a message; a transmission process by voice, articulation, and bodily accompaniments; and immediate and larger outcomes of the speaking event.

Rhetoric is the art of using words effectively, while oratory is the practice of that art in public speech. Oratory, like rhetoric, is instrumental and practical, as distinguished from poetic or literary composition, which traditionally aims at beauty and pleasure. Rhetoric and oratory are of the marketplace and are not so much concerned with the universal and permanent. The orator in his purpose and technique is primarily persuasive rather than informational or entertaining. He attempts to change human behaviour or to strengthen convictions and attitudes. The orator would correct wrong positions of the audience and establish psychological patterns favourable to his own wishes and platform. He uses argument and rhetorical devices: evidence and lines of reasoning, and appeals that support his aims. He uses exposition to clarify and enforce his propositions. He includes anecdotes and illustrations to heighten response.

Union of thought and feeling

The orator need not be a first-rate logician, though a capacity for good, clear thought helps to penetrate into the causes and results of tentative premises and conclusions, and to utilize analogy, generalizations, assumptions, deductive-inductive reasoning, and other types of inference. Effective debaters, who depend more heavily on logic, however, are not always impressive orators because superior eloquence also requires strong appeals to the motives, sentiments, and habits of the audience. Audience behaviour is guided by imaginative and emotional suggestion at each phase of the address. Oratorical greatness is invariably identified with strong emotional phrasing and delivery. When the intellectual qualities dominate with relative absence of the affective appeals, the oration fails just as it does when emotionality sweeps aside reason.

The ideal orator is personal in his appeals and strong in ethical proofs, rather than objective or detached. He enforces his arguments by his personal commitment to his advocacy. William Pitt, later Lord Chatham, punctuated his dramatic appeals for justice to the American colonies with references to his own attitudes and beliefs. So were personal appeals used by the Irish orator Daniel O'Connell, the French orators Mirabeau and Robespierre, and the Americans Wendell Phillips, Robert G. Ingersoll, and Woodrow Wilson.

The eloquent speaker looks to the basic foundations of reasoning and emotive expression. His utterances are concrete, but they argue broad principles. The orator, as illustrated by Edmund Burke, has a catholic attitude. Burke's discussion of American taxation, conciliation, Irish freedoms, justice for India, and the French Revolution show analytical and intellectual maturity, the power of apt generalization, and comprehensiveness of treatment.

Types of oratory. Oratory has traditionally been divided into legal, political, or ceremonial, or, according to Aristotle, forensic, deliberative, or epideictic.

Forensic oratory. Typically, forensic, or legal, oratory is at its best in the defense of individual freedom and resistance to prosecution. It was the most characteristic type of oratory in ancient Athens, where laws stipulated that litigants should defend their own causes. In the so-called Golden Age of Athens, the 4th century BC, great speakers, in both the law courts and the assembly, included Lycurgus, Demosthenes, Hyperides, Aeschines, and Dinarchus.

In the first century BC of ancient Rome, Cicero became the foremost forensic orator and exerted a lasting influence on later Western oratory and prose style. Cicero successfully prosecuted Gaius Verres, notorious for his mismanagement while governor of Sicily, and drove him into exile, and dramatically presented arguments against Lucius Sergius Catiline that showed a command of analysis and logic, and great skill in motivating his audience. Cicero also delivered 14 bitter indictments against Mark Antony, who was to him the embodiment of despotism.

Among the great forensic orators of later times was the 18th- and 19th-century English advocate Thomas Erskine (1750–1823), who contributed to the cause of English liberties and the humane application of the legal system.

Deliberative oratory. According to Aristotle, the chief problems for deliberation and policy settlement are those involved with the domestic and foreign affairs of the government, and its economic concerns. Demosthenes, the Athenian lawyer, soldier, and statesman, was a great deliberative orator. In one of his greatest speeches, "On the Crown," he defended himself against the charge by his political rival Aeschines that he had no right to the golden crown granted him for his services to Athens. So brilliant was Demosthenes' defense of his public actions and principles that Aeschines, who was also a powerful orator, left Athens for Rhodes in defeat.

Epideictic oratory. The third division of persuasive speaking, epideictic, or ceremonial, oratory was panegyrical, declamatory, and demonstrative. Its aim was to eulogize an individual, a cause, occasion, movement, city, or state, or to condemn them. Prominent in ancient Greece were the funeral orations in honour of those killed in battle. The outstanding example of these is one by Pericles, perhaps the most finished orator of the 5th century BC, in honour of those killed in the first year of the Peloponnesian War.

Webster—master of all forms

The 19th-century American speaker Daniel Webster excelled in all three major divisions—forensic, deliberative, and epideictic oratory. He brought more than 150 pleas before the U.S. Supreme Court, including the Dartmouth College Case (1819) and the *Gibbons v. Ogden* case (1824); he debated in the U.S. Senate against Robert Young Hayne and John Calhoun on the issues of federal government versus states' rights, slavery, and free trade, and he delivered major eulogies, including those on the deaths of Jefferson and John Adams.

Religious oratory. Another major type of persuasive speaking that developed later than ancient Greek and Roman rhetoric was religious oratory. The decline of Roman republican government, with its Senate and Forum, saw the passing of significant political oratory. Roman law and order prevailed, but there was little scope for provocative public address. For more than 1,000 years after Cicero the important orators were churchmen rather than politicians, lawyers, or military spokesmen. This tradition derived from the Judean prophets, such as Jeremiah and Isaiah, and in the Christian Era, from the Apostle Paul, his evangelistic colleagues, and such later fathers of the church as Tertullian, Chrysostom, and St. Augustine. Ecclesiastical speaking became vigorously polemical. The rhetorical principles of Aristotle and Cicero were adopted by ecclesiastical leaders who challenged rival doctrines and attacked the sins of the communities.

In the Middle Ages, Pope Urban II elicited a great response to his oratorical pleas for enlistment in the First Crusade. The Second Crusade was urged on with great eloquence by Saint Bernard, abbot of Clairvaux. In the 15th and 16th centuries the revolt against the papacy and the Reformation movement stimulated the eloquence of Huldrych Zwingli (1484–1531), John Calvin (1509–64), Hugh Latimer (*c.* 1485–1555), and, most notably, Martin Luther (1483–1546). At the Diet of Worms, as elsewhere, Luther spoke with courage, sincerity, and well-buttressed logic. Religious controversies in the 17th century engaged such great oratorical skills as those of Richard Baxter (1615–91), the English Puritan, and Catholic Bishop J.B. Bossuet (1627–1704) of France. In the 18th century the Methodist George Whitefield (1714–70) in England and America, and the Congregationalist Jonathan Edwards

(1703–58) in America, were notably persuasive speakers. Preachers of oratorical power in the 19th century included Henry Ward Beecher (1813–87), famous for his anti-slavery speeches and his advocacy of women's suffrage from his Congregational pulpit in Plymouth Church, Brooklyn, New York, and William Ellery Channing (1780–1842), American spokesman for Unitarianism.

Oratory as a reflection of its audience. Because the orator intuitively expresses the fears, hopes, and attitudes of his audience, a great oration is to a large extent a reflection of those to whom it is addressed. The audience of Pericles in ancient Greece, for example, was the 30,000 or 40,000 citizens out of the state's total population of 200,000 or 300,000, including slaves and others. These citizens were sophisticated in the arts, politics, and philosophy. Directing their own affairs in their Assembly, they were at once deliberative, administrative, judicial. Speaker and audience were identified in their loyalty to Athens. Similarly, the senatorial and forum audience of Cicero in ancient Rome was an even smaller elite among the hundreds of thousands of slaves and aliens who thronged the Roman world. In the Forum the citizens, long trained in law, and with military, literary, and political experience, debated and settled the problems. The speeches of Cato, Catiline, Cicero, Julius Caesar, Brutus, Antony, Augustus, and the others were oratory of and for the Roman citizen.

In the Christian Era, however, the religious orator often found himself addressing an alien audience that he hoped to convert. To communicate with them, the Christian often appealed to ancient Greek and Roman thought, which had achieved widespread authority, and to Judean thought and method, which had the sanction of scripture. By the time of the Reformation, however, Christian dogma had become so codified that most of the disputation could be carried on in terms of doctrine that had become well known to all.

Trend toward common speech

The history of the British Parliament reveals a continuing trend toward common speech and away from the allusions to ancient Greek and Roman thought that abounded when the members consisted largely of classically educated aristocrats.

In the golden age of British political oratory of the late 18th century, greater parliamentary freedom and the opportunity to defend and extend popular rights gave political oratory tremendous energy, personified by such brilliant orators as the elder and younger William Pitt, John Wilkes, Charles James Fox, Richard Sheridan, Edmund Burke, and William Wilberforce. Parliamentary reforms of the 19th century, initiated and promoted by Macaulay, Disraeli, Gladstone, and others of the century, led to more and more direct political speaking on the hustings with the rank and file outside Parliament. Burke and his contemporaries had spoken almost entirely in the Commons or Lords, or to limited electors in their borough homes, but later political leaders appealed directly to the population. With the rise of the Labour Party in the 20th century and the further adaptation of government to the people, delivery became less declamatory and studied. The dramatic stances of the 18th-century parliamentary debaters disappeared as a more direct, spontaneous style prevailed. As delivery habits changed, so did the oratorical language. Alliteration, antithesis, parallelism, and other rhetorical figures of thought and of language were sometimes carried to extremes, in speeches addressed to those highly trained in Latin- and Greek-language traditions. These devices gave way, however, to a clearness of style and vividness consonant with the idiom of the common man and later with the vocabulary of radio and television.

Similarly, American speech inherited and then gradually discarded British oratorical techniques for its own speaking vernacular. John Calhoun, in his addresses to Congress on behalf of the South, absorbed much of the Greek political philosophy and methods of oral composition and presentation, and his principal opponent in debate, Daniel Webster, too, had the marks of British communicative tradition. This inheritance was absorbed into the speaking adjustments indigenous to those later peoples of New England, the West, and the South. The orator whose speech preceded Lincoln's at Gettysburg—Edward Everett, statesman and former professor of Greek literature at Harvard—was a classical scholar. Lincoln, on the same platform, had address born of his native Middle West yet expressed with authentic eloquence.

In the 20th century, the era culminating in World War II saw the development of two leaders who applied oratorical techniques in vastly different ways with equal effect. It was primarily through his oratory that Adolf Hitler whipped the defeated and divided Germans into a frenzy of conquest, while Winston Churchill used his no less remarkable powers to summon up in the English people their deepest historical reserves of strength against the onslaught. Subsequently, though the importance of persuasive speech in no way diminished, television so reshaped the method of delivery that much of the theory of traditional oratory often seemed no longer to apply. In the televised debates of John F. Kennedy and Richard Nixon during the U.S. presidential campaign in 1960, for example, the candidates might be said to have been most persuasive when they were least oratorical, in the traditional sense of the term. Nonetheless, even conventional oratory persisted as peoples in newly developing nations were swept up into national and international political struggles.

BIBLIOGRAPHY. General collections include: D.J. BREWER (ed.), *Crowned Masterpieces of Eloquence That Have Advanced Civilization, as Presented by the World's Best Orations, from the Earliest Period to the Present Time*, 10 vol. (1908); T.B. REED *et al.* (eds.), *Modern Eloquence*, 10 vol. (1901), rev. by A.H. THORNDIKE, 12 vol. (1923); and H. PETERSON (ed.), *A Treasury of the World's Great Speeches*, rev. ed. (1965), designed for classroom study.

Individual orators and periods: R.C. JEBB, *The Attic Orators from Antiphon to Isaeos*, 2nd ed., 2 vol. (1893); J.F. DOBSON, *The Greek Orators* (1919); DEMOSTHENES, *Orations*, trans. by C.R. KENNEDY, 5 vol. (1852–63); PLUTARCH, *Lives*, trans. by JOHN DRYDEN, rev. by A.H. CLOUGH (1932); CICERO, *The Orations*, 4 vol., trans. by C.D. YONGE (1913–21); F.V.A. AULARD, *Les Grands orateurs de la Révolution* (1914); D.C. BRYANT *et al.* (eds.), *An Historical Anthology of Select British Speeches* (1967); C.K. ADAMS, *Representative British Orations*, 4 vol. (1900); A. JOHNSTON (ed.), *Representative American Orations to Illustrate American Political History*, 3 vol. (1884); A. CRAIG BAIRD (ed.), *Representative American Speeches*, annually (1937–59); L.W. THONSSEN (ed.), *ibid.* (1960–70); and R.T. OLIVER, *History of Public Speaking in America* (1965).

Theory of oratory and rhetorical-oratorical criticism: C.S. BALDWIN, *Ancient Rhetoric and Poetic* (1924); D.C. BRYANT (ed.), *The Rhetorical Idiom* (1958); W.N. BRIGANCE (ed.), *History and Criticism of American Public Address*, 2 vol. (1943); M.K. HOCHMUTH (ed.), *A History and Criticism of American Public Address*, vol. 3 (1955); L.W. THONSSEN and A. CRAIG BAIRD, *Speech Criticism: The Development of Standards for Rhetorical Appraisal* (1948); G. KENNEDY, *The Art of Persuasion in Greece* (1963); R.F. HOWES (ed.), *Historical Studies of Rhetoric and Rhetoricians* (1961); A. CRAIG BAIRD (ed.), *Goodrich's Essays from Select British Eloquence* (1963). Important also are the contributions of later and contemporary theorists, including H. BLAIR, *Lectures on Rhetoric and Belles Lettres*, ed. by H.F. HARDING, 2 vol. (1965); G. CAMPBELL, *The Philosophy of Rhetoric*, ed. by L.F. BITZER (1963); JOSEPH PRIESTLEY, *A Course of Lectures on Oratory and Criticism*, ed. by V.M. BEVILACQUA and R. MURPHY (1965); R. WHATELY, *Elements of Rhetoric*, ed. by D. EHNINGER (1963); H.H. WICHELNS, "The Literary Criticism of Oratory," in *Studies in Rhetoric and Public Speaking*, ed. by A.M. DRUMMOND, pp. 181–216 (1925, reprinted 1962); E.T. CHANNING, *Lectures Read to the Seniors in Harvard College*, ed. by D.I. ANDERSON and W.W. BRADEN (1968); I.A. RICHARDS, *The Philosophy of Rhetoric* (1936); K.D. BURKE, *A Rhetoric of Motives* (1950); and A. CRAIG BAIRD, *Rhetoric: A Philosophical Inquiry* (1965).

(A.C.B.)

Orchestration and Instrumentation

Most authorities make very little distinction between the terms orchestration and instrumentation. Both deal with musical instruments and their capabilities of producing

various timbres or colours. Orchestration is somewhat the narrower term since it is frequently used to describe the art of instrumentation as related to the symphony orchestra. Instrumentation, therefore, is the art of combining instruments in any sort of musical composition, including such diverse elements as the numerous combinations used in chamber groups, jazz bands, rock ensembles, ensembles employing chorus, symphonic bands, and, of course, the symphony orchestra. Included under this designation are the various instrumental groups that play non-Western music, such as the gamelan orchestras of Bali and Java and the traditional ensembles of India, Africa, the Far East, and the Middle East.

Constituents of the orchestra

In Western music there are many standard or traditional groups. The modern symphony orchestra usually comprises the following instruments, although they are not necessarily used in every composition:

1. Woodwinds: four flutes (one doubling, or duplicating the part of the piccolo), four oboes, English horn (cor anglais), three clarinets, bass clarinet, three bassoons, contrabassoon (double bassoon);
2. Brass: four trumpets, four or five French horns, three trombones, tuba;
3. Strings: two harps, first and second violins, violas, violoncellos, double basses;
4. Percussion; four timpani (played by one player) and at least three general percussion players.

The orchestra has arrived at this complement through centuries of evolution; the present size is needed to perform repertory from the Baroque, Classical, Romantic, and Impressionistic periods, as well as the repertory of the 20th century.

The various sections, with the exception of percussion, divide themselves in somewhat the same manner as a choir. The woodwinds, for example, divide into flutes (sopranos), oboes (altos), clarinets (tenors), and bassoons (basses), although this distinction must be greatly qualified. Instrumental range is larger than vocal range, and the clarinets of an orchestra may play higher than the flutes in a woodwind passage.

The standard instrumental groups of Western chamber music include the string quartet (two violins, viola, and violoncello), the woodwind quintet (flute, oboe, clarinet, French horn, and bassoon), the combinations employed in sonatas (one wind or string instrument with piano), and the brass quintet (frequently two trumpets, French horn, trombone, and tuba). In addition to these standard groups there are, however, hundreds of other possible combinations.

Other groups that deserve mention are those used in the popular music of the 20th century. The dance band, popular in the 1930s and 1940s, consisted of five saxophones, four trumpets, four trombones, double bass, piano, guitar, and drums. The basic rock ensemble consists of two electric guitars, electric bass, electric organ (doubling electric piano), drums, and frequently includes one or more singers. The concert band, which is particularly popular in North America, consists of mixed wind and percussion players totalling from about 40 to well beyond 100 players.

The music of the non-Western world is most frequently performed by groups of chamber music size. In this category would fall the music played by the Javanese gamelan orchestra (consisting mainly of tuned gongs and other metal instruments), Japanese *gagaku* music (performed on flutes, mouth organs, lutes, drums, and gongs), and Chinese music (with a traceable history of about 4,000 years) consisting of sacred, folk, chamber, and operatic music.

TYPES OF INSTRUMENTATION

The approach to the art of instrumentation is naturally greatly influenced by the type of group for which the composer is writing. He cannot treat a string quartet or a group of brass instruments in the same manner as he would a symphony orchestra. In general, the larger and more diverse the instrumental group, the more coloristic possibilities it presents to the composer. The smaller instrumental groups often have a sound character of their own, and the composer is challenged to find new and interesting ways to deal with this limitation.

The symphony orchestra has had definite traditions in relation to orchestration. The composer of the 18th century was likely to use the orchestral instruments at least part of the time in the following manner: the flutes doubling the same part as the first violins (frequently the melody); the oboes doubling the second violins or the first violins in octaves; the clarinets (by the end of the century) doubling the violas; and the bassoons doubling the violoncellos and double basses. French horns were often used as harmonic "filler" and in conjunction with every section of the orchestra because of their ability to blend easily with both string and wind instruments.

These traditional doublings were not so often used in the orchestration of the 19th and 20th centuries because of the great improvement in the making of wind instruments and their consequent ability to function in a solo capacity. Wind instruments became used more and more for colouring; the flutes, for instance, were noted for their bright tone quality and great technical agility, the clarinets for all the aforementioned qualities, and the bassoons for their special tone quality. Brass instruments had to await the development of valves, which increased greatly the musical proficiency of brass players and overcame previous typecasting of these instruments as bugles and hunting horns.

Development of wind instruments

String techniques. The string quartet has long been considered one of the greatest challenges to the composer because the contrast to be achieved by changing from one type of instrument when writing for a full orchestra is simply not available. The composer has had to rely on varying timbres to be arrived at by different playing techniques, such as pizzicato (plucking the strings), tremolo (the quick reiteration of the same tone), sul ponticello (bowing near the bridge of the instrument), sul tasto (bowing on the fingerboard), the use of harmonics (dividing the string in such a way as to produce a high flute-like tone), col legno (striking the strings with the wood of the bow), and many special bowing techniques.

Wind techniques. Special playing techniques also can alter the timbres of wind instruments. For instance, on many, tremolos can be played on two different notes. Some wind instruments—and the flute is particularly agile in this respect—can produce harmonics. Flutter tonguing (produced by a rapid rolling movement of the tongue) is possible on most wind instruments; so are many other tonguing techniques that affect the quality of sound in orchestration.

Muting. The string mute is a device that softens the tone of the instrument. Muting is also used by brass instruments, particularly the trumpet and trombone, a development that took place in 20th-century popular music and then came into common use in all types of music. Mutes—of which there are various kinds—provide the trumpet and trombone with a different tone colour. Mutes on woodwind instruments have been experimented with, but the results have not been satisfactory.

Percussion instruments. Percussion instruments have become a favourite source of colour for the 20th-century composer, both in the concert and popular fields. Instruments from all over the world are now commonly available and are divided into two categories: of definite and of indefinite pitch. The former include the xylophone, marimba, vibraphone, glockenspiel, timpani, and chimes. Instruments of indefinite pitch exist by the hundreds. Some of the more common ones are the snare drum, tenor drum, tom-tom, bass drum, bongos, Latin American *timbales*, many types of cymbal, maracas, claves, triangle, gongs, and temple blocks.

The availability of these instruments and the great improvement in percussion playing has resulted in an enormous increase in the number of compositions for percussion instruments. The percussion ensemble, a group of from four to eight players, is a chamber group that has existed only in the 20th century, particularly since the late 1940s. One of the interesting features of such an ensemble is that each player in it is capable of playing many instruments. An ensemble of four players, for in-

stance, can easily handle 25 or 30 instruments, once again showing the rich palette available in a single composition.

Keyboard instruments. Since the 17th century, keyboard instruments have played an important role in orchestration. Those commonly available today are the harpsichord, celesta, organ (both pipe and electric), and electric piano, in addition to the instrument for which most of the standard literature has been written—the piano. Keyboard instruments vary greatly in the manner in which they produce a sound: the harpsichord has quills that pluck the strings; the piano has hammers that strike the strings; the celesta has hammers that strike a metal bar; the pipe organ sends air through a pipe; the electric organ employs electronic oscillators to produce its sound. The resulting colours are naturally very different.

The piano, with its wide range (more than seven octaves), has been used in conjunction with virtually every instrument and instrumental combination. In the 18th century it gradually replaced the harpsichord as the common keyboard instrument because of the piano's ability to alter dynamics rapidly and its ability to sustain sounds. There is a vast amount of literature for the piano as the accompanying instrument in sonatas, partly because the piano can function as a "one-man orchestra." Many composers of the 20th century have discovered facets of the piano that had been previously ignored. The inside of the grand piano is a harplike body that has presented many new possibilities to the composer, such as the "prepared" piano. To prepare a piano, objects such as bolts, pennies, and erasers are inserted between the strings, thus producing many different sounds. The piano strings can be plucked or played with percussion mallets and can produce harmonics in the manner of non-keyboard stringed instruments, much to the dismay of piano tuners and traditional pianists.

The "prepared" piano

Electronic instruments. The electric piano is one of a number of instruments that have gained in popularity in recent times. These instruments either produce sound by means of electronic oscillators or are amplified acoustic instruments. The sound produced by ensembles playing this type of instrument is distinctive. The rock ensemble is the best known, but rock musicians are by no means the only instrumentalists to employ electric instruments. For the composer, amplified or electric instruments pose certain problems. Balances can be achieved or ruined simply by turning an amplifier up or down. The timbres produced by rock ensembles and other groups employing electronics are unusual for a number of reasons. The electric guitar has such devices as reverberation controls, "wa-wa" pedals, and filters that enable the performer to change timbre radically in the middle of a performance. Composers since the early 1960s, being much concerned with coloristic possibilities of instruments, have found the electronic ones most attractive.

Voices. The largest quantity of literature in Western music has been written for the chorus. The choir, an instrument capable of great subtleties of colour, has been a favourite of composers for centuries. The range of most individual singing voices is rather limited. Choral singers, who usually have a limited amount of training, are capable of a range of about an octave and a fifth, which is considerably smaller than the range of individual instruments. Singers are usually not capable of singing wide leaps, that is to say, notes that are far apart in range. Great skill is required in the musical setting of the text in a choral work. Attention must be paid to the vocal qualities of vowel sounds as well as to the way in which the consonants are treated.

For centuries composers have been intrigued with the combination of voices and instruments, and many of the most important compositions in Western music have been written for chorus and orchestra. Almost every major composer of the past three centuries has written for choir and large instrumental ensembles.

HISTORY AND DEVELOPMENT OF WESTERN INSTRUMENTATION

The development of the art of using instruments for their individual properties did not really begin in Western music until about 1600. The known history of musical instruments, however, has been traced back 40,000 years, although nothing is known about the music these early instruments produced. The Greeks have left mostly musical theories and only a very small amount of extant music. The Romans used instruments particularly in military bands, but, again, little is known of their specific use. The music of the Middle Ages and Renaissance was primarily vocal, although instruments were frequently used in compositions to accompany or reinforce the individual vocal line. String, brass, woodwind, and percussion instruments were added not so much for their coloristic potential but because of their availability. Another practice in the Middle Ages was to make literal instrumental versions of vocal compositions, which, of course, has rather little in common with the modern art of instrumentation.

Baroque. Orchestration in a modern sense probably began in the 16th century with Giovanni Gabrieli, organist of St. Mark's in Venice. He was the first composer to sometimes designate specific instruments for each part in a composition, as in his *Sacrae symphoniae* (1597). Claudio Monteverdi, one of the great composers in Western music, made important contributions to the art of orchestration. His opera *Orfeo* was first performed at Mantua (now Mantova, Italy) in 1607 with an orchestra of about 40 instruments including flutes, cornetts, trumpets, trombones, strings, and keyboard instruments. For the first time, a composer, in order to heighten certain dramatic moments, specified exactly which instruments were to be used.

Monteverdi's orchestra

The century after the first performance of *Orfeo* was characterized by a rise in the use of string instruments that were similar to the modern ones. Although that trend helped set the stage for the modern orchestra, it was not a period that made great strides in the art of orchestration: the prevalent practice of writing out only the melody and the bass line of a composition did not lend itself easily to creative scoring. By the end of the 17th century, however, the groundwork had been laid for new developments. Instruments and instrumentalists had improved steadily. Johann Sebastian Bach created works that occasionally exploited the coloristic capabilities of instruments, but in a rather limited way. In some of Bach's music the stringed instruments are played pizzicato, although this practice had already been employed by Monteverdi. Bach also wrote for muted strings. Wind instruments were treated occasionally for their special sounds, although more frequently they were simply employed on a musical line that their range happened to fit.

Handel, whose life covered the same period as Bach's, had a keener sense of orchestral effect. He introduced the clarinet into his orchestra, although it was not to become standard until the 19th century, and in his operas Handel often used instrumental colour in a way that did not become common practice until much later. Jean-Philippe Rameau, the leading French composer of the 18th century, also contributed much to the development of orchestration. Rameau, like Handel, was principally famous as an opera composer, and the overtures and dances of his operas represent the most advanced uses of instruments during that period. Rameau was probably the first composer to treat each instrument of the orchestra as a separate entity, and he introduced interesting and unexpected passages for the flutes, oboes, and bassoons.

By the middle of the 18th century the symphony orchestra was beginning to resemble the modern instrumental group, yet it was still considerably smaller. The orchestra at the court of Mannheim, Germany, consisted of 20 violins, four violas, four cellos, two double basses, two flutes, two oboes, two bassoons, four French horns, one trumpet, and kettledrums. Baroque composers frequently could not count on a fixed orchestra and therefore had to write the various parts so that they could be played on more than one instrument. The contrapuntal style that prevailed from the time of Monteverdi until the mid-18th century usually meant simply assigning instruments to each line in a composition; the basic considera-

The court orchestra in Mannheim

tion was whether that line stayed within the range of the chosen instrument. The fixed personnel of such orchestras as the Mannheim group, therefore, freed the composers to experiment with the capabilities of the instruments within the group. Musical style was also changing, the contrapuntal style of the Baroque giving way to a style that relied more heavily on melodic invention supported by harmony.

One of the more important composers of the period between the Baroque and Classical eras was Johann Sebastian's son, Carl Philipp Emanuel Bach. In C.P.E. Bach's symphonies the strings become melodic instruments, and the winds—two flutes, two oboes, one or two bassoons, two horns—fill out chords and provide body to the orchestration.

Classical. The Classical period, which covers roughly the second half of the 18th century, is one of the most significant periods in the development of orchestration. The most talented composers of this period were Mozart and Haydn. Many important developments took place during this time. The orchestra became standardized. The Classical orchestra came to consist of strings (first and second violins, violas, violoncellos, and double basses), two flutes, two oboes, two clarinets, two bassoons, two or four French horns, two trumpets, and two timpani. Toward the end of his career, in the *London Symphonies*, Haydn introduced clarinets as part of the woodwind section, a change that was to be permanent. Haydn also introduced the following innovations: trumpets were used independently instead of always doubling the horns, cellos became separated from the double basses, and woodwind instruments were often given the main melodic line. In the *Military Symphony* (*No. 100*) Haydn introduced some percussion instruments not normally used in the orchestras of this time, namely, triangle, hand cymbals, and bass drum, and, what is still more unusual, they are employed in the second movement, which in the Classical tradition is normally the slow movement.

Standardization of the orchestra

In Haydn's music a method of composition appeared that had a bearing on orchestration. This consisted of the conscious use of musical motives; motive is defined in the *Harvard Dictionary of Music* as: "The briefest intelligible and self-contained fragment of a musical theme or subject." Perhaps the best known musical motive in Western music is the four-note group with which Beethoven's *Fifth Symphony* begins. These musical cells became the musical building blocks of the Classical period, particularly in the middle or development section of a movement, with the composer moving the musical motive from instrument to instrument and section to section, giving a new facet to the orchestration. The art of orchestration was thus becoming a major factor in the artistic quality of the music.

Mozart's special contribution

Mozart, too, was responsible for great strides in the creative use of instruments. His last two symphonies (*Nos. 40*, K. 550, and *41*, K. 551) are among the most beautifully orchestrated works of this or any period. For his 17 piano concertos, Mozart exhaustively explored the combination of piano and orchestra.

Romantic. Beethoven began his career under the influence of the Classical composers, particularly Haydn, but during his lifetime he transformed this heritage into the foundation of a new musical practice that was to become known as Romanticism. The Classical composers for the most part attempted to orchestrate with a sense of grace and beauty. Beethoven occasionally made deliberate use of new, intense, often even harsh orchestral sounds. He also, in his later symphonies, augmented the orchestra with a piccolo, contrabassoon, and third and fourth horn. The *Ninth Symphony* has one passage calling for triangle, cymbals, and bass drum, a combination identified with the imitations of Turkish Janissary music in vogue in previous years.

The Romantic period was characterized by great strides in the art of instrumentation, and, in fact, the use of instrumental colour became one of the most salient features of this music. The piano really came into its own as a source of interesting sonorities; the orchestra expanded in size and scope; new instruments were added; and old instruments were improved and made more versatile. The Romantic period saw the appearance of the first textbook on the subject of orchestration. It was the French composer Hector Berlioz' *Traité d'instrumentation et d'orchestration modernes* (1844; *Treatise on Instrumentation and Orchestration*, 1856). Berlioz was one of the most individual orchestrators in the history of music, and his *Symphonie fantastique* (1830) is one of the most remarkable pieces of music to come out of this era. Berlioz made use of colour to depict or suggest events in his music, which was frequently programmatic in character. He called on large forces to express his musical ideas, an idea that persisted throughout the 19th century and into the 20th. Berlioz' *Grande Messe des morts* (*Requiem*, 1837) calls for four flutes, two oboes, two English horns, four clarinets, 12 French horns, eight bassoons, 25 first violins, 25 second violins, 20 violas, 20 violoncellos, 18 double basses, eight pairs of timpani, four tam-tams (a type of gong), bass drum, and ten pairs of cymbals; four brass choirs placed in various parts of the hall, each consisting of four trumpets, four trombones, two tubas, and four ophicleides (a large, now obsolete brass instrument); and a chorus of 80 sopranos, 80 altos, 60 tenors, and 70 basses.

First treatise on orchestration

The coloristic ideas in Berlioz' music were carried on in various ways by other important 19th-century composers and reached a culmination in the music of the German composer Richard Strauss and the Austrian Gustav Mahler—both of whom demanded a virtuoso orchestra—and were orchestrated in a complex fashion, although Mahler was capable of very delicate effects.

Post-Romantic and 20th century. Claude Debussy in France was probably the most important composer of the Impressionistic period, which lasted roughly from 1880 until the turn of the century. The Impressionist composers attempted to describe scenes and evoke moods by the use of rich harmonies and a wide palette of timbre. No composer has ever handled the colours of the orchestra with greater subtlety. Naturally, this is also dependent on his use of harmony, melody, and rhythm, but the dominant impression of a Debussy work is focussed on his use of orchestral instruments to create light and shadows. Works that exemplify his techniques are *Prélude à l'après-midi d'un faune* (*Prelude to the Afternoon of a Faun;* 1894), *Nocturnes* (1899), and *La Mer* (*The Sea;* 1905). In *Nocturnes* he uses a wordless women's chorus as a section of the orchestra, functioning as another source of timbre rather than as the transmitter of a text.

The Impressionist school

Many of the composers who followed Debussy and Mahler brought about radical changes in the use of the orchestra. A good example of some of these changes is in *The Rite of Spring* (1913), by the Russian-born composer Igor Stravinsky. The strings frequently do not assume a dominant role but, rather, often play music that is subservient to the brass or woodwind. Percussion instruments greatly increased in importance and have continued to do so. In 1931, Edgard Varèse composed an important work, *Ionisation*, for 13 percussion players, a landmark in the emergence of percussion instruments as equal partners in music.

The period between World War I and World War II was dominated by two main schools of composers with vastly differing results for orchestration. One was responsible for the Neoclassic style; the other, gathered around the Austrian composer Arnold Schoenberg, drew heavily on the Romantic movement for its direction. The Neoclassic composers sought to free music from the influence of Impressionism. Whereas the Impressionist and Romantic composers had frequently employed the instrumental forces at hand to create a deliberate sense of vagueness, the Neoclassic composers, beginning in about 1917 with a group of Frenchmen known as Les Six, attempted to recreate the clarity of the classic period by turning to models found in the popular music of the period, the music of the dance halls and cabarets. The Neoclassic composers also turned away somewhat from the orchestra as a medium, finding the forces of chamber music more suitable for their ideals. Neoclassic music returned

to a clearer concept of "sections" in orchestration. The music of a composer such as Paul Hindemith in Germany is closer to the music of Mozart in its sense of instrumentation than it is to Romanticism or Impressionism.

The music of Schoenberg and the Austrian Alban Berg drew heavily on the Romantic movement and eventually became known as Expressionism, which stressed inner experience. Emphasis on the inner man produced a music that was thick, dark, and intense.

Importance of electronic music

The first half of the 20th century saw the emergence of electronic music, although it did not become important until after 1950. The principal reasons for the inclusion here of electronic music, which is dealt with in the article ELECTRONIC MUSIC AND INSTRUMENTS, are that electronic sounds, either taped or live, frequently are included in a composition combined with traditional instruments, and it has had a decided influence on orchestration. By the 1960s many composers were writing works for electronic sounds and instruments. The electronic sounds provide a dimension to instrumentation never before possible. A number of things are noteworthy. Electronic sounds are capable of incredibly subtle changes of timbre, pitch, and mode of attack. When combined with traditional instruments they add a rich new spectrum of colour. This in turn has influenced the composer to attempt to produce "electronic" sounds with standard instruments. The result has been a great extension of the sound possibilities of Western instruments.

Another 20th-century trend was away from large orchestras and toward chamber ensembles, often of nontraditional combinations. Compositions for such ensembles may excel in economy of means and focussing on individual instrumental timbres. To achieve this, unusual playing techniques may be required.

NON-WESTERN INSTRUMENTATION

Much of music outside the West has entirely different aesthetic aims; the music of the Hindu world, best known to the West through the classical music of India, provides an example. Indian music always has had strong ties with mythology and religion and thus produced an art that is as different from Western music as Hinduism is from Christianity. It achieves unity through similarity rather than through change and is based on a more purely sensual approach that does not lend itself to detailed intellectual analysis as does much Western music. Hindu music is divided, for example, into *rāgas*, or melody types. The word *rāga* means colour or mood. Combined with the *rāgas* are *tālas*, or rhythmic patterns. The possible combinations of *tālas* and *rāgas* are many, producing a music that is wonderfully subtle.

Instruments of Indian and Balinese music

The instruments for this music consist of various drums made of terra-cotta, wood, or metal; cymbals also serve as percussion instruments. Probably the instrument best known to Western audiences is the *ṭablāh* a two-drum set capable of very subtle changes in sound. The two best known string instruments are the sitar (plucked) and the tambura, a four-stringed instrument that provides the omnipresent drone accompaniment. In addition, there are various wind instruments, such as the bamboo flute, the *shahnā'ī* (oboe), and a trumpet that frequently measures more than two yards long.

Balinese and Javanese music is centred on the gamelan orchestra, the instruments of which include the *saron* and *gender* metallophones (like xylophones but with metal, not wooden, keys), the *gambang kayu* xylophone, tuned gongs, flutes, and the *rebab*, a violin-like instrument with two strings. All the instruments follow the same nuclear melody but elaborate it in different ways. The heavy reliance on tuned percussion instruments has given this music a brilliant quality that Western audiences have found extremely attractive. The gamelan orchestra, for instance, influenced Debussy, who first heard the music at the Paris Exposition in 1889.

The approach to instrumentation in the music of India and Bali is quite different from that of Western music. The concept of contrast created through the various "choirs" of the Western orchestra is not a primary concern. In Indian music a sameness of colour is created through the use of the drone played on the tambura. This is not to say that this music is uncolourful, but that a specific timbre is established for an entire composition. Since the time of Debussy, Western composers have come increasingly into contact with, in particular, the music of India, Bali, and Japan. A comparison of Balinese gamelan music with the *Sonatas and Interludes* for prepared piano by the 20th-century American composer John Cage would show how profound this influence can be.

ARRANGEMENT AND TRANSCRIPTION

A practice that has been much employed in the 20th century, although by no means confined to it, has been the writing of transcriptions and arrangements. Though little distinction is made between the two terms, there are, at least in current practice, differences. A transcription is essentially the adaptation of a composition for an instrument or instruments other than those for which it was originally written. An arrangement is a similar procedure, although the arranger often feels free to take musical liberties with the original score. This is especially true of arrangements for jazz or rock groups and of popular compositions or songs from musical comedies.

In the 18th and 19th centuries, chamber and orchestral music was transcribed for the piano for the purpose of study and, of course, for the pleasure of playing at home the music that had been heard at a concert. This practice has continued to the present day. Piano versions of many 18th- and 19th-century orchestral works exist in two- and four-hand arrangements. Another common practice is to reduce the orchestral parts of concertos to a keyboard version to enable students to study and play these works without an orchestra.

Piano versions of orchestral works

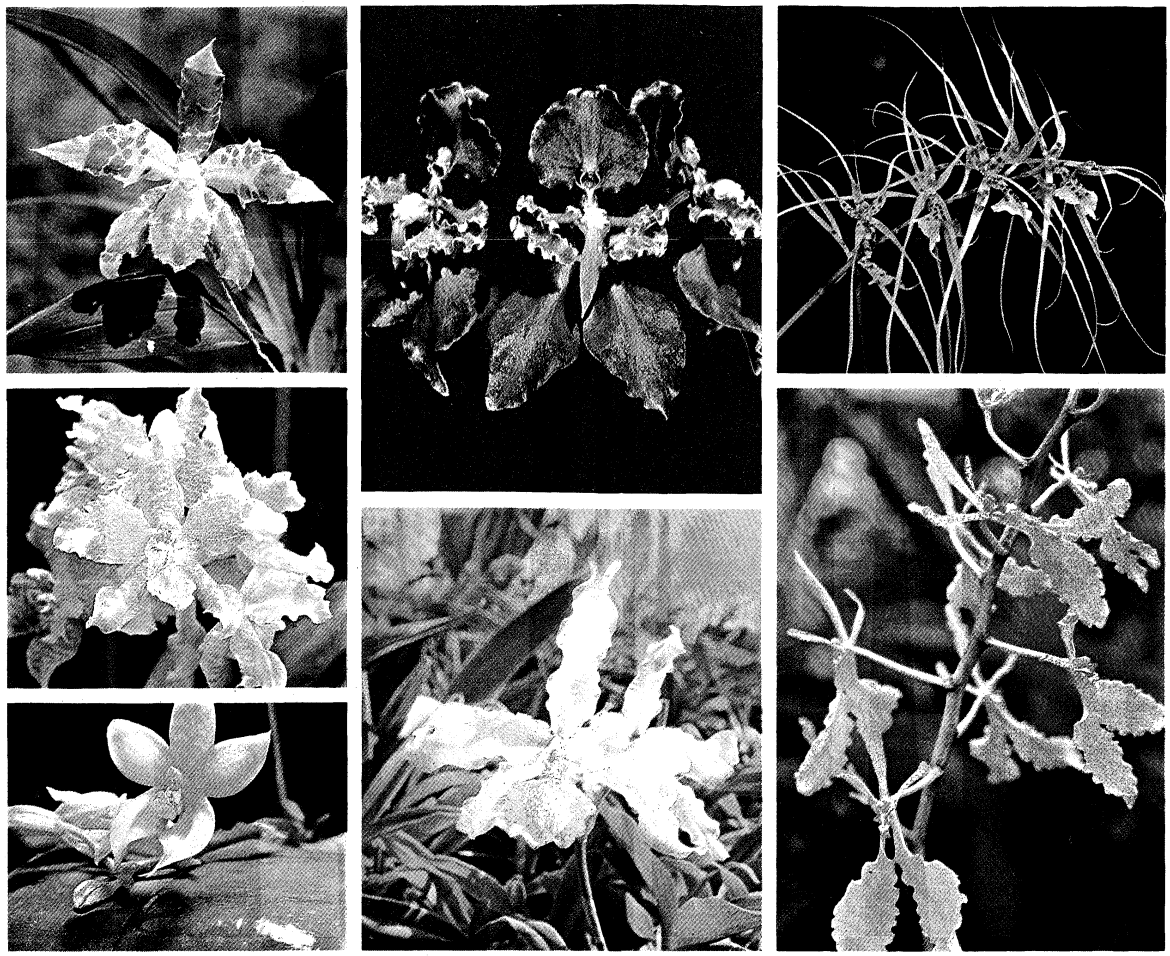
The symphonic band, despite its popularity in North America, was faced with a dearth of repertory written specifically for it. In the past, one answer was to transcribe orchestral works for band, substituting particularly the clarinets, with their wide pitch range, for the strings of the symphony orchestra. The necessity for that substitution is no longer so great because in recent times much more music has been written specifically for the symphonic band.

The jazz or rock arranger has done much more than simply transcribe the keyboard version of a song. All forms of popular music in the 20th century have been involved in the art of improvising. Musicians working in this field almost always embellish the music as they perform it. The jazz or rock arranger in a sense improvises on manuscript paper. In making an arrangement for a group of musicians the arranger will embellish both the harmonic structure and the melody of the composition; or the arrangement will be worked out in rehearsal and memorized or written down later. Usually, the arranger keeps enough of the original material to enable the listener to recognize the source. His skill depends on how well he can manipulate the materials of the original and on his originality in scoring the composition for the group at his disposal. The men who work in this field are frequently composers of popular music themselves.

The dance band predominant in the 1930s and 1940s is treated roughly in the following way by arrangers: the saxophones carry the melody more frequently than the other sections; the trumpets provide embellishment or figures that work around the melody; the trombones either are combined with the trumpets or serve as a melodic instrument; the piano and guitar provide harmonic filler and the double bass and set drums the rhythm.

BIBLIOGRAPHY. WILLI APEL, *Harvard Dictionary of Music*, 2nd ed. rev. (1969), a good source on any musical subject; DONALD JAY GROUT, *A History of Western Music* (1960), the best general history of music to date; ADAM CARSE, *The History of Orchestration* (1925, reprinted 1964), a detailed look at the evolution of the orchestra and musical instruments; NICOLAS RIMSKY-KORSAKOV, *Principles of Orchestration, with Musical Examples Drawn from His Own Works*, ed. by MAXIMILIAN STEINBERG, 1 vol. (1964; orig. pub. in Russian, 1910), still one of the best texts for the serious student; ROMAIN GOLDRON, *Ancient and Oriental Music* (1968), examples of non-Western music and instruments.

(D.Er.)



Diversity among orchids.
(Top left) *Odontoglossum triumphans.* (Top centre) *Oncidium falcipetalum.* (Top right) Spider orchid *(Brassia brachiata).* (Centre left) *Odontoglossum crispum.* (Bottom left) Moth orchid *(Phalaenopsis violacea).* (Bottom centre) *Dendrobium chrysotoxum.* (Bottom right) *Renanthera Imshootiana.*
(Top left, top centre) G.C.K. Dunsterville, (top right, centre left, bottom left) A to Z Botanical Collection—EB Inc., (bottom centre) Sven Samelius, (bottom right) Walter Chandoha

Orchidales

Orchidales is an order of monocotyledonous flowering plants (monocots comprise one of two great groups of flowering plants and are roughly equivalent to plants having parallel-veined leaves, grasses being a familiar example) considered by most authorities to contain only one family, the Orchidaceae, or orchid family. The small family Burmanniaceae has sometimes been included within the orchid order, however, and the subfamily Apostasioideae, included here, has sometimes been considered as a separate family (Apostasiaceae) in the order Hemodorales.

The orchid family is regarded by most botanists as the largest family of flowering plants, although it is not yet possible to give an accurate estimate of the size of the family. Educated guesses range from 15,000 to 35,000 species in 400 to 800 genera, but the lower estimates usually fail to consider the diversity of the family in the tropical regions.

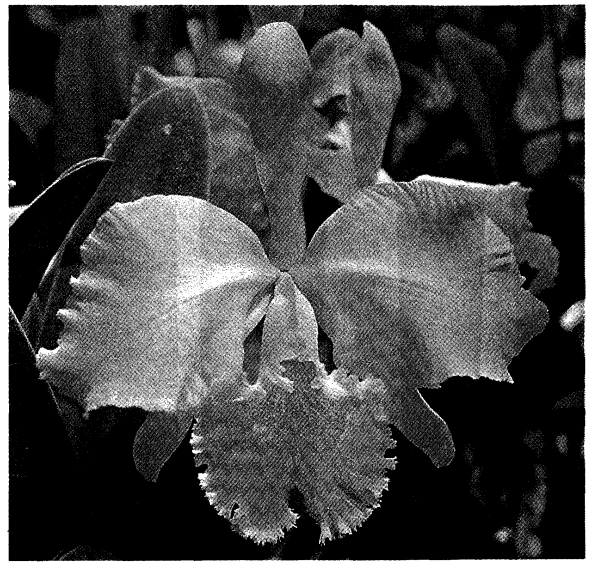
GENERAL FEATURES

Discussions about orchids, whether among professional botanists or amateur orchid enthusiasts, often leave the impression that orchids are "somehow different" from other plants. As a group, they are different from other plants but only in the morphological (structural) characters that tend to distinguish them. These characters are associated with the flower and its organization. Even the special characters of orchid flowers, such as the masses of pollen called pollinia, the joining of the stamens (male organs) and pistil (female organ) to form the column, and the tiny seeds without endosperm (starchy nutrient tissue) or organized embryos, are found individually in other groups of flowering plants. Therefore, no single character distinguishes the orchids; however, through the combination of several characters, a family of flowering plants, the Orchidaceae, emerges.

Diversity of structure. Orchids are primarily herbaceous (nonwoody), although some species may be vines, vinelike, or somewhat shrubby. The plants may be free-living or, in a few cases, saprophytic (*i.e.*, obtaining their food from the organic matter of dead and decaying plants), and they may be terrestrial or epiphytic (not rooted in soil but living among tree branches or on other aerial supports). They may even be parasitic on fungi. Orchid flowers vary tremendously in size from the minute flowers of some species of the genus *Pleurothallis*, which are no more than 2 millimetres (about 0.1 inch) in diameter, to the large ones of *Brassia*, which may be more than 15 inches (about 38 centimetres) from the tips of the lateral sepals (petallike structures) to the tip of the dorsal sepal. Although vegetatively orchid plants are somewhat uniform and do not show the wide range of diversity that may be found in some other families, the plants do offer a number of different growth types. Growth habits vary from those in which the plant is reduced to no more than roots (*Dendrophylax*) to saprophytic plants apparently lacking chlorophyll (*Corallorhiza*) to gigantic plants (*Arundina*) that superficially resemble a bamboo.

Growth forms

The diversity of structure among orchid flowers can be attributed mainly to the methods of pollination found in the family or to the fact that the family is adapted for the utilization of a number of different types of pollinators (animals that carry pollen from one flower to another).



Large, attractive orchids: wild and cultivated.
(Left) The yellow lady's slipper *(Cypripedium calceolus)* is one of the largest northern temperate orchids. (Right) *Cattleya,* a genus with thousands of hybrids, contains many popular ornamental and corsage orchids.
(Left) Sven Samelius, (right) Derek Fell

Orchids range from above the Arctic Circle to the tropics, but the majority of species are found in the latter.

Importance to man. The orchid family is probably one of the most important of plant families from a horticultural point of view. Nevertheless, of the 400 or more genera in the family, only about 50 genera are commonly cultivated. A plant group as large as the orchid family might be expected to contain enough different types of plants to satisfy any horticultural desire, but in the past 100 years over 60,000 orchid hybrids (crosses between different species) have been produced. A number of these hybrids are used in the cut-flower trade, and a large number of the plants are grown and exhibited with the flowers intact. The large number of corsage orchids and flowers for the cut-flower trade, however, derive from only about 150 species. Several species from distinct genera are grown as hedges in tropical areas, and a number of species are used as ornamentals in warmer regions.

Other than the horticultural uses to which orchids are put, the family is notably lacking in species from which products are derived. The only commercially important product derived from orchids is vanilla. Most vanilla is produced from one species, *Vanilla planifolia,* although two additional species are also cultivated commercially (*V. pompona* and *V. tahitensis*). In recent years the use of artificial vanilla has even reduced the use of natural vanilla.

The principal vanilla-growing areas are Madagascar, Mexico, French Oceania, Réunion, Dominica, Indonesia, French West Indies, Seychelles islands, and Puerto Rico. Vanilla is grown from sea level to about 2,000 feet (600 metres) in elevation. The plant is a climbing vine that is indigenous throughout the tropical regions of the Western Hemisphere.

Although the flowers of vanilla are rather attractive, they last only for a day or so and are not useful for the floral industry.

Folk uses of orchids

Primitive peoples throughout the world use various other orchids for a variety of folk medicines and cures. In the West Indies, the bulbs of *Bletia purpurea* are boiled, and the liquid is thought to cure poisoning from fish. In Malaya, women take a drink made from the boiled leaves of *Nervilia aragoana* to prevent sickness after childbirth. In Malacca, boils are treated with a poultice made from the entire plant of *Oberonia anceps.* In Chile, *Spiranthes diuretica* is known to be a strong diuretic. In certain parts of Ecuador, the mucilage from *Catasetum* is thought to be good for broken bones. In various parts of the world, certain orchids are also used for food or food supplements. In Malaya, the leaves of one species of *Anoectochilus* are sold as a vegetable, and the leaves of *Dendrobium salaccense* are cooked as a seasoning with rice. In certain parts of the Asian tropics, the tubers of some species of *Gastrodia* are eaten like potatoes. Throughout the world several species of orchids are used as a glue substitute. In most cases the glue is derived from the pseudobulbs. Salep is derived from the tubers of several species of *Orchis.* The tubers are boiled, then dried and powdered. The resulting preparation is often used as a flour substitute. It is also used as a substitute for gum arabic.

FORM AND FUNCTION

The flower. The primary characters that distinguish the orchids as a group are found in the flower. At the bottom of an unspecialized non-orchid flower is the stem that supports it, called the pedicel. Directly above, and at the base of the flower itself, is a whorl of green, leaflike organs called sepals. Above and inside the sepals is a second whorl of coloured petals. These are the asexual parts and are developed to protect the flower or to attract pollinators. Inside (also arranged in whorls) are the sexual portions of the flower. First are the male stamens, which consist of a filament (the slender stalk that supports the anther) and an anther (usually an arrangement of four sacs filled with loose pollen grains). There may be several whorls of stamens. In the centre of the flower is the female pistil, which is composed of an enlarged portion (the ovary) topped by a stalklike style with a stigma at its apex. The pollen in the anther is powdery and is usually carried from the anther of one flower to the stigma of another flower by a pollinating agent.

Structure of orchid flowers

The previous description is of a generalized flower, and it must be kept in mind that many types of variations may be found. Even though it is a generalized flower, it is not to be considered a primitive flower. The generalized orchid flower is also supported on a pedicel. The ovary, however, with its ovules seems to be an integral part of the pedicel because it is embedded within the upper portion of the pedicel below the attachment of the petals. The orchid flower is thus said to have an inferior ovary. The sepals and petals are usually similar, often highly coloured, and in sets of three. One petal is developed as a landing platform for the pollinator and is called the lip (or labellum).

Sexual organs. The sexual portions of the orchid flower are quite different from other generalized flowers and they tend to characterize the family. The filaments, anthers, style, and stigma are reduced in number and are usually fused into a single structure called the column. In

one relatively small group (the lady's slippers), two anthers are present, one on each side of the column (some members of the primitive subfamily Apostasioideae have three anthers). The majority of the orchids, however, retain only a single anther at the apex of the column.

In the orchid the ovary is composed of three structures called carpels. The carpels have fused so that the only outward evidence of their existence is the three ridges on the outside of the seed pods. The mature seed pod does not open along the lines of juncture but rather down the middle between the lines of juncture. The ovules, which are arranged along the ridges inside the ovary, do not develop until some time after the flower has been pollinated, thereby contributing to the long delay between pollination and the opening of a ripened pod.

Orchid petals. In most monocots the sepals and petals are so similar that they are not distinguished but fall under the collective term perianth. In the orchids, however, they are usually quite distinct and therefore retain their separate identification. The petal opposite the fertile stamen is the one called the lip, or labellum. Often two, or even all three, of the sepals are joined, and the lip, petals, or the sepals may be joined to the column for some distance. One of the characteristic differences between the orchid family and other advanced monocots is that the fertile stamen or stamens (that is, the one or more that are not completely reduced) are all on one side of the flower opposite the lip. This makes the flower bilaterally symmetrical.

The lip is oriented upward in the bud, but as it later develops, twisting takes place in the pedicel or ovary so that the lip is usually oriented downward by the time the flower opens, a process called resupination.

The presence of the labellum as a landing platform for insect pollinators and the reduction of the stamens and pistil of a flower to a single structure, the column, is certainly the apex of floral adaptation to insects as pollinating agents. Once achieved, this combination provides a foundation for all kinds of specializations for attraction of specific pollinators. The development of the strange and complex reorganizations found in the flowers of many orchid genera are directly attributable to evolutionary forces interplaying on the basis of such a pre-existing foundation.

Nectar-producing organs. There are several types of nectaries in the orchids, including extrafloral types that secrete nectar on the outside of the buds or inflorescence (flower cluster) while the flower is developing. Shallow cuplike nectaries at the base of the lip are common. Some nectaries are in long spurs produced either from the joined sepals or from the base of the lip. Members of the *Epidendrum* complex have long tubular nectaries embedded in the base of the flower alongside the ovary. Nectaries on the side lobes of the lips are known, and general nectar secretion along the central groove of the lip is common. The nectaries of many species of the Liliales order, a possible ancestral group, are within the septa of the ovary, but the nectaries of the Orchidales are present on the sepals or petals, if they are present at all.

In most of the orchids, the stamen and style are fused completely into a single organ—the column; however, in *Cypripedium, Apostasia,* and some higher orchids such as *Spiranthes,* they are only partially united. In some cases an extension of the base of the column with the lip attached at its apex forms the column foot. In many instances this structure exudes nectar.

Anthers and pollen. In most of the orchids the anther is a caplike structure located at the apex of the column. The anther of some of the more primitive orchids is superficially similar to that of a lily or amaryllis. In *Habenaria* and its allies the anther projects beyond the apex of the column but is thoroughly attached. Sometimes the two halves of the anther are widely separated and have led some botanists to believe that they represent two anthers. In one subfamily, Cypripedioideae, two anthers do occur and are attached on each side of the column. These orchids are usually considered to be primitive.

In most orchids the pollen grains are bound together by threads of a clear, sticky substance (viscin) in masses called pollinia. Two basic kinds of pollinia exist: one has soft, mealy packets bound together to a viscin core by viscin threads and is called sectile; the other kind ranges from soft, mealy pollinia, through more compact masses, to hard, waxlike pollinia. *Goodyera, Haberaria,* and *Spiranthes* have the former, and the latter are found in *Sobralia, Phajus, Cattleya,* and *Oncidium* (becoming harder in that sequence). The hard, waxy pollinia usually have a small amount of mealy pollen with viscin strands that attach the pollinia to each other or to a viscidium. This portion of the pollinium is termed the caudicle.

Structure of orchid pollinia

The stigma is usually a shallow depression on the innersides of the column. It is composed of the three stigmatic lobes found in the typical monocot flower, but the three lobes are thoroughly grown together. Faint lines often can be seen on the surface of the stigma, dividing it into three parts.

In the majority of the orchids, a portion of one of the three stigma lobes is specialized and forms a structure called the rostellum. The rostellum projects down in front of the anther in the form of a flap. As the visiting insect begins to back out of the flower, it brushes the rostellum, which is covered with sticky stigmatic liquid. The pollinia are then picked up from the anther and stick to the body of the insect. Some primitive species have no rostellum, and the pollinia simply stick to stigmatic liquid that is first smeared on the back of the insect. A further specialization occurs in more advanced orchids in which the caudicles of the pollinia are already attached to the rostellum and a portion of it comes off as a sticky pad called a viscidium. In the most advanced genera a strap of nonsticky tissue from the column connects the pollinia to the viscidium. This band of tissue is called the stipe and should not be confused with the caudicles, which are derived from the anther. Orchids that have a stipe also have caudicles that connect the pollinia to the apex of the stipe. The pollinia, stipe, and viscidium are called the pollinarium.

Seeds and their dependence on fungi. Orchid seeds are extremely small and contain an undifferentiated embryo that lacks endosperm. A single seed pod produces a large number of small seeds, which are ideally suited for dispersal by wind. Orchid seeds need the presence of a fungal mat in order to germinate and grow in nature. This mat is called a mycorrhiza. The fungus apparently penetrates the seed and contributes to the growth of the seedling by producing or supplying some of the necessary nutrients for growth. It has not yet been firmly established whether or not the fungus is necessary for the continued growth of the mature plant, but it appears likely that the presence of the fungus aids in the uptake of nutrients and prevents the leaching of nutrients from the root material of epiphytic species. Although it is possible to grow and germinate some orchids without the fungus in artificial cultures, it is thought that a fungus is necessary for germination and growth in nature.

Mycorrhizal fungi and seed germination

In some cases the presence of a specific species of fungus is necessary, while in other cases several fungi may have the ability to become involved in the process. The fungi that are involved with mycorrhiza in orchids are of two taxonomic groups: the Fungi Imperfecti (Deuteromycetes) and the Basidiomycetes (fungus order Agaricales), the latter group being that to which the familiar mushrooms belong. Although most, if not all, orchids have mycorrhiza, mycorrhiza are not limited to the orchid family but are found in a large number of other plant families.

Growth habits. The majority of the tropical orchid species are epiphytes; that is, they grow on trees without connection with the ground. Nearly all of the orchids in the temperate zones, however, are terrestrial.

The predominant growth form in a wide range of monocots has been shown to be a type known as the sympodium, and this may reasonably be considered to be the primitive condition in the orchids. Sympodial growth is a creeping habit and it consists of an axis that appears continuous but is, rather, made up of a succession of elements that each originate not from a terminal bud but

Extreme variations employing habitat.
(Left) *Neottia nidus-avis*, bird's nest orchid, is a nongreen saprophytic orchid. (Right) *Taeniophyllum zollingeri* is epiphytic, with flattened green roots that act as organs of attachment and in place of leaves for photosynthesis.
(Left) C. Foord—EB Inc., (right) W.H. Hodge

as a fork of a dichotomy, the other fork being weaker in growth or suppressed entirely. The usual form of a sympodium is a horizontal rootlike stem structure called a rhizome that turns up at the ends of each "branch." Most primitive orchids have a rather ordinary monocot appearance with a short rhizome stem and erect, nonthickened annual stems having scattered, spirally arranged leaves and a terminal inflorescence (flower cluster).

Another major type of growth form found in the orchid order is the monopodial habit, in which the stem has unlimited apical (terminal) growth and roots are not restricted to its basal portion. In some systems of classification this has been considered to be a distinguishing mark of the subtribe Sarcanthinae; however, the monopodial habit occurs in the subtribes Vanillinae, Cymbidiinae, Maxillariinae, Pachyphyllinae, Thelasiinae, and Oncidiinae, and possibly in other groups as well. In some of these groups there exists a graded series from sympodial plants to related monopodial types.

A great many orchids, especially the epiphytic groups, show variously thickened stems, or "pseudobulbs." While these structures are quite diverse in form, they fall into a limited number of morphological types and seem to show some evolutionary trends. One of these seeming trends is from pseudobulbs (or corms, bulblike stem structures) of several or many internodes (nodes are stem regions at which leaves attach, internodes are the stem areas between such nodes) to pseudobulbs of a single internode (as in *Bulbophyllum* and the subtribes Maxillariinae and Oncidiinae). Thickened stem bases may be found in either terrestrial or epiphytic groups, but pseudobulbs of a single internode are restricted to primarily epiphytic groups.

Saprophytic orchids

Saprophytic orchids, those that obtain their food from dead organic matter instead of by photosynthesis, are found in at least 12 subtribes of orchids. The majority of orchids pass through a saprophytic seedling stage, which may last for months, especially in terrestrial species. Thus, the evolution of a completely saprophytic life cycle in different groups of orchids is not surprising. The wholly saprophytic orchids pose special classification problems. The adaptations for saprophytism drastically change the vegetative features of the plant, thus obscuring some of the characteristics normally used for determining relationships. Even the reproductive features may be affected by these adaptations. In the genus *Corallorhiza*, for example, the pollen bearing structures are more simple than in the related genera *Oreorchis* and *Tipularia*. In the absence of these two closely allied genera, *Corallorhiza's* relationship to the Cyrtopodiinae subtribe would be much less clear. The saprophytes are difficult to cultivate and are poorly represented by herbarium specimens, which further complicates their study. Autogamy (self-fertilization) is frequent, and an autogamous saprophyte is nearly the ultimate in taxonomic difficulty.

NATURAL HISTORY

Life. The life cycle of an orchid is not essentially different from any other flowering plant. When the pollinator leaves the pollinia on the stigma (the pollen receiving structure of the female flower parts), the pollen tubes germinate and grow down the centre of the column to reach the ovary. This often causes the sides of the stigma to swell around the stigma and the enclosed pollinia. When the pollen tubes reach the ovules (this often takes six weeks or more, during which time the ovules have been developing), one or more enters and the sperm unites with the egg. The fertilized egg, or zygote, begins to divide to form the mass of cells destined to become the embryo. Most other plants have a slightly different development from that of the orchid. There is usually a double fertilization, one of which forms the zygote and the other an endosperm that nourishes the developing plant upon germination until the plant is able to manufacture its own food.

Orchid seeds

The orchid seed, however, has no endosperm. The seed consists of a simple, dry, outer coat enclosing a small mass of undifferentiated cells that form a pro-embryo. This extremely small and light unit can easily be carried in air currents and may travel long distances before coming to rest. The number of seeds in a single orchid capsule is nearly astronomical: close to 2,000,000 seeds have been reported from a single capsule of *Catasetum*. These extreme numbers undoubtedly are correlated with the unlikelihood of falling in a favourable spot.

Ecology. Orchids have a rather wide ecological distribution when the entire family is considered. Although the family is primarily tropical, a number of species are found in the northern and southern temperate zones. At least four or more species have been reported from north of the Arctic Circle. A number of species of the North Temperate Zone are found in bog situations, as well as on the prairies, grasslands, and in the hardwood forests.

Several species of *Spiranthes*, *Habenaria*, and other orchids are found in roadside ditches, often in wet, boggy situations. In some areas of the United States, *Habenaria ciliaris* might almost be considered a weed. The introduced Asian species *Zeuxine strateumatica* in southern Florida is now widespread and may be considered a weed.

Orchids grow from sea level to at least 15,000 feet (4,600 metres) in elevation. The greatest number of orchid species are found in cloud forest associations in tropical

regions. These are usually on forested mountainsides where the clouds brush the mountain day and night. Such forests are literally covered with mosses and lichens, and the inclination of the ground permits sunlight to penetrate through the vegetation to the ground. This is a perfect habitat for epiphytic orchids as well as aroids (family Araceae), ferns, gesneriads (family Gesneriaceae), and numerous other epiphytic groups of plants.

Contrary to popular belief, rain forests on generally flat terrain are poor localities for orchids. At Iquitos, on the Amazon River in Peru, only about 125 species of orchids have been reported, and many of these are quite rare. Orchids that do occur in rain forests tend to be in the tops of large trees with little variation in each tree. Often great quantities of a single species inhabit a tree, while another nearby tree may have only one or two additional species. The tropical-deciduous seasonal hardwood forests of the tropics, where marked wet and dry periods occur, often have numerous orchid species; however, this also tends to be the best farmland and therefore does not last long after colonization by man.

A wide range of ecological tolerances are shown by orchids. In addition to those species that are found above the Arctic Circle, a few species of orchids are found in desert conditions; for example, in the dry areas of northern Peru, several species of orchids are found epiphytic on cacti. On the Peninsula de Santa Elena in western Ecuador, two species of *Oncidium* as well as one species of *Brassia* are found on cacti. In parts of Central America one species of *Brassavola* is found growing on mangrove roots, often at, or only slightly above, the level of high tide. In Jamaica several species are found growing on bare rocks. In the Everglades, on the other hand, one species of *Habenaria* is almost aquatic. In western Mexico, one species of *Pleurothallis* is epiphytic on lichens.

Desert-dwelling orchids

Orchids vary from those species that are very widespread, such as the species found throughout most of the tropical regions of the Western Hemisphere (*e.g., Ionopsis utricularioides*), to some species that seem to be restricted to a single mountain (some species in the subtribe Pleurothallidinae). In the West Indies, each of the major islands seems to have a fair number of such restricted species.

Pollination. The total orchid flower is an instrument adapted for exact modes of pollination. The function of pollen deposition is centralized precisely in the median plane opposite the lip, or labellum (the large central petal of the flower). Being directed back toward the centre of the flower, the anther deposits pollen on the most advantageous side of the visitor, considering efficiency in reception and deposition. This precision is also expressed by a tendency of orchids to deposit the pollen as one mass, the pollinarium.

The published data are insufficient to establish accurately the relative importance of pollinating agents in the family; however, a rough estimate may be given as in the Table. Such extreme divergence in pollinators testifies for youth, continuing evolution, and adaptive radiation in the family.

Pollinator Spectrum of the Orchid Family* (percentage)

Hymenoptera		other agents	
Wasps	5	Moths	8
Lower bees	16	Butterflies	3
Carpenter bees	11	Birds	3
Euglossine bees	10	Flies	15
Social bees	8	Mixed agents	8
Mixed bees	10	Self-pollination	3

*The numbers represent an estimate of the relative importance of various pollinating agents in terms of the percentage of orchid species pollinated by each group.

Orchids as a group have nectar as the major attractant, whereas pollen, sought by pollinators among lower plant families as a protein-rich food, has been withdrawn. This is tied to the exactness of the mechanics of pollination, but it also means that orchids can no longer provide the only source of sustenance for the pollinator, and other flowers must be present in the biosphere to maintain visitors.

Nectar in orchids is provided in tubular nectaries (*Brassavola, Angraecum, Comparettia,* and other genera), in grooves on the labellum (*Listera, Epipactis,* etc.), and at the base of the column and lip (*Dendrobium, Scaphyglottis,* etc.). It has been estimated, however, that as many as 8,000 species of orchids, one-third of the family, are nectarless. These orchids have developed other means of attracting pollinators, largely consisting of deceptive attractants in one form or another.

Many orchids produce pseudopollen for the attraction of pollinators. A powdery mass resembling pollen occurs on the labellum of a number of species of *Maxillaria* and *Polystachya.* Sometimes the grains are detached outgrowths called papillae, sometimes disintegrated multicellular hairs filled with starch. It has been demonstrated that this pseudopollen is collected by bees at least from the orchid genus *Maxillaria.* In none of the species producing such pseudopollen is nectar produced.

Pseudopollen

Bee pollination. Flowers pollinated by bees tend to produce agreeable odours, have bright colours, to be open during the day, to provide a landing platform, to have nectar guides in the form of coloured lines running into the depths of the flower, and to have concealed nectaries. The basal portions of the orchid lip are usually formed into a tunnel with the column forming its upper side. The bee enters the tunnel to get at the nectary, and in backing out, some of the stigmatic fluid may be rubbed on the back of the bee. As the bee backs farther, the pollinia become attached to the sticky material and are carried with the bee to the next flower. In more advanced orchids the pollinia may be attached to a sticky pad, the viscidium, which becomes detached from part of the stigma and sticks to the pollinator.

Some species of orchids are pollinated by bees that are attracted by deceit. The flowers of the large genus *Oncidium,* for example, are pollinated by male *Centris* bees in what appears to be a case of pseudo-antagonism. The flower appears to simulate an enemy insect, which the male bee tries to drive away from his territory. In the process of striking at the flower, the pollinia are attached to the head of the bee.

The most exciting and unusual examples of deceit, traps, and manipulation of pollinators are to be found in those orchids that are pollinated by male euglossine bees (species of the bee tribe Euglossini). The syndrome of flowers pollinated by male euglossini is based on the attraction of the male bees to the odour of the flower. In no case does the male euglossine bee receive food from the orchid that it visits for the purpose of collecting odour. So far as has been determined, all members of the orchid subtribes Stanhopeinae, Catasetinae, Lycastinae, and Zygopetalinae are pollinated by male euglossine bees, as also are certain genera of the subtribes Oncidiinae and Ornithocephalinae. The euglossine male bees visit other nectar-producing flowers for their food, but when the male bees visit the non-nectar-producing orchid flowers, they rub the surface of the lip with their front feet and collect in special tarsal brushes the odour that is produced there. The bees then launch into the air and transfer the odour to their hind tibiae (leg segments), which are noticeably swollen. It is in the process of transferring the odour to the hind tibiae that the bee is manipulated by the orchid.

Odour-collecting bees

The pollination phenomena of the Stanhopeinae subtribe are based on the same principles as in the Catasetinae, but with the separation of sexes in the flower being in time rather than in space. Self-pollination in the Stanhopeinae subtribe is averted by differences in the maturing time of the male and female reproductive parts, whereas in the Catasetinae there are separate flowers of each sex. The stigma is narrow when the bee first visits, and the pollinia are thick before drying and therefore cannot be inserted into the stigma of the same flower. Powerful fragrances for attraction are present. In the Catasetinae, the male flowers are organized to position the bee perfect-

Manipulation of bees by orchid flowers

ly so that when the pollinarium is discharged the viscidium becomes attached at a specific location on the bee. In this manner it will fit properly in the stigmatic cleft of the female flower when the bee visits it. In most species of the Stanhopeinae, the flowers are organized to take advantage of the erratic flight habits and slowed reflexes of the male bees that are collecting odour components. Their pollination mechanisms, in the more spectacular forms, manipulate the bees so as to make them fall either through a chute, down a slide, or into a "bucket" of water.

In the genus *Stanhopea*, for example, some species face downward with the lip and column pendent while the sepals and petals are reflexed upward. The base of the lip is saccate (baglike) and is separated from the arched column. The bee enters here rather than at the apex of the lip and column, which are close together, forming a chute. The rostellum (beaklike apex of the female flower parts), with the attached viscidium, projects into the chute. After the bee has landed and brushed at the base of the lip in order to collect the odour, it attempts to fly out of the flower. The column interferes with its flying equilibrium, and it falls, abdomen first, down through the chute, picking up the viscidium under its body (metathorax) as it passes.

Calaway H. Dodson

Manipulation of pollinators.
(Left) *Coryanthes albo-purpurea* has a water bucket into which a visiting pollinator bee falls. The pollinarium is deposited on the abdomen of the bee as it exits by way of a special tunnel, and pollination occurs when the bee visits another plant. (Right) After rubbing for odour at the base of the lip, a visiting bee *(Eulaema meriana)* falls through the specialized chute of *Stanhopea gibbosa* and either picks up or delivers the pollinarium.

Another variation of pollinator manipulation by orchids is found in the genus *Coryanthes*. The flowers are very large and might even be considered grotesque. The sepals and petals fold back out of the way when the flower opens —like sails on a boat—revealing the strangely formed lip. The lip is divided into three parts: a globular- or hood-shaped portion called the hypochile above; an elongate, sometimes fluted part, the mesochile; and a bucket-shaped epichile. The epichile is partially filled with water during the last few hours before the flower opens and for a short time afterward by two faucet-like organs located at the base of the column, which drip water. Male euglossine bees are attracted by the strong odour produced by the hypochile, where they scratch. In trying to launch into the air to transfer the odour to their hind legs, the bees occasionally fall into the water-filled bucket. The sides of the bucket are vertical and are very waxy, so that the bee is not able to climb out of the bucket. The only way out is through a small tunnel formed by the apex of the column and the apex of the epichile of the lip. As the bee forces its way out of the tunnel, the pollinarium is deposited on its thorax. The pollinia may then be deposited in the stigma of another flower on a next visit, provided the original pollinia of that flower have already been removed and the stigmatic cleft has opened sufficiently to become receptive.

Moth and butterfly pollination. Moths normally fly at night and are attracted to flowers that produce strong odours and are white or light coloured. Moths that visit flowers are usually strong fliers and are capable of hovering in front of the flower while extracting the nectar. The typical moth-pollinated flower has a long, slender nectar tube containing abundant nectar. The fragrance produced is typically sweet or musky and the flowers usually are horizontal or hanging. Butterflies, on the other hand, are day fliers and go to highly coloured flowers that may or may not be fragrant. Butterflies tend to be somewhat erratic fliers and, lacking the ability to hover, usually land on the flower. The flowers are, therefore, usually erect and provide platforms for landing. Often the platform simply consists of a head of erect, densely packed flowers. Butterflies detect colours well, and butterfly-pollinated flowers are usually brightly coloured with red, orange, or yellow predominating. Nectar is commonly abundant and is hidden in deep nectar tubes.

Characteristics of butterfly-pollinated orchids

In an often cited case, Charles Darwin, the English evolutionist, predicted that a moth with a 10-inch- (26-centimetre-) long proboscis would eventually be found on the island of Madagascar as the pollinator of the orchid *Angraecum sesquipedale*, since a moth would need a long tongue to reach the nectar hidden in the very long nectary of this species. Such a moth has been found but has not yet been observed to pollinate the flowers of this amazing orchid.

The nectar tube of *Angraecum sesquipedale* is formed by the base of the lip. In most of the moth-pollinated, butterfly-pollinated, as well as bird-pollinated orchids, the nectar tube is arranged to guide the tongue or beak in such a manner that the pollinia are correctly attached to the pollinating organism.

Bird pollination. Flowers adapted to pollination by birds are usually brightly coloured, with reds, blues, and yellows predominating. They are usually tubular in form, often with a long nectary, and nectar is almost always present. Birds have little or no sense of smell, and bird-pollinated flowers tend to lack odour; however, the bright colours serve to attract the birds. Bird-pollinated orchids tend to follow the pattern of other bird flowers; however, in some cases they diverge considerably. Many orchids of the Western Hemisphere appear to have adapted to bird pollination as an extension of butterfly pollination, and, as in the case of *Epidendrum secundum*, birds and butterflies act as copollinators. In such cases, orchid flowers already adapted to butterflies are not greatly changed morphologically. On the other hand, orchids that have adapted directly to hummingbirds from bee-pollinated ancestors have changed fundamentally. The genera *Cochlioda*, *Sophronitis*, *Elleanthus*, *Isochilus*, *Comparettia*, *Hexisea*, and *Meiracyllium* are all bird pollinated and are remarkably similar in certain aspects. All have bright colours, tubular form, and a callus or hump in the interior of the tube, on the lip, which acts to force the beak of the bird against the column.

Fly pollination. Some flies are important pollinators of flowers and certain families of flies (*e.g.*, the Syrphidae and Bombyliidae) are restricted to flowers for their food. Unspecialized flowers may attract flies to nectar, which is present in open, shallow nectaries and may emit sweet odours. The flies eat the nectar and do not store it as do bees. More specialized fly flowers may attract flies through deception, imitating decaying substances, dung, or carrion. Many kinds of flies are then attracted and act as effective pollinators. Fly-pollinated flowers have often developed traps for catching and holding unadapted visitors. They commonly have large landing surfaces and "tails" produced from the flower parts, which function as guides. Their colours are usually checkered or blotched and tend toward dull green, brown, purple, or red. The odours produced are commonly putrescent. Orchids pollinated by flies are common throughout the world. In most fly-pollinated orchids, special adaptations have developed—superimposed on the basic pattern of the bee-pollinated orchid flower—to guide the somewhat poorly oriented flies. Certain of the five petals may be long and taillike (as in *Bulbophyllum* and *Masdevallia*) or joined to form a flat radial flower (as in *Stelis*). The flowers themselves may be arranged to form a larger radial, compound "flower" (as in *Cirrhopetalum*). The petals or lip may be fringed with motile clublike hairs that vibrate in the wind and attract the flies (*Bulbophyllum* and *Cirrhopetalum*). Often the sepals are joined or the lip is saccate to form a trap (*Pragmopediium* and *Pterostylis*) into which the flies fall and from which they must crawl by

Adaptations to fly pollination

way of a tunnel that passes the stigma and anther. A common contrivance by which orchids exploit flies is a hinged, balanced lip that tips with the weight of the fly and launches the pollinator into the flower.

In orchids it is often difficult to decide when simple fly pollination ends and deception based on carrion mimicry begins. Many genera have some species whose pollination is based simply on attracting various flies on the basis of sweet odours and nectar production, while others attract flies on the basis of rotten odours but provide no food.

Three major groups of orchids have become predominantly fly pollinated: the subtribe Pleurothallidinae in tropical America, containing about 1,000 species pollinated by flies; the *Bulbophyllum* group of about 1,000 species found mainly in the Old World; and the large genus *Pterostylis* and its relatives (in Australia).

Several reports of fly pollination in species from genera otherwise known for bee- or moth- and butterfly-pollinated flowers have been made. *Epidendrum fimbriatum* is pollinated by flies of the family Tachinidae. In the northern United States, several species of *Habenaria* have been shown to be pollinated by mosquitoes.

Self-pollination in orchids. Self-pollination occurs in a significant number of orchids. Several degrees of this phenomenon may be found in a single genus, from species in which accidental self-pollination results in fertilization to those in which the flowers never open, yet produce fertile seed. In many orchids, self-fertilization is not possible due to genetically controlled self-incompatibility, in which pollen from a plant having a particular combination of genetic factors will not fertilize its own ovules or those of any other plant having the same combination.

In most species the pollen is kept separate from the stigma by a structure called the rostellar flap. This physical barrier is normally quite effective; but in some species, forms occur in which the rostellum degenerates or becomes stigmatic, and self-pollination results when the pollen germinates on the stigmatic fluid. In most of these forms, normal plants are also found in the same population. Self-pollination may also occur as a result of simple falling of old pollinia, a means of averting sterility at the end of a long normal period when the flower is open but no pollinator arrives.

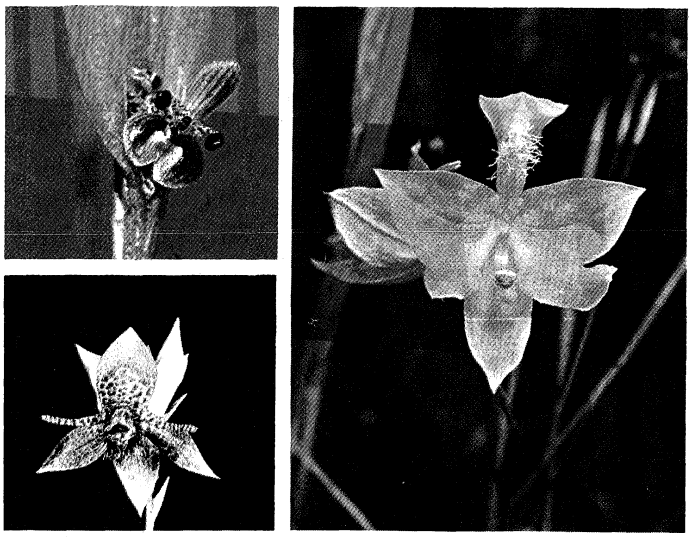
A kind of mechanical self-pollination occurs in some orchids in which the tissues connecting the viscidium and the pollinia bend downward and carry the pollinia into place on the stigma. Generally speaking, self-fertilization in orchids seems to be a means of averting extinction in plants growing under conditions adverse for normal pollination relationships. Examples include some species of *Orchis* in Europe and several orchids in Florida, such as *Epidendrum nocturnum, Encyclia cochleata,* and *Bletia purpurea,* all of which have cross-pollinated forms in other areas. Seeds blow into Florida from the Caribbean area where normal pollinators exist, but those plants that survive tend to be self-pollinated because they lack their customary pollinators, which are not found in Florida.

Self-pollination for survival

Probably more than 200 species of orchids have been reported to be more or less regularly self-pollinated, and undoubtedly many more exist.

Mimicry and deception—pseudocopulation. Flowers of the genus *Ophrys* deceive and manipulate pollinators mainly through odours, imitating those produced by the abdominal glands of female bees or wasps. Flower shapes, colours (including ultraviolet reflection), and tactile stimuli by the hairs on the lip operate on the sensory organs of the visiting males leading ultimately to the same behaviour as that observed during the initial phases of copulation with female bees. No ejection of sperm occurs, however, but the supernormal olfactory stimulation ensures that the male will remain for a long stay on the flower.

This act of pseudocopulation takes place in such a way that the pollinia are carried off and redeposited on a different plant. Four genera of solitary bees and wasps appear to be the principal pollinators. The species of *Ophrys* that are pollinated by the wasps *Trielis, Gorytes,* and the bee *Eucera* induce the insects to attempt copulation with the apex of the lip. Those pollinated by *Andrena* appear, for the most part, to stimulate the bee to reverse its position and copulate with the base of the lip. In the former group the pollinarium is affixed to the head of the pollinator, while in the latter it is attached to the abdomen. Only the introductory behaviour is necessary for pollination of the flower, and the bees do not encounter structures that lead to ejection of sperm. The behaviour is elicited by tactile stimulation from the hairs on the labellum, but the male "suitor" requires simultaneous and continued olfactory stimulation. The glistening pseudonectaries apparently imitate the eyes of the female bee. Metallic-blue mirrorspots similar to those found in the females enhance the effect. Dimensions of the flowers in the various species of *Ophrys* help in determining specificity and success.



Deception of pollinators.
(Top) *Pleurothallis raymondi* and (bottom) *Trichoceros antennifera* resemble female insects and attract the corresponding males, which carry out pollination while attempting to copulate with the flowers. (Right) *Calopogon pulchellus* has false stamens that act as an attraction to pollen-eating insects.
(Top) G.C.K. Dunsterville, (right) Walter Dawn, (bottom) Robert L. Dressler

Australian orchids of the genus *Cryptostylis* are pollinated by ichneumon wasps of the genus *Lissopimpla.* The wasp, after backing into the stigma, attempts to copulate with the flower by bending its body into an arch, with the base of the lip of the flower held by the claspers of the wasp. The upper side of the apex of the abdomen comes in contact with the viscidium, and the pollinarium becomes cemented in place. The wasp, after a short pause, then flies to another flower and the same behaviour delivers the pollinia to the stigma.

The South American orchid *Trichoceros antennifera* has flowers that simulate the female flies of the genus *Paragymnomma* to a remarkable degree. The column and base of the lip are narrow, barred with yellow and red brown, and they extend laterally to simulate the extended wings of a sitting fly. The base of the lip has no particular similarity to the head and thorax of a fly, but this is probably not necessary to complete the illusion. The stigma of the flower is located more or less at the apex of the "false abdomen" of the flower and reflects sunlight, as does the genital orifice of the female fly. The viscidium, extended over the stigma on the slender rostellum, projects up through the bristles and becomes attached to the basal portion of the abdomen of the fly. The viscidium is flat and padlike in this genus. The male flies, deceived by the similarity to the female fly and stimulated by the signal from the genital-orifice-like stigma, strike the flower for only a moment and then pass on to other flowers in the same area. The action is sufficient, however, to pick up the pollinarium. The long, slender stipe of the pollinarium bends down slightly and is forced into the stigma when the fly visits a succeeding flower.

Orchid flowers resembling female flies

Insects attracted for protection. A number of species of orchids have developed nectaries located in places

other than the flower. These seem to act to attract ants, which in turn scare away bees that would otherwise rob the flower of its nectar by cutting into the back of the flower. Such extrafloral nectaries also attract wasps, which function in scaring away insects that might rob the nectar from the flower. These ants and wasps are also thought to keep away grasshoppers, crickets, caterpillars, and other insects that eat the flowers. The symbiotic relationship (*i.e.*, two organisms living together with mutual benefit) between ants and the orchids that grow only in ant nests (*e.g.*, *Coryanthes* and some species of *Gongora, Epidendrum*, and *Schomburgkia*) may depend on this defensive mechanism. The plants and flowers of these orchids are extremely susceptible to damage by chewing insects if the ants are destroyed. The ants involved in this relationship do not pollinate the orchids; indeed, no ants are known to pollinate orchids.

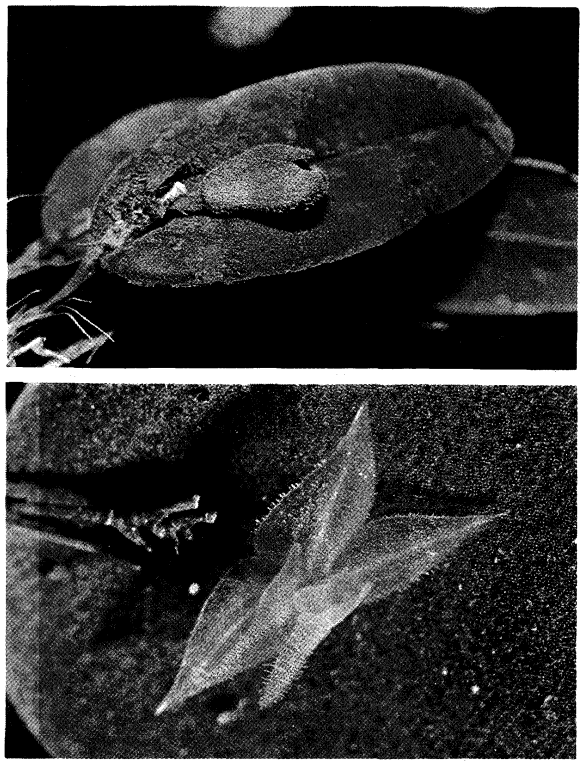
EVOLUTION

There are two extremes in the variation patterns shown by living plant groups, apparently depending on rates of evolution and amount of extinction. At one end of the spectrum stand such families as the Magnoliaceae, Annonaceae (both of order Magnoliales), and Nymphaceae (order Nymphaeales). These families have the appearance of being "old" groups, in which evolution is proceeding at a leisurely pace and extinction has greatly affected the pattern of variation. The delimitation of genera and tribes within the Magnoliaceae or the Nymphaeaceae is not difficult; however, students of evolution are at a loss in trying to understand relationships or phylogeny (evolutionary descent) in these groups.

Rapid evolution in orchids

The orchids are near the other end of the spectrum. They show little evidence of great age, show signs of relatively rapid evolutionary diversification in geologically recent time, and give much less evidence of extinction. Genera are often difficult to define, and higher categories within the family often seem even more difficult. These hazy boundaries between tribes and subtribes, however, may give clues to the patterns of phylogeny within the family. This is not to say or imply that living groups are derived from other living groups, but excellent evolutionary series can be found for nearly every morphological feature within the orchids. In recent years there has been a healthy skepticism concerning phylogenetic schemes. Even when a good morphological series can be found in some feature, it is often difficult to decide in which direction or directions evolution has occurred. This problem is not so serious in a highly derived group such as the orchid order. In comparing the genera *Cephalanthera* and *Oncidium*, there can be little doubt as to which is derived and which is primitive. In nearly every feature in which these two differ, it is *Cephalanthera* that is the ordinary monocot, easily comparable with other monocot families, while *Oncidium* is so advanced that it can scarcely be understood without comparing it with the less specialized members of the family.

In dealing with the primitive orchids and their evolution from groups now extinct, the bewildering parallelisms to be found in the more advanced groups must be borne in mind. It is probable that the subfamilies Neottioideae, Orchidoideae, and Epidendroideae evolved from a series of related species or genera having partial union of filaments and style, androecial zygomorphy (bilateral symmetry in the male parts), and a close association of the stigma and the median anther—all features that would predispose the groups to parallel patterns of evolution. Some of the ancestral populations may have been quite similar to the subfamily Apostasioideae, while others were similar to the genus *Selenipedium*, and yet others quite unlike either. The orchid family is not "unnatural" or polyphyletic (derived from more than one ancestral group) in the strict sense, since the ancestral group was, itself, a natural and closely interrelated group, even though it may have differed from the modern orchids in a number of features. It is fairly certain that the stipe, divided pollinia, and the viscidium have evolved independently in different groups of orchids. It is quite possible that the rostellum has evolved independently in two or more separate lines and that the reduction to a single median anther is similarly polyphyletic in the unusually "natural" family Orchidaceae.



Extremely small orchids.
The flowers of (top) *Pleurothallis pertusa* and (bottom) *Lepanthes* are only a few millimetres long.
Norris H. Williams

None of the living tribes of orchids could readily be derived from another living tribe, but their derivation from similar or common ancestors is easily visualized. The subfamily Cypripedioideae (lady's slipper orchids) is not very closely related to any other group, its relationship to the subtribe Limodorinae of the Neottioideae subfamily being perhaps quite as marked as its few resemblances to the subfamily Apostasioideae. The lady's slippers clearly diverged early from the main lines of orchid evolution. The Apostasioideae subfamily more nearly approaches the hypothetical ancestral type, the median anther being functional in *Neuwiedia*. It, too, represents a small relic group of somewhat isolated phyletic position, though perhaps closer to the other orchids than to the Cypripedioideae. The Orchidoideae subfamily presumably is derived from *Cephalanthera*-like types, but it would stand quite isolated if it were not for the relic subtribe Epipogiinae, which shows some relationships to both the Neottioideae and the Epidendroideae subfamilies. The Epidendroideae subfamily might be derived from somewhat Neottioideae-like ancestors, but it shows more primitive seed structure in the Vanillinae subtribe than any living species of Neottioideae, and the two subfamilies have apparently diverged at an early level in orchid evolution. The subtribes of the Neottioideae subfamily that are grouped with the Limodorinae subtribe are among the more generalized and primitive of the living orchids and would constitute a typically relic group if it were not for the great evolution of this group in Australia, where several striking specializations occur.

The subfamily Epidendroideae has the most examples of adaptive radiation to various types of pollinators. Apparently evolution in this subfamily has been accentuated by adaptive radiation to different types of pollinators. In this subfamily are found all types of pollinator classes with the exception of bat, rain, and wind pollination.

CLASSIFICATION

Distinguishing taxonomic features. The Orchidales order may be separated from the Liliales, a closely relat-

ed order, on the basis of their having an undifferentiated embryo and little or no endosperm, very tiny seeds, the stamens reduced to one (rarely two or three), the fertile stamen adnate (fused) to the style to form a column, the flowers usually strongly zygomorphic (bilaterally symmetrical), the pollen usually coherent in masses called pollinia, and the ovary inferior always. The relationships between the subfamilies of the Orchidaceae are based on the number of fertile stamens, the character of the leaves, the consistency of the pollinia, the number of pollinia, and the associated structures of the pollinarium.

Criteria used to classify orchid groups

Annotated classification. The following classification is a recent one of wide acceptance among orchid taxonomists. Each subfamily is broken down into tribes, and in those cases in which more than one subtribe is found in a given tribe, the tribe is broken down into subtribes.

ORDER ORCHIDALES

Perennial herbs; terrestrial, epiphytic, or saprophytic, sometimes vinelike. Stems leafy or scapose, of sympodial or monopodial growth. Flowers solitary or in spikes, racemes, or panicles; bisexual, rarely unisexual, zygomorphic, always bracteate, mostly resupinate; perianth typically of 6 segments in 2 series, the outer series of 3 sepals green or often coloured and petaloid and usually all similar, the inner series of 3 petals with the central petal usually larger, often highly modified in colour and shape. Ovary inferior, of 3 carpels, ovules very numerous and minute. Style, stigmas, and stamens variously adnate into a singly highly complex structure (the column). Stamens 1 and then terminal on the column, or 2 and lateral (rarely 3), each anther 2-celled and introrse, the pollen usually in pollinia of 2 to 8 per anther. Fruit a capsule. Seeds abundant, minute, without endosperm, the embryo undifferentiated. A large order, variously estimated at 400 to 800 genera with 15,000 to 35,000 species, all in the single family Orchidaceae.

Family Orchidaceae

The only family in the order, it has the characters of the order.

Subfamily Apostasioideae

Perianth essentially regular, lip never deeply saccate; fertile anthers 2 or 3, elongate; style slender. One tribe with 10 to 15 species, restricted to Indo-Malayan region.

Tribe Apostasieae. Two genera: *Apostasia* and *Neuwiedia,* mostly terrestrial.

Subfamily Cypripedioideae

Perianth irregular with a deeply saccate lip; fertile anthers 2, subglobose; a conspicuous, flattened median staminode present; style relatively thick. One tribe with about 100 species of worldwide, but mostly tropical, distribution.

Tribe Cypripedieae. Four genera: *Cypripedium, Paphiopedilum, Phragmopedium,* and *Selenipedium,* mostly terrestrial in habitat.

Subfamily Neottioideae

Fertile anther 1, more or less erect, often dorsal, pollinia 2 to 4, soft and mealy; stems without corms or other thickenings, leaves not jointed at base. One tribe with about 1,000 mostly terrestrial species distributed mainly in the temperate zones of the world.

Tribe Neottieae. Nine subtribes: Chloraeinae, Cryptostylidinae, Diuridinae, Limodorinae, Neottiinae, Prasophyllinae, Pterostylidinae, Rhizanthellinae, and Spiranthinae.

Subfamily Orchidoideae

Fertile anther 1, erect or reclinate (rarely incumbent), persistent, usually broadly joined to the column, pollinia in soft masses, caudicles arising from the base of the pollinia. One tribe with about 1,000 temperate zone terrestrial species.

Tribe Orchideae. Four subtribes: Coryciinae, Disinae, Epipogiinae, and Orchidinae.

Subfamily Epidendroideae

Fertile anther 1, often incumbent, pollinia 2 to 12, usually hard, waxy, sometimes mealy, without caudicles or these terminal; leaves often jointed at the base. Four tribes with numbers variously estimated at 12,000 to 30,000 species. Distributed worldwide, but mostly in the tropics; both terrestrial and epiphytic forms occur.

Tribe Gastrodieae. Pollinia mealy, 2 or 4; plants saprophytic or green; leaves (if present) more or less fleshy (except in *Nervilia,* with fan-shaped leaves), not jointed at the base. Three subtribes: Gastrodiinae, Pogoniinae, and Vanillinae.

Tribe Epidendreae. Pollinia mealy to compact and hard, 2 to 12, usually club shaped or laterally flattened (except in *Coelogyne* and some species of *Polystachya*); viscidium present (usually more or less liquid) or absent; stipe rarely present (in *Polystachya*); leaves usually jointed at base. Eleven subtribes: Adrorhizinae, Arethusinae, Bletiinae, Coelogyninae, Eriinae, Laeliinae, Pleurothallidinae, Ridleyellinae, Sobralinae, Thelasiinae, and Thuniinae.

Tribe Malaxideae. Pollinia 2 to 4, completely naked, without caudicles of any sort; viscidia or stipes (usually double) rarely present. Four subtribes: Dendrobiinae, Genyorchidinae, Malaxidinae, and Thecostelinae.

Tribe Vandeae. Pollinia 2 to 4, dorsoventrally flattened if 4; viscidium normally always present; stipe almost always present. Ten subtribes: Catasetinae, Cymbidiinae, Cyrtopodiinae, Lycastinae, Maxillariinae, Ornitocephalinae, Oncidiinae, Sarcanthinae, Stanhopeinae, and Zygopetalinae.

Subtribes of uncertain affinity. Grobyinae and Pachyplectrinae.

Critical appraisal. Although the classification presented here is a recent one, there is still some need of examination of relationships in the orchids. The evolution of similar vegetative and floral structures in different orchid groups (convergent evolution) leading to the exploitation of the same types of pollinators has resulted in a situation in which many groups appear to be superficially similar. This same adaptation to pollinators has often obscured the actual relationship of numerous other groups.

BIBLIOGRAPHY

General references: C.L. WITHNER (ed.), *The Orchids: A Scientific Survey* (1959), contains a complete bibliography of orchid literature by geographical regions; F.S. SHUTTLEWORTH, H.S. ZIM, and G.W. DILLON, *Orchids* (1970); C.H. DODSON and R.J. GILLESPIE, *Biology of the Orchids* (1967); A.D. HAWKES, *Encyclopaedia of Cultivated Orchids* (1965).

Floristic works (New World): D.S. CORRELL, *Native Orchids of North America, North of Mexico* (1951); L.O. WILLIAMS, "The Orchidaceae of Mexico," *Ceiba,* vol. 2, no. 1–4 (1951, reprinted 1965); O. AMES and D.S. CORRELL, "Orchids of Guatemala," *Fieldiana, Bot.,* vol. 26, no. 1–2 (1952, 1953); R.E. WOODSON, JR., *et al.,* "Flora of Panama (Orchidaceae), *Ann. Mo. Bot. Gdn.,* vol. 33, no. 1–4 (1946), vol. 36, no. 1–2 (1949), reprinted 1965; G.C.K. DUNSTERVILLE and L.A. GARAY, *Venezuelan Orchids Illustrated,* 4 vol. (1959–66); C. SCHWEINFURTH, "Orchids of Peru," *Fieldiana, Bot.,* vol. 30, no. 1–4 (1959–61); F.C. HOEHNE, *Iconografia de Orchidaceas do Brasil* (1949); W. FAWCETT and A.B. RENDLE, *Flora of Jamaica (Orchidaceae)* (1910, reprinted 1963).

Floristic works (Old World): R.E. HOLTTUM, *Flora of Malaya,* vol. 1, *Orchids* (1953); F. PIERS, *Orchids of East Africa,* rev. ed. (1968); A.W. DOCKRILL, *Australian Indigenous Orchids,* vol. 1 (1969).

Orchid pollination: CHARLES DARWIN, *On the Various Contrivances by Which British and Foreign Orchids Are Fertilised by Insects,* first and second editions (1862, 1877); B. KULLENBERG, *Studies in Ophrys Pollination* (1961); L. VAN DER PIJL and C.H. DODSON, *Orchid Flowers: Their Pollination and Evolution* (1966).

Orchid classification: R.L. DRESSLER and C.H. DODSON, "Classification and Phylogeny in the Orchidaceae," *Ann. Mo. Bot. Gdn.,* 47:25–68 (1960); L.A. GARAY, "On the Origin of the Orchidaceae," *Bot. Mus. Leafl. Harv. Univ.,* 19:57–96 (1960). These two references contain excellent reviews of the literature on orchid classification.

Orchid Journals: American Orchid Society Bulletin; Orchid Review; Orchid Digest; Die Orchidee (in German); *Orquideologia* (in Spanish); *Malayan Orchid Review; Australian Orchid Review.*

(C.H.D.)

Ordovician Period

The Ordovician System was proposed in 1879 by Charles Lapworth for rocks exposed in the Arenig Mountains and eastward across the Bala District of North Wales (Figure 1). This area was part of that inhabited by the ancient British tribe of the Ordovices. It lies between an anticlinorium on the west, which exposes Cambrian rocks, and a synclinorium on the east, in which Silurian rocks are contained. Thus the type section is dominantly eastward-dipping, but it is complicated by subsidiary faulting and folding and, particularly in lower ground, is exposed only sporadically through a cover of Pleistocene glacial deposits. The sequence is approximately 3,450 metres (11,380

feet) thick and consists of graywackes, fine sandstones, and siltstones. Interbedded lavas and ashes occur in the lower portion, and some thin (two metres maximum) limestones are present in the upper part. Fossils occur at intervals throughout the sequence.

In proposing a new systemic name for the geological time scale (*q.v.*), Lapworth was, in part, offering a solution to the classic controversy over limits of the older Cambrian and younger Silurian systems. Lapworth was primarily concerned, however, with formally recognizing that Lower Paleozoic rocks contained three major faunas. This had become increasingly clear to contemporary students of these rocks. The basis for his Ordovician System was thus the distinctive characters of the fossils it contained, and this remains the basis for the system today, because the rocks have otherwise no unique characters for determining stratigraphic boundaries.

Strata of the Ordovician System are widely recognized in the continents of the Northern Hemisphere (but not in Africa from the Sahara southward or in peninsular India), in the Andes of South America, in Australia, in limited areas in New Zealand, but not in the Antarctic. Following the appearance and evolution of varied kinds of animals in the Cambrian Period, the Ordovician is notable for a great rise in numbers and variety of marine animals with a mineralized skeleton or shell. Some of these groups evolved rapidly, and the changes shown in the fossil record give the basis for classification and correlation of the Ordovician System. Few radiometric dates can be accurately related to this evolutionary record, but it is generally believed that the Ordovician Period began about 500,000,000 years ago and was about 70,000,000 years in duration.

Ordovician rocks of economic importance include a variety of building stones; road metals; slates from fine-grained, cleaved rocks; and decorative marbles (that is, massive limestones of unusual colour that often are derived from reef mounds). Some of the pure limestones are used in cement making. Some gypsum deposits are of Ordovician age, and various metalliferous deposits also occur. The latter include syngenetic iron ores, gold, and lead-zinc ores in limestones. Certain porous Ordovician sandstones are important as sources of water.

This article also treats the rocks, life forms, and environments of Ordovician time. See CAMBRIAN PERIOD and SILURIAN PERIOD for related coverage of the prior and subsequent time intervals, and FOSSIL RECORD for further details of Ordovician in vertebrate paleontology.

ORDOVICIAN ROCKS

Figure 2 shows areas in which Ordovician rocks crop out and which were covered by the sea for all or part of the period. No land plants or animals are known in Ordovician rocks so that terrestrial sediments cannot be recognized, if they exist. The history of Ordovician lands therefore is unknown. The marine successions are of two major types: those in geosynclines, which are thick and may be deformed and metamorphosed, and the thinner, less complete, shelf successions, which transgress over shield areas and are little deformed and unmetamorphosed. The geosynclines are situated outside the edges of the shield areas of Precambrian rocks: in North America on all sides of the shield; in South America and Australia on the western and eastern edge, respectively; and in northern Europe on the northwestern and southern edges of the Baltic shield. The folded Paleozoic rocks of the Urals lie between the Baltic Shield and the Angara Shield of Siberia, and a central Asian geosyncline lies to the south of the latter shield. Although the main deformation in these geosynclines was in the Middle or Late Paleozoic, intense deformation did take place earlier in some geosynclines during the Ordovician.

Geosynclines

One of the best-known geosynclinal regions is that which runs through Norway, northwestern Britain, Newfoundland, the Appalachians of northeastern North America, and parts of east Greenland and Spitsbergen. In this long and relatively narrow, tectonically active downwarp, there may have been a central zone in which fine-grained sediments accumulated slowly. In southern Scotland, for example, 30 metres of rock represent the entire Upper Ordovician. This central belt was flanked by intensely active zones, in which sequences, of the order of 4–5 kilometres, of graywackes and volcanic rocks accumulated. These sediments were derived in part from rising, narrow welts within the basin and in part from piles of volcanic lava and ash, rapidly extruded and building up to and above sea level and accompanied by intrusions.

Faults were active in these geosynclinal regions. Thick wedges of conglomerates may have accumulated on the downthrown side of a fault bounding an area being rapidly eroded. The trough between two such faults may have preserved a far thicker sequence than the relatively raised areas on either side. Rapid changes in sedimentary facies (*q.v.*) thus characterize the rocks of such zones. During the Ordovician Period the most active part of the mobile zone was strongly folded, metamorphosed, and upraised; and in some areas slices of rock were carried many kilometres laterally on thrust planes. Outside the zone of intense deformation, graywackes and volcanic sequences were strongly folded; and these pass laterally into successions that, although still thick as a result of continued subsidence, were built of limestone, sandstone, and shale. These rocks were deposited on the shield side of the geosyncline, adjacent to areas of low relief not undergoing active uplift and erosion. Here the supply of sediment was less and sands were better sorted. There is every transition between these sequences, with wedges of graywacke extending into the limestone-shale succession or of limestones extending basinward during periods of lesser activity.

In Britain, the Ordovician rocks of Wales and the Lake District include graywackes and volcanics, with rare, thin limestones that are strongly folded and faulted; in northern Scotland, Ordovician and older rocks were deformed and metamorphosed during the Early Ordovician and

Adapted from H. Fullard and H. Darby (eds.), *The Library Atlas* (1967); George Philip and Son Ltd., London

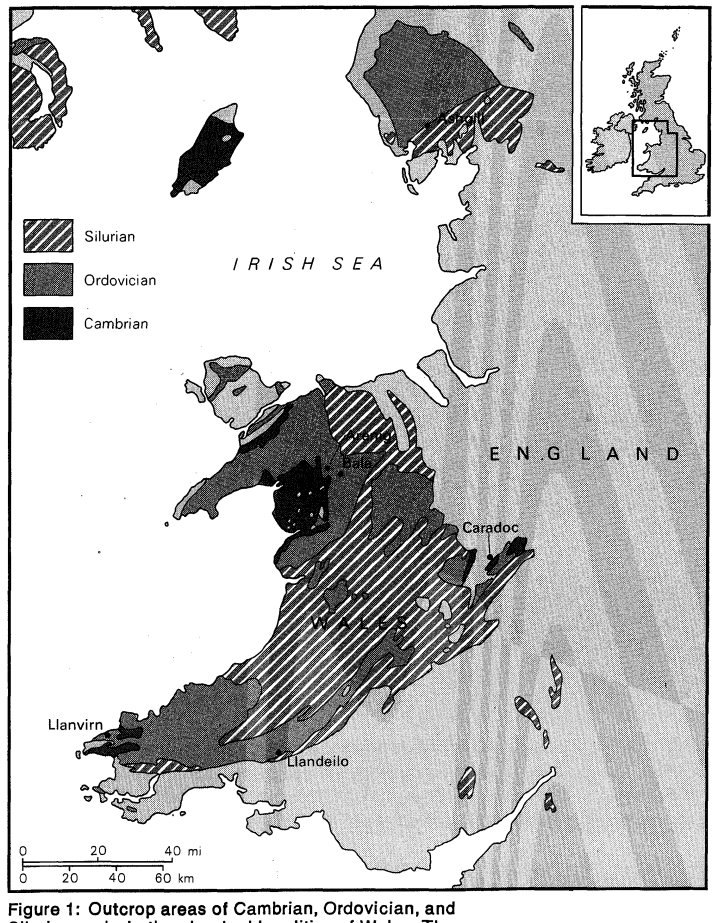


Figure 1: Outcrop areas of Cambrian, Ordovician, and Silurian rocks in the classical localities of Wales. The differentiation of these systems was initially based on studies of this region.

represent the most active zone of the geosyncline. The central and southern Appalachians of eastern North America are partly formed by a thick Ordovician limestone and shale sequence into which extend tongues of graywackes and red sandstones derived from welts raised in the active geosynclinal zone to the east. In the Great Basin area of the western United States, the Ordovician is represented by 500 metres (1,600 feet) of limestone and quartzite, over which dark shales and cherts of a different sequence were thrust from the west. Geosynclinal regions in other continents show the same broad features; *i.e.*, the graywacke-volcanic-chert facies that characterizes the most tectonically active zones and beyond this zone the limestone-sandstone-shale successions.

Shelf areas

At the margins of the geosynclines in the Northern Hemisphere and Australia, the seas spread onto the shields; and a thin and incomplete series of sediments, dominantly limestone, was lain down. These were the shelf areas. In central North America and the Arctic islands, Upper Ordovician limestones typically contain corals and large cephalopods. There are associated evaporite deposits. These shelf sediments were transgressive on Precambrian rocks of the shield.

On the Baltic Shield, the Ordovician limestone and shale sequence is relatively complete and rests on essentially undeformed Cambrian strata. In central Sweden, small algal reefs developed in the shallow Upper Ordovician seas, associated with a rich trilobite-brachiopod-mollusk-an fauna. Notably, angular deposits at the margin of the reefs were formed by debris broken from them, and between the reefs were mudstones and muddy limestones with a different trilobite-brachiopod fauna and a variety of echinoderms. On the Siberian Platform a moderately thick sequence of carbonate rocks was deposited with abundant shallow-water faunas and associated gypsum-bearing beds; similar rocks are found in central and western Australia.

ORDOVICIAN LIFE

Ordovician invertebrate faunas differ from those of the Cambrian in the much greater variety of groups with calcareous shells. Most localities in fossiliferous Cambrian rocks yield abundant trilobites accompanied by species of inarticulate brachiopods and primitive mollusks. Other groups occur more rarely. In contrast, Ordovician rocks yield quite different trilobites, which at many localities are outnumbered by a variety of brachiopods, bryozoans, mollusks, echinoderms, corals, and representatives of other groups.

Marine fossils

Trilobites underwent a major evolutionary change during the latest Cambrian and earliest Ordovician, so that only a few representatives of Cambrian families persisted, outnumbered by a variety of new groups, each of which evolved rapidly. Broadly speaking, Ordovician trilobites had fewer thoracic segments and a relatively larger pygidium (tail section) than Cambrian ones, and families with proparian sutures (a type of molt line on the head) were an important part of the faunas. Other arthropods are rare in Ordovician rocks, except for the calcareous, bivalved shells of ostracods, which became abundant in the later Lower Ordovician and were varied throughout the Upper Ordovician.

A great increase in kinds and numbers of articulate (hinged) brachiopods is evident in the Early Ordovician, and succeeding strata yield the shells of a variety of rapidly evolving groups. Bryozoans are almost unknown in the Cambrian, but by the late Lower Ordovician their massive and branching colonies, supported by a calcareous skeleton, became an important element in shallow-water faunas, even building small reeflike mounds on the sea bottom. The expansion of these extinct bryozoan groups continued through the period. The abundance of gastropod shells and of straight, curved, and coiled shells of cephalopods (distantly related to the living pearly *Nautilus*) is characteristic of Ordovician rocks from the earliest part of the system; pelecypods became abundant in the Late Ordovician.

In the Early Ordovician, species of groups that survive today—the starfish, brittle stars, and crinoids—were present, and the number and variety increased during the period. Of extinct echinoderm groups, the cystoids were the most varied and important during the Ordovician. Thus, by the late Lower Ordovician, echinoderms were sufficiently numerous for parts of their skeletons to accumulate and form limestone beds. The calcite of echinoderm skeletons is lain down in crystal continuity and when broken shows cleavage planes, so that these beds have a distinctive character not seen in older limestones. Corals, both solitary forms with prominent radially arranged septa (tetracorals) and colonies in which horizontal plates divide the corallites (tabulate corals), were present in the late Lower Ordovician and widespread before the end of the system. Bryozoa, colonial corals, and sponges with thick calcareous or siliceous skeletons played their part in building small reeflike mounds or in forming thin but extensive sheets of limestone in late Lower and Upper Ordovician time.

Graptolites were colonial, mainly pelagic, animals with a skeleton of tough, organic, chitin-like material. They evolved rapidly throughout the Ordovician, and while they are conspicuous in dark shales because they may be the only abundant fossils, they occur in all other types of sedimentary rocks.

Small (2–44 millimetres) toothlike fossils of organic material (scolecodonts, the jaw apparatus of worms) or calcium phosphate (conodonts, a group of uncertain affinity since the complete animal is unknown) are abundant fossils. Conodonts show a succession of widespread types indicating a considerable evolution. Dissociated and broken plates of true bony structure indicate the presence of vertebrates—probably primitive, jawless fish. These are so rare and fragmentary that the form of the animal is unknown.

The remains of vascular plants are not known in the Ordovician, but a variety of calcareous algae have been recognized. Microscopic organic fossils such as flask-shaped Chitinozoa or spherical, spiny acritarchs have been extracted from limestones, siltstones, and sandstones by special techniques.

Stratigraphically important fossils

Fossils of importance for zonation and correlation are those that are abundant in a wide variety of rocks over large areas and are of organisms that evolved rapidly. The hard parts must be preserved in a manner that enables them to be separated from the rock and studied in detail. For the Ordovician Period, brachiopods and trilobites fulfill these conditions and have been studied long and intensively so that, in most areas in the world, stages and zones are based on them. The small bivalved shells of ostracods are also abundant and widespread and show considerable evolution but have not been studied so intensively. In limestones, where brachiopods and trilobites may be abundant, nautiloid cephalopods and gastropods may also occur in considerable numbers. The cephalopods evolved rapidly. Thus, in the Lower Ordovician limestone-dolomite facies in North America, zonation and correlation is in part based on the remains of these animals.

In the Upper Ordovician, corals became widespread and varied and are also used to a limited extent in correlation. In the late Lower and Upper Ordovician sections, Bryozoa are common in rocks of shallow-water facies, but stratigraphic usefulness requires detailed preservation of the complete skeleton so that it may be studied in thin section. Since only casts of the skeleton are usually preserved in shales, mudstones, and sandstones, they cannot be so studied; and the use of these fossils in correlation is limited.

In recent years there has been increased emphasis on study of fossils obtained by dissolving rocks in acid. Such work has shown the apparently worldwide distribution and abundance of conodonts. Though each conodont is only one element in the assemblage of an individual, study of relative numbers and kinds of these elements shows groupings that are proving valuable in correlation. Other potentially useful fossils revealed by acid digestion of rocks are chitinozoans and acritarchs.

It was discovered long ago that in black shale and fine silts graptolites were the most conspicuous and abundant

fossils. A scheme of graptolite zones for the Ordovician of Britain was erected by Charles Lapworth and his students in the early years of this century. Many of the forms appear to be worldwide in distribution, so that the British zones are often quoted as a world standard, though a quite different scheme for the Lower Ordovician is used in Australia. Graptolites do occur in rocks of all types but generally are rare and poorly preserved in siltstones and sandstones. In limestones they occur much less commonly, though certain limestones and associated cherts have yielded the best-preserved and most completely known graptolites.

This rather general restriction of abundant graptolites to dark shales and the restriction of varied brachiopod-trilobite-mollusk-bryozoan faunas to limestones, sandstones, and siltstones has led to the greatest problem in Ordovician correlations: that between these two sedimentary facies and their separate zonal schemes. The relations between facies are revealed where they interfinger, but such interfingerings are not well enough known to establish detailed correlations. It appears that the best-developed graptolite shale sequences were in parts of the geosyncline far from the shelves, and on the shelves strata rich in brachiopods and trilobites were deposited. These two general regions have had different structural histories, so that the original relation between them is complicated and obscured by folding and faulting. Thus in the western United States the graptolite facies is best known from thrust plates overlying limestones, and the transition zone, if it existed between the facies, is unknown.

Faunal realms

A faunal realm is a large region, which may include parts of more than one continent, in which the assemblage of genera of one or more groups of fossils was similar during a substantial portion of the time considered. In general, during the Ordovician Period, such realms were marked in the lower part but in the upper became progressively less distinct. Study of the distribution of trilobite genera has shown that in the Lower Ordovician there were two realms, one including almost all of North America and the Arctic islands, Greenland, western Ireland, Scotland, Spitsbergen, northern Europe, and the U.S.S.R. The other realm included the Andean region of South America, northern Florida, central Europe (including England and Wales), the Mediterranean area, the Himalayas, China, Southeast Asia, Australia, and New Zealand. Certain family groups and particular genera of other families were confined to one region or the other, and few were common to both regions.

In the Upper Ordovician the distinctiveness of these realms began to break down. Genera previously confined to one region then appeared in both. In the Late Ordovician the genera of a single fauna were widely distributed in the Northern Hemisphere. The distribution of brachiopods exhibits this pattern. In the early Upper Ordovician the brachiopod faunas of Scotland and the Appalachians were remarkably similar but very different from the contemporaneous faunas of central Europe.

The worldwide distribution of genera of other groups of shallow-water fossils is not well known; cephalopod distribution in the Lower Ordovician shows some parallels and some contrasts with that of trilobites. Modern shallow-water faunal realms are partly determined by water temperature and ocean currents. The latter influence distribution by transporting the floating larval stages. Barriers to such distribution are land masses and wide stretches of deep water. The larvae cannot survive the required transportation and thus cannot settle and colonize the sea floor.

The paleogeography of the Ordovician, the oceanic circulation, and water temperatures presumably, were important factors in determining the faunal realms, but because there is as yet no clear understanding of the distribution and shape of Ordovician continents and oceans, it is impossible to determine oceanic circulation. Organisms that float either by attachment to other organisms (*e.g.*, graptolites attached to floating masses of seaweed) or by possessing structures containing gas or fat or organisms that swim in the surface oceanic waters may be distributed widely. In the Upper Ordovician, graptolites were so distributed, the rapidly evolving successions of kinds being closely similar all over the world. Lower Ordovician graptolites were probably also floating forms but do show a provinciality. The kinds and their stratigraphical arrangement are remarkably similar in the Americas, Australia, and China but differ in both kinds, relative abundance, and stratigraphical range from those of Europe. This does not mean that the oceans in these areas were completely separated from each other but that some factors restricted migration and mixing, factors that were not operative in the Upper Ordovician. Patterns of ocean currents are obviously such a factor, and it may be that these patterns changed markedly between Early and Late Ordovician time, so that by the Late Ordovician both floating and benthonic faunas were similar over large areas of the world.

Evolution during the Ordovician Period

The Ordovician was preeminently the time of major expansion of the invertebrates, and many new species, genera, and families of major groups (classes or orders) appeared. Trilobites, which dominated the Cambrian record in individual numbers as well as kinds, began to be rivalled and then overshadowed in the Ordovician by corals, bryozoans, brachiopods, mollusks, and echinoderms. This reflects the evolutionary success of these groups in colonizing the shallow sea floor, a success that came long after their first appearances. The nautiloid cephalopods were expanding rapidly in the earliest Ordovician and, if feeding like the living *Nautilus*, they would have been formidable predators. Their varied forms suggest a variety of living habits.

It may be that the increase in forms with a hard protective shell is related to the rise of these (or other) predators. If such a relationship existed, however, it may have been far more complex than this. An increase in food supplies and changes in the salts in solution in the oceans may have created the conditions for this evolution, and predatory forms may have evolved as a response to an increased food supply of bottom-dwelling invertebrates. The fossil record is biassed in favour of forms with a mineralized shell. Since soft-bodied animals are rarely preserved, their numbers and kinds cannot be estimated. Thus the picture is incomplete, revealing but not explaining this great expansion of hard-shelled invertebrates. The Ordovician is notable for the first records of vertebrates in the form of bone fragments.

CLASSIFICATION AND CORRELATIONS OF ORDOVICIAN ROCKS

In the quarter century following Lapworth's proposal of the Ordovician System it gained acceptance, and in Britain the names for the series composing it, shown in the Table, became established. Each series is a sequence of rocks exposed in a particular area (Figure 1). The Arenig rocks were mentioned by Lapworth as the basal unit of his system, resting unconformably (a break in continuous deposition) on the Tremadoc Series of the youngest Cambrian. The rocks and fossils of the Llandeilo and Caradoc areas had been described earlier by Sir Roderick Murchison as part of his Silurian System and were familiar to British geologists. The Llanvirn Series contained distinctive graptolite faunas, and the Ashgill section was recognized as including the youngest faunas. These standard series are in different areas, not part of a single continuous section; in the Arenig-Bala area (the type sequence of the system) no rocks of Llandeilo age are recognized, and the strata of Arenig, Llanvirn, Caradoc, and Ashgill age are fossiliferous at only a limited number of widely spaced horizons. Thus the standard section is a composite, and all but the Llanvirn Series are rocks mainly in the shelly facies, that is, shallow-water sediments with brachiopod-trilobite faunas.

Facies problems

Early in the 20th century the graptolites of Britain were described and a zonal scheme for Ordovician rocks proposed. These zones are commonly quoted as a standard, but they were based on studies of dominantly black shale or black shale and graywacke sequences exposed in areas other than those of the type series. These zones are difficult or impossible to define exactly in the type Arenig, Llandeilo, or Caradoc areas. Such difficulties of corre-

Major Divisions of Ordovician Rocks and Their Correlation

Britain			North America			Baltic	South China	U.S.S.R.
ORDOVICIAN	UPPER	Ashgill	ORDOVICIAN	UPPER	Cincinnatian	Harju	Chientang -kiang	Upper
		Caradoc		MIDDLE	Champlainian	Viru	Neichia -shan	Middle
	LOWER	Llandeilo						
		Llanvirn		LOWER	Canadian			
		Arenig				Oeland	Ichang	Lower
Upper Cambrian		Tremadoc	Upper Cam-brian					

lating between the shelly and graptolitic facies are common in the study of Ordovician rocks in all continents. Sections that show some interfingering of the facies are important in correlation but are rare and do not show the relationship clearly or completely. Correlation is also hampered by the lateral changes in shelly faunas and the existence of faunal realms; it is almost impossible, for example, to correlate the type Caradoc Series with rocks of the same age in Scotland because the shelly faunas have little in common. Rare occurrences of graptolites in the shelly sequences have to be used for correlation. Similarly, intercontinental correlations between the shelly standard divisions of other countries (see the Table) and Britain are largely based on graptolites; and even with these relatively widespread fossils there are difficulties arising from faunal realms, particularly in the Lower Ordovician.

System boundaries

Definition of the upper boundary of the Ordovician System has offered relatively little difficulty. It is based on the changes in the widespread Late Ordovician brachiopod-trilobite fauna and the graptolite fauna from Ordovician to earliest Silurian.

The lower boundary (Table) is variously defined in different countries. Lapworth originally designated the base of the Arenig Series as the base of his system, but he subsequently and erroneously modified his view and included part of the Tremadoc Series in the Ordovician. His Scandinavian colleagues adopted his modification and included all the Tremadoc Series in the Ordovician, whereas the United Kingdom Geological Survey adhered to his original view. In the United States the Canadian Series is taken as the earliest Ordovician. It now appears likely that the earliest Canadian rocks may correlate with the upper part of the Tremadoc Series only. Thus in existing practices the base of the system appears to be placed at three different levels (Table). In Britain it has been customary to divide the Ordovician System into an upper and lower portion, the base of the upper being at the base of the most widely recognized graptolite zone. A threefold division has been used in other countries, but the boundaries between these divisions are placed at different positions. Thus, care has to be exercised in interpreting these terms, and the twofold division has been used here.

ORDOVICIAN PALEOGEOGRAPHY AND CLIMATE

The areas of the present continents on which shelf or geosynclinal sediments were deposited in Ordovician time are shown in Figure 2. Seas covered these areas and probably other areas from which Ordovician rocks have been eroded, as well as areas in which Ordovician strata may be concealed by younger rocks. Further, the elongated geosynclines appear narrower than they originally were because of the contraction resulting from folding and faulting. A sea probably covered the Cordilleran area of South America and was connected to one covering much of North America, western Europe, and North Africa. From the Mediterranean region a sea probably extended continuously along the present mountain areas into the Himalayas, southeastern Asia, and on to cover parts of present Australia and New Zealand. The Urals and probably much of northern Asia were also covered. The similarities of faunas between the deposits suggest these connections. The major land areas of the Ordovician appear to have been the eastern half of South America, Africa from the Sahara southward, peninsular India, and the Antarctic.

Figure 2 is plotted on present-day geography, for lack of any real alternative. In the north Atlantic areas the rocks, their faunas, and structures reveal evidence suggesting that during the Ordovician Period they were much closer to each other than at present. There is also evidence accumulating to suggest that the Atlantic Ocean is geologically young and that the ocean of Ordovician time was narrower and in a different position. For example, it may have separated northern England from Scotland and parts of present northeastern North America from each other.

It has long been suggested by geologists considering continental drift (*q.v.*) that in the Late Paleozoic the present areas of eastern South America, Africa, peninsular India, Australia, and the Antarctic were in close juxtaposition. A large part of this continent appears to have been land or covered only by shelf seas (Australia) during the Ordovician Period. Did it form a single continental mass in the Ordovician? If so, what were the positions relative to it of the Canadian, Baltic and Angara shields and of the marginal geosynclines?

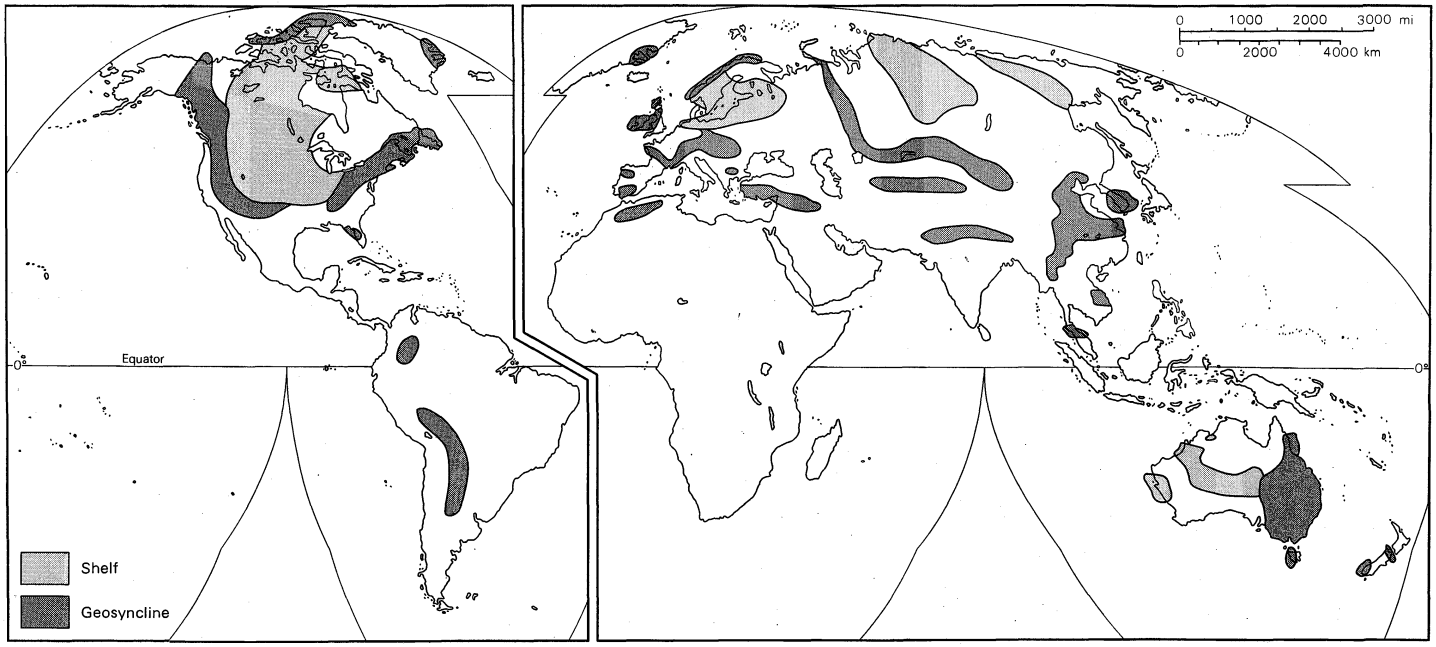


Figure 2: World distribution of Ordovician deposits showing the location of shallow seas (shelf areas) and geosynclinal areas during Ordovician time.

These unanswered questions make it impossible to draw a meaningful paleogeographical map of the world in Ordovician times. Further, other lines of evidence from rock magnetism suggest that the magnetic poles have changed in position during geological time. If there has always been a close relation between the magnetic and rotational poles of the earth, then the poles of rotation have varied in position relative to particular areas of today's continents. Since Ordovician geography cannot be reconstructed, the positions of the oceans and land areas relative to the Ordovician poles cannot be determined. Hence Ordovician climate remains mysterious, except as one interprets evidence within a single present continental area.

In the Late Ordovician shelf seas of northern North America, large nautiloid cephalopods and coral faunas were characteristic, and gypsum-bearing rocks (formed by evaporation of sea water) were associated. On the Siberian Platform limestone was the dominant Ordovician rock, and gypsum deposits occur. Evidence of this kind suggests that these present northern regions were warm water areas in the Ordovician, lying adjacent to the Equator. The variety and abundance of certain faunas in these areas also suggest warm waters. Which regions may have been cooler is difficult to establish. There is no evidence of extensive glaciation during the Ordovician, so that ice caps do not appear to have existed; and climate may have been less extreme than that of today.

BIBLIOGRAPHY

General: B. KUMMEL, *History of the Earth* (1961), a general introduction and some stratigraphical columns; H.B. WHITTINGTON and A. WILLIAMS, "The Ordovician Period," in *The Phanerozoic Time-Scale*, pp. 241–254 (1964), a discussion of the history, subdivision, and correlation of the system.

North America: M. KAY and E.H. COLBERT, *Stratigraphy and Life History* (1965), a recent introduction.

United Kingdom: A. WOOD (ed.), *The Pre-Cambrian and Lower Palaeozoic Rocks of Wales* (1969), a review and discussion of recent work; *British Regional Geology*, a series of handbooks of the Institute of Geological Sciences, London, that give accounts of the various regions.

Northern Europe: L. STORMER, "Some Aspects of the Caledonian Geosyncline and Foreland West of the Baltic Shield," *Q. Jl. Geol. Soc., Lond.*, 123:183–214 (Dec. 1967), a general account, with many references to the Scandinavian region.

Central Europe: J. SVOBODA *et al.*, *Regional Geology of Czechoslovakia*, part 1, *The Bohemian Massif* (1966), a classic account of the Bohemian area.

U.S.S.R.: D.V. NALIVKIN, *The Geology of the U.S.S.R.*, trans. by S.I. TOMKEIEFF (1960), a short outline; B.S. SOKOLOV *et al.*, *Stratigraphy, Correlation and Palaeogeography of the Ordovician Deposits of the USSR*, Rept. 21st Session, Nordon, Int. Geol. Congr., pt. 7, pp. 44–57 (1960).

Australia and New Zealand: D.A. BROWN, K.S.W. CAMPBELL, and K.A.W. CROOK, *The Geological Evolution of Australia and New Zealand* (1968).

Paleontology: R.C. MOORE (ed.), *Treatise on Invertebrate Paleontology* (1953–), the available volumes have the most recent information on many Ordovician fossils; H.B. WHITTINGTON, "Phylogeny and Distribution of Ordovician Trilobites," *J. Paleont.*, 40:696–737 (May 1966), illustrations of many trilobites and discussion of evolution, faunal provinces, and climate.

(H.B.W.)

Ore Deposits

The word ore is derived from an Anglo-Saxon expression referring to metals such as copper and alloys such as bronze. It is used to mean a concentration of metals or metalliferous minerals that can be separated from the associated rock and mined at a profit. Present usage includes elements such as sulfur, when the sulfur is extracted from the mineral pyrites, and nonmetalliferous minerals such as fluorite and barite that occur in veins and require separation from the associated rock.

The profitability of mining any concentration of minerals can, and frequently does, change from year to year depending on a number of technological and economic factors. For these reasons the quantity and quality of ore in any particular concentration, or association of concentrations, must be defined on the basis of tonnages and grades of proven, probable, and possible ore. Grades refer to the actual metal content of the ore, commonly expressed in pounds, ounces, or metric units of weight of a particular metal per unit weight of ore. Proven ore refers to a known volume of ore of defined limits that has been blocked by drilling. Probable ore relates to extensions of proven ore, and the limits are not precisely defined. Possible ore is classed as a prospective volume where spatial or geological relationships to known ore deposits warrant the assumption that mineral concentrations of economic value may be found.

Associated with the ore minerals in an ore deposit is worthless material called gangue, consisting of rock and unwanted minerals, which is mined with the ore minerals, subsequently separated in the milling processes, and discarded as dumps of waste. The unwanted minerals contained in these dumps may eventually become valuable, and, if the grade of such dumps is sufficiently high, they may be reworked at a profit. Also associated with an ore

deposit is the country rock, or host rock, enclosing the deposit. The country rock can be of any type, although igneous and metamorphic rocks (those formed at high temperatures and pressures) are generally better host rocks for metalliferous ore deposits than are sedimentary rocks. Many exceptions are now known, but regardless of rock type, mineral prospecting and development generally require drilling, open-cut operations, or underground mining before a reliable assessment can be made of the reserves present and the feasibility of mining them for profit.

Metals and metalliferous minerals

Some metals, including copper, silver, gold, and platinum are mined in the native state. Others are rarely found in this state and only in quantities too small to be of economic importance. The outstanding deposits of native copper are in the Lake Superior area of Michigan, where the copper occurs in interbedded conglomerate (sedimentary rock with particles of diverse size) and basaltic lava flows, many of which are vesicular or amygdaloidal (that is, the minerals were deposited in cavities, or amygdules, after the flow formed). In the conglomerate the copper fills interstices and replaces matrix and pebbles; in the lava flows the copper fills the vesicles. Originally discovered by Indians who manufactured ornaments from the malleable copper and later noted by the Jesuits in the 17th century, these native copper deposits were first mined in 1845.

Native silver was found in quantity in the Cobalt District of Ontario during the early years of the 20th century. Some of the ore was phenomenally rich, ranging from 200 to 1,000 ounces of silver per ton. Slabs of up to 725 kilograms (1,600 pounds) containing 40 percent by weight have been extracted from the Cobalt silver mines. The silver occurs in veins that have filled fissures in conglomerate and other sedimentary rocks.

Gold commonly occurs in the native state alloyed with copper and silver. It originally formed in fracture-filling veins associated with chalcopyrite, pyrite, and other sulfides, with silicate minerals such as quartz, sericite, and tourmaline, and with the carbonate minerals calcite and ankerite. Erosion of such veins results in the eventual transportation of the released gold by streams and rivers in the form of fine particles, flakes, and nuggets; these become concentrated, as in crevices of the bedrock underlying the gravel or at the upstream end of sandbars, and are called placer deposits.

Platinum is mined chiefly in the native state; it occurs alloyed with other metals of the platinum group, namely, osmium, iridium, palladium, rhodium, and ruthenium. It also occurs as platinum arsenides and sulfides. Invariably associated with nickel and chrome mineralization in basic igneous rocks (those rich in iron and magnesium), it is mined mainly in the Ural Mountains of the U.S.S.R., in the Sudbury District of Ontario, and in the Bushveld District of South Africa. Production from the U.S.S.R. is derived principally from placer deposits that have been derived from magmatic concentrations of platinum in olivine and pyroxene-rich igneous rocks called dunite and pyroxenite. In the Sudbury District, the platinum is associated with massive copper–nickel sulfide ores, which are the principal metals sought. In South Africa, platinum is mined from magmatic concentrations associated with copper–nickel sulfides in dunite pipes.

Metalliferous minerals are chemical compounds of metals, commonly sulfides and oxides, that are found in the rocks. Some are common and some are rare. Copper, nickel, lead, and zinc are chiefly mined as sulfides, some of which have a metallic appearance; iron, aluminum, and uranium are mined as oxides, many of which have an earthy appearance. Where these minerals are concentrated they form metalliferous deposits that can be defined as ore deposits. High concentrations, such as veins of copper and lead ore, may warrant underground mining; low concentrations, such as porphyry copper deposits, where the copper minerals are disseminated throughout a large mass of fractured rock, can be mined economically only by open-pit operations handling large tonnages. Some deposits of iron and most deposits of aluminum lie at the surface.

Nonmetallic industrial minerals

Nonmetallic industrial minerals include those extracted from ore, such as diamonds and asbestos. They also include minerals that form deposits in themselves, such as beds of gypsum, salt, and sulfur. Other nonmetallic minerals used in industry are feldspar, quartz, mica, spodumene, gemstones, fluorite, barite, magnesite, sylvite, and many others. Spodumene and magnesite are also metalliferous because they are mined for the extraction of the metals lithium and magnesium, respectively.

A distinction must be drawn between industrial minerals and industrial materials such as soapstone, pumice, limestone, and shale, which are rocks, and clays such as kaolin, which are mixtures of hydrous aluminum silicates derived from the weathering of rocks.

Strategic minerals

Essential minerals that are in short supply, or that are found mainly in certain areas of the world from which they must be imported, are known as strategic minerals. The list of such minerals will vary from country to country and from time to time. Some countries are dependent on others for supplies of strategic minerals or their concentrates. Withholding such minerals from the world markets can cause serious shortages and consequent price rises on the international metal exchanges. The movement of minerals is affected not only by the amounts produced but also by political and economic considerations. Stockpiles of strategic minerals, concentrates, or the metals derived from them are hedges against the possibility of world shortages and high prices.

Worldwide exploration by mining companies, and the consequent discovery of new ore deposits, tends to modify the dependence of one country on another for strategic minerals. Antimony and titanium are examples of essential metals derived from strategic minerals. For many years China supplied two-thirds of the world's requirements for antimony, a metal that has the property of expanding upon cooling, a property that is used for type metal and hard lead alloys. China is still an important producer on the world market, but discoveries of the principal antimony sulfide mineral (stibnite) in other parts of the world have tended to lower the dependence on exports from China. Titanium, a light, strong, extremely heat-resisting metal of increasing importance to the aerospace industry, is mainly derived from the titanium oxide mineral rutile, which is supplied largely from beach sand-mining operations in Australia, South Africa, and India.

Conservation and future sources

Conservation of an ore deposit means the utilization of the deposit in such a way that the maximum amount of ore is ultimately mined. This means further that within reasonable economic limits the lower grades of ore are not sacrificed by mining methods that deplete too rapidly the higher grades for the sake of immediate monetary gain. Higher grades of ore are usually mixed with lower grades to average a uniform grade for the daily intake of a mill. The milling of highgrade ore only is seldom justified and frequently results in mining practices that preclude the subsequent recovery of large tonnages of lower grade. Such practices are infrequent, and mine management now generally fosters the long life of a mine.

Conservation is particularly important because ore deposits are nonrenewable natural resources. Although new ore deposits may continually be formed, they do so over periods of many thousands of years, commonly at depths of hundreds or thousands of feet; thus it is probable that the existing ore deposits are the only ones that man will ever be able to use. As known deposits have been depleted in the past, new ones have been found. To date these have been either extensions of known deposits, deposits that crop out in explored parts of the world but which hitherto have escaped discovery, or unexposed deposits lying just below a surface covered by soil or vegetation. As time goes on fewer of the exposed deposits will be found, and discovery will depend increasingly on geophysical and geochemical methods to find the hidden deposits. In those parts of the world that have not been explored and mapped in some detail on the ground, exposed ore deposits may be found for many years to come. Whether or not such deposits are mined will depend on their economic viability at the time of discovery.

Technological developments increasingly have made possible the economic mining of low-grade deposits. This trend may continue and in the future make available large low-grade deposits that are not mined today. Technological developments involving dredging, suction, and other methods may in the future be used to mine phosphate nodules from deeper parts of the ocean floor, and rutile, zircon, ilmenite, and gold from ancient beaches now under a shallow sea. Magnesium has been commercially extracted in large quantities from seawater for many years, but the vast volume of seawater required to produce relatively small quantities of most metals is an inhibiting factor.

The manufacturing of synthetic minerals used for their crystallographic or mineralogical properties will continue to be developed in the future; such synthetics include diamonds, sapphires, rubies, quartz, and certain micas.

This article treats the nature, origin, and classification of ore deposits, the kinds of earth processes that are involved in their formation, and the world distribution of the various types of deposits. For further information on the kinds of rocks and minerals of concern see IGNEOUS ROCKS; METAMORPHIC ROCKS; SEDIMENTARY ROCKS; MINERALS; and the several articles on individual mineral groups, such as NATIVE ELEMENTS; and SULFIDE MINERALS. See also GEOCHEMICAL EQUILIBRIA AT HIGH TEMPERATURES AND PRESSURES; ROCK METAMORPHISM, PRINCIPLES OF; and ROCK DEFORMATION for additional details on earth processes of relevance to ore formation.

NATURE, GENESIS, AND CLASSIFICATION OF ORE DEPOSITS

Size, shape, and depth distribution. Ore deposits show considerable variations in size, shape, and depth of occurrence. They can range in size from a few hundred to many hundreds of millions of tons of ore. High-grade vein deposits of the copper minerals chalcopyrite, bornite, and chalcocite, for example, will generally have considerably less tonnage than low-grade copper porphyry deposits containing the same minerals. The latter commonly contain tens of millions (some contain hundreds of millions) of tons of ore averaging a fraction of 1 percent copper by weight. This fraction, or average grade, is an important factor in evaluating the worth of the deposit and the viability of developing it as a mine. The range of grades found that can economically be mined are commonly 0.4–0.8 percent copper, but the economic limit or cutoff grade may subsequently be lowered by technological developments.

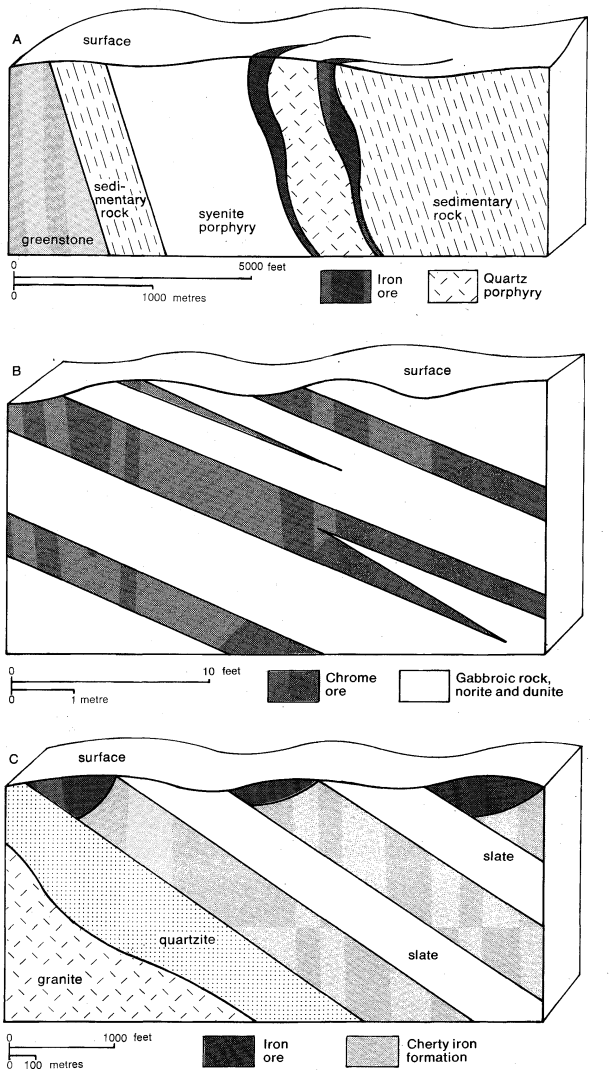


Figure 1: Deposits of (A) magnetic iron ore, Kiruna, Sweden; (B) chrome ore, Bushveld, South Africa; and (C) residual iron ore, Mesabi, Minnesota.

Iron-ore deposits

Iron-ore deposits, largely consisting of the iron oxide minerals magnetite and hematite, also show a great range in size depending on their nature and genesis. Deposits of magmatic origin such as the magnetite ores at Kiruna (Figure 1A) in Sweden, and of metasomatic (replacement of minerals by solution and redeposition) origin such as the magnetite ores at Cornwall, Pennsylvania, are small compared to the bedded deposits of residual origin such as the hematite ores of Lake Superior (Figure 1C) and Western Australia. The Kierunavaara Deposit, the largest in the Kiruna District, probably contains a few hundred million tons of iron averaging somewhat more than 60 percent iron by weight. The ore body is sill-like (that is, in the form of a tabular body that parallels the structures of host rocks) and has been emplaced between Precambrian (older than 570,000,000 years) syenite porphyry in the footwall and quartz porphyry in the hanging wall. The iron-ore deposits at Cornwall, Pennsylvania, lie within the contact-metasomic zone of Cambrian (500,000,000 to 570,000,000 years ago) limestone intruded by Triassic (190,000,000 to 225,000,000 years ago) igneous dikes (tabular intrusive body that cuts across host rock structures) and sills of quartz diabase. Associated with silicate minerals such as diopside, actinolite, and phlogopite, the iron ore consists mainly of magnetite that constitutes 40–60 percent of the ore. The ore body contains several tens of millions of tons of ore averaging slightly more than 40 percent iron by weight. The Lake Superior iron-ore deposits of hematite are residual ores formed in iron-rich beds of Precambrian metamorphosed sedimentary beds called banded iron formation (BIF). The iron originally was present within the beds as hematite and in the form of the iron silicate greenalite and the iron carbonate siderite. Weathering of the outcropping beds of iron formation and consequent alteration of the iron-bearing minerals to oxides has resulted in a concentration of hematite at and near the surface. These concentrations that form the ore deposits contain several thousand million tons of ore ranging in grade from 50 to 60 percent iron by weight. Similar iron-ore deposits in Western Australia have been formed by concentrations in Precambrian iron-bearing sedimentary beds. Several of these contain thousands of millions of tons of ore averaging more than 60 percent iron; the known cumulative reserves amount to approximately 20,000,000,000 tons of ore.

Fractures, fissures, and veins

Ore deposits have a variety of shapes. They may be formed as discrete veins and lenses, irregular masses, pipes or carrot-shaped bodies, networks of fractures, disseminations of minerals in crushed rock and sedimentary beds, or as sedimentary deposits themselves. Their shapes depend on their nature and genesis and on the stratigraphic and structural conditions that existed during their time of formation. Subsequent earth movements such as faulting and folding may displace or contort ore bodies, resulting in the development of zones of fracturing or crushing within the ore body in which further mineralization can take place. This sequence of formation of the minerals of an ore deposit is termed paragenesis.

A fissure vein (Figure 2A) is a fracture in the country rock that has been filled with minerals. Some of these minerals may be valuable, whereas others constitute

gangue. The vein material and the minimum quantity of adhering wall rock that must be mined with it constitutes ore. Mineralization of the space formed by a fracture may result in some cases from the forceful intrusion of magmatic differentiates; but most examples are hydrothermal in origin; that is, they have been formed by deposition of minerals from hot water and gaseous solutions moving through fractures in the country rock at depths of hundreds or thousands of feet. The force exerted by growing crystals, particularly by those aligned perpendicular to the walls of the fracture can force apart the walls of a fracture, thus creating more space for growth. The asbestos minerals chrysotile, crocidolite, and tremolite commonly form in this way.

Because the original fractures occupied by veins were formed by tension within the rock, they commonly show a relationship to the vectors of tension, or directions of forces, that caused them. This arrangement, known as a fracture pattern, indicates the orientation of these vectors and, consequently, the sense of displacement of the country rock by faulting or folding. Where there have been several periods of earth movements it may be difficult or impossible to unravel the sequence of fracture patterns. A common type of vein fills tension or gash-fractures, which may have an en echelon arrangment (essentially parallel and overlapping in plan view) due to the tension created by rotation of the rock mass in which they were formed. Other types of vein fillings are anastomosing (braided), subparallel, or fan-shaped, and commonly they are offset by numerous minor faults. The grade of ore within larger veins varies spatially. It may be higher near the footwall or hanging wall, and it may show variations with depth. Where the ore deposit consists of an interlacing network of veinlets less than an inch wide and a few feet long the entire rock mass is mined. Concentrations of mineralization within the veins that can be mined as richer ore are termed ore shoots.

From A.M. Bateman, *Economic Mineral Deposits* (1954); John Wiley & Sons, Inc.

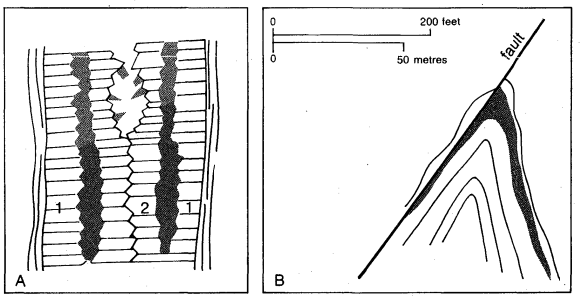


Figure 2: *Ore deposit shapes.*
(A) Fissure vein of (1) white quartz, sphalerite (shaded), and (2) amethyst formed in that order, illustrating paragenesis. Sphalerite crystals have formed in the vug (cavity) at top centre.
(B) Saddle reef of gold ore (shaded) within a faulted fold of quartzite and slate.

Because they may be conspicuous at the surface, vein deposits have been mined since ancient times, particularly for copper and gold. Other minerals found in vein deposits in important mines in various parts of the world include native silver, lead, zinc, antimony, nickel, cobalt, mercury, uranium, and tungsten minerals.

Lenticular, irregular, and pipelike bodies

All veins terminate and many pinch out to form irregular, sheetlike bodies that are markedly elongate but lenticular in cross section. For this reason a clear-cut distinction cannot always be made between veins and lenses of ore. In some circumstances the ore deposits may be formed in distinct lenses or in lenticular bodies of quartz, examples being the saddle reefs (Figure 2B) of Bendigo, Australia. These gold-bearing deposits form inverted V-shaped lenses in the domes of anticlinal folds of Lower Ordovician (about 500,000,000 years ago) quartzites and slates. The quartzite tends to fracture rather than bend with folding, and the saddle reefs containing the ore are commonly found at or near the intersection of a fault with the trend of the fold. Saddle reefs containing gold ore are also found in association with folded beds of slate in the Salmon River District of Nova Scotia.

In some cases ore deposits form irregular masses that do not show any clear-cut stratigraphic or structural relationships. Some of these are the result of secondary and tertiary periods of deformation and mineralization that have obscured the original pattern. Other irregular masses, including some deposits of the manganese oxide minerals hausmannite and psilomelane, and the aluminum oxide minerals diaspore, gibbsite, and boehmite, which are collectively known as bauxite, have formed by a process of lateritization, or long-continued weathering (*q.v.*) *in situ* near the surface (Figure 3). Such irregular mass-

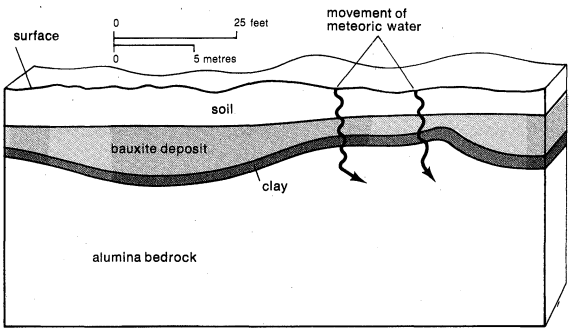


Figure 3: Relationship of bauxite deposit to alumina bedrock and overlying soil. Arrows indicate water percolation.

es may form disconnected pockets in the underlying bedrock, partly connected pockets, or irregular sheets commonly ranging in thickness up to 15 metres (50 feet). Lateritic deposits of the nickel silicate minerals garnierite and noumeite are also found.

Important lateritic deposits of manganese, aluminum, and nickel are mined in the following countries: manganese in the U.S.S.R., India, Brazil, South Africa, Morocco, China, Australia, and Ghana; aluminum in Australia, Guinea, and Jamaica; and nickel in New Caledonia and Cuba.

Pipes are irregular, cylindrical or funnel-shaped ore bodies that commonly extend vertically, or nearly so, to depths of hundreds or thousands of feet. They have been formed either as volcanic necks (central conduit of a former volcano) or as crushed zones bordering the intersection of the planes of two steeply dipping faults, fissures, or dikes. The diamond pipes of South Africa are carrot-shaped volcanic necks, up to 700 metres (2,300 feet) deep, tending to be oval near the surface but becoming narrower at depth. They consist of fragmental and weathered basic igneous rock termed kimberlite in which the diamonds are disseminated. Mining was carried out initially by open-pit operations and subsequently by underground methods. Other pipes, sometimes referred to as chimneys, are formed by various combinations of structure and stress that cause disruption of the country rock, by upward stoping (removal of rock) caused by solutions, or by the washing of rock debris into limestone caves. Consisting of brecciated rock, or rubble, these pipes have been channels for mineralizing solutions. Ore-bearing pipes are commonly less than 30 metres (100 feet) in diameter but may be several hundred feet in length.

Ore bodies formed by mineralizing solutions also can consist of disseminations of minerals in crushed or finely fractured rock and in porous limestone and sandstone. Porphyry copper deposits are in the former category. These consist of disseminated copper minerals in the crushed or cracked periphery of intrusive igneous bodies of granitic rocks such as quartz monozonite and diorite porphyry. Mineralization also extends into the intruded and disrupted country rock. These porphyry copper bodies, which commonly contain valuable amounts of molybdenum sulfide minerals, are large, low-grade deposits that are usually mined by giant open-pit operations. Irregular in shape, but tending to be tabular, porphyry copper deposits may have dimensions ranging up to several thousand feet horizontally and several hundred feet vertically. Large porphyry copper deposits are mined in the United States, the Solomon Islands, and Canada.

Ore bodies in sedimentary rocks

Ore bodies are also formed as irregularly shaped mineralized masses that have developed in zones of higher permeability through which mineralizing solutions have permeated beds of limestone or sandstone. In the U.S., Missouri, Oklahoma, and Kansas have notably large deposits of the zinc sulfide mineral sphalerite, associated with galena, a lead sulfide, in Cambrian dolomite. The ore deposits occur as patchy replacements in silicified zones of the dolomite; they are commonly tabular bodies 3 to 4.5 metres (10 to 15 feet) thick and several hundred feet wide.

Lead–zinc deposits of similar shape are found in Triassic dolomites of the Silesian region of Czechoslovakia and Poland. An irregular and tabular body of lead–zinc ore is also mined from a Devonian (345,000,000 to 395,000,000 years ago) limestone reef at Pine Point in northern Canada.

Ore deposits also are found in sandstone beds where they have formed by precipitation of ore minerals in the interstices between the grains of sandstone. The loci of deposition have been determined by variations in permeability of the sandstone, the mineralizing solutions moving more freely through certain portions of the bed and possibly in some cases by selective precipitation caused by the catalytic action of particular variations in the sedimentary content of the sandstone. As these factors are stratigraphic, they result in the formation of fairly regular, usually elongate and tabular bodies of ore lying parallel to the bedding.

In Zambia the disseminated copper sulfide mineralization, chiefly chalcocite, chalcopyrite, and bornite, within Lower Roan Group feldspathic sandstones, is characterized by uniformity over considerable distances, the ore deposits following the configuration of the folded sandstone beds. Other examples of ore deposits formed by mineralization of particular portions of a sandstone bed are the uranium—vanadium deposits in the Jurassic (136,000,000 to 190,000,000 years ago) Morrison Formation and the Triassic Chinle Formation and Shinarump Formation of the Colorado Plateau in the United States. The deposits were selectively developed within fluvial channels, forming sinuous belts of elongate ore bodies. Commonly associated with carbonized fossil wood fragments and plant material, the minerals form a complex association that includes the uranium minerals uraninite and brannerite and the vanadium minerals roscoelite and montroseite. The ore bodies are elongate and tabular, or podlike, commonly ranging in thickness up to 6 metres (20 feet) and in length up to 304 metres (1,000 feet). They are more or less parallel to the bedding but locally discordant.

Some very large ore deposits consist of minute particles of minerals disseminated in layers parallel to the bedding planes of a sedimentary rock. A few examples are the lead–zinc deposits containing the sulfide minerals galena and sphalerite disseminated in argillaceous sedimentary rocks, in the Gulf of Carpentaria and Mt. Isa areas of Australia; the Kupferschiefer copper deposits of Germany, which consist of black, bituminous shale that mainly contains finely divided particles of the copper sulfide minerals bornite, chalcocite, and chalcopyrite but also contains galena and sphalerite; and the cherty banded iron formations called taconite, itabirite, or jaspilite, consisting of disseminated hematite, magnetite, siderite, or greenalite, found in many parts of the world, including the Lake Superior region. These deposits (BIF) are tabular to sheetlike in shape, usually only a few feet thick, and in places may be traced along the layers of sediments for several miles.

Little is known about the precise depths at which ore deposits are formed or the maximum ranges of depth. Diamonds probably were formed as crystals in a basic kimberlite magma that was extruded upward, possibly from a depth of several miles, into volcanic pipes that form the ore deposits today. Sulfide minerals are formed within a depth range of several hundred to many thousands of feet. The limit of depth probably varies geographically and certainly depends on the limiting depth of fracturing and circulation of water vapour. Many ore deposits probably form within a temperature range of 300°–500° C (570°–930° F) and at depths of up to 16 kilometres (10 miles).

Geochemistry and mineralogy. The geochemistry of an ore deposit includes the physicochemical reactions involved in the formation of individual minerals, and mineralogy is the study of those minerals and their physical and crystallographic properties. Studies of the genesis of ore deposits, of the relationships of ore deposits to the depths and temperatures at which they formed, and to the petrology of country rock, therefore involve geochemistry and mineralogy. Many variations of physicochemical processes can be involved in the genesis of an ore deposit, which may involve several episodes of mineralization. A few examples are discussed here.

Physicochemical processes of importance

The Kiruna iron ore deposits in Sweden consist of magnetite and fluorapatite that appear to have first formed as a magmatic differentiate in the form of an immiscible magnetite–apatite liquid melt. Forceful intrusion of this melt into fractures in the country rock resulted in the formation of the ore deposits. The geochemistry of this deposit is clearly related to the physicochemical properties of solutions in molten rocks and the minerals that can crystallize from them under certain conditions of temperature and pressure.

The hot volatile gases, including water vapour, given off by molten magma intruding the country rock are important mineralizing agents. Permeating the country rock through fissures and brecciated zones above the intruding magma, these aqueous and gaseous solutions deposit minerals as they become cooler. There are many examples of ore deposits formed as veins, stockworks, or disseminations by high-temperature (hypothermal), medium-temperature (mesothermal), and low-temperature (epithermal) solutions.

Residual ore deposits of the manganese oxides psilomelane and hausmannite, and of the aluminum oxides collectively referred to as bauxite, are formed by chemical processes in the weathering *in situ* of soil, subsoil, and underlying alumina-rich or manganiferous rock. These processes are not thoroughly understood but are known to be the result of meteoric water (water of atmospheric origin) penetrating downward, removing silica and leaving a concentration of manganese or aluminum oxides. The drainage of meteoric waters through the soil and into streams and rivers results in the solution and transportation of minute amounts of elements such as copper, lead, and zinc that are in the surficial material derived from the weathering of underlying mineralized rocks; these elements can be detected by chemical or spectrometer analyses. In the search for ore deposits, particularly those that are hidden by soil and vegetation, geochemical surveys that collect and analyze soil and water samples are among the principal methods employed.

Geochemistry also is a factor in the formation of sedimentary ore deposits, referred to as syngenetic deposits. Some of these may have formed originally as mechanical concentrations of mineral particles within the sediments. There is evidence, for example, that the gold in the conglomerate beds of Witwatersrand in South Africa was emplaced as a placer deposit, although it subsequently has been partly redistributed. Similar evidence can be shown for the origin of the uranium in conglomerate beds of the Blind River area in Canada. But other deposits, with the possible inclusion of the lead–zinc mineralization in the Gulf of Carpentaria and Mt. Isa areas of Australia, appear to have formed by the contemporaneous deposition of metallic sulfides with the sediments. Deposition of lead and zinc sulfides in mud are known to be taking place in certain low areas of the Red Sea floor where water of higher salinity and density is concentrated. The physicochemistry and possible biochemistry involved in such processes is under study, and the stratigraphic implications, insofar as exploration for ore deposits is concerned, are far-reaching.

Temperatures and pressures of formation. Ore deposits have been formed under great ranges of temperature and pressure, depending on depth. Magmatic segregation deposits resulting from the differentiation within a molten

magma of immiscible melts, such as that believed to have formed the magnetite–apatite deposits at Kiruna in Sweden, have originated under conditions of very high temperature and pressure. Another example of an ore magma is a sill-like body of magnetite and hematite with minor amounts of apatite at Lago Sur in northern Chile. These metallic facies (aspects) of magmas are not common at the eroded surface of the earth's crust but may be more abundant at depth. It is not known whether they also include metallic sulfides or are confined to the oxides.

Allied to deposits formed by ore magmas are the tabular to lenticular chromite deposits of Cuba, which are thought to have originated as a crystal mush of ore magma and chromite crystals injected into the country rock. Other chromite deposits that seem to have had a similar origin are those of South Africa (Figure 1B) and Turkey.

The temperature ranges under which ore magmas are emplaced exceed 500° C, in many cases probably by several hundred degrees. The pressure ranges are in the order of several thousand pounds per square inch. Temperature and pressure ranges obtained during the formation of ore deposits of hydrothermal origin are generally lower.

Hydrostatic gradients and mineral precipitation

T.S. Lovering, U.S. economic geologist, says

> Of many physical factors that may induce precipitation, changes in temperature and pressure seem the only ones likely to be important during ore deposition. It is generally assumed that pressure increases directly with depth, and rock pressure at any specified depth below the surface is usually calculated as equal to the weight of the column of rock above it. Similarly, it has been assumed that the pressure of a mineralizing solution at any point below the surface corresponds to the hydrostatic head existing at that depth. It should be noted that pressure on a small area is transmitted in an outward flaring cone of diminishing intensity and it is thus possible for pressure to build up in a localized area that is greatly in excess of the rock pressure calculated for that depth. Although the pressure gradient in an open channel filled with liquid is approximately the hydrostatic head of any given depth, this is not true of a hypogene solution passing through a conduit that has marked constrictions between the point of entrance and the exit of the fluid. At any place the pressure is that of the hydrostatic head plus the driving force required to overcome friction. Below a marked constriction the pressure may be very much in excess of that above it, resulting in a rapid change of pressure. In any conduit the steepest pressure gradient would be at the greatest constriction. The places of greatest constriction may change from time to time because of mineral deposition, mineralization stoping, or intramineralization movements, and the sudden access to a new channel may move the position of the steepest pressure gradient either toward or away from the source of the solution.

From the foregoing it is evident that the pressure–temperature relationships that obtain during the formation of an ore deposit can be complex. Geothermal gradients vary from place to place and may be in excess of 1° C (2° F) per 100 feet of depth. Hydrostatic gradients depend on the density, or salinity, of the underground water and are commonly in the range 40–45 pounds per square inch per 100 feet (3 kilograms per square centimetre per 30 metres) of depth. Geostatic or rock pressures are two to three times the hydrostatic pressures at the same depth.

Paragenesis of ore minerals. Paragenesis means the sequential order in which various minerals within an ore deposit are formed. Variations in temperature, pressure, and chemical constituents of a mineralizing solution will result in the precipitation of different minerals at different times within the same deposit. Further complications in paragenesis arise where the ore deposit has been formed by more than one period of hydrothermal activity. A simple example of paragenesis can be seen where minerals line the walls of a fissure or cavity (Figure 2A). The earliest formed minerals grow inward from the wall rock forming an outer layer, and later formed minerals grow as inner layers or rings filling the void. More complex relationships are seen by examination of polished sections of ore under the microscope, which shows the manner and sequence in which crystallization and replacement of one ore mineral by another have occurred.

Studies of many hydrothermal ore deposits throughout the world have established a general sequence of mineral deposition based on mineral stability ranges. Commonly, the first minerals formed are silicates and carbonates, which constitute gangue. Of these, quartz and calcite, both of which are common in ore deposits, continue to be deposited during the later stages of development. The minerals chlorite, sericite, tourmaline, albite, adularia, barite, siderite, fluorite, ankerite, and rhodochrosite are also formed at an early stage. Secondly, the oxide minerals magnetite, ilmenite, chromite, specularite, and uraninite are formed. Thirdly, the minerals pyrite, arsenopyrite, cassiterite, tantalite, wolframite, molybdenite, pyrrhotite, pentlandite, and the nickel and cobalt arsenides are crystallized. Fourthly, the minerals chalcocite, bornite, chalcopyrite, sphalerite, galena, native silver, gold, the tellurides, stibnite and cinnabar are precipitated. This arrangement shows that in general the oxides are deposited early; that the sulfides and arsenides of iron, nickel, cobalt, and molybdenum are contemporaneous with or slightly younger than the oxides; and that the sulfides of lead and zinc are even younger. Later, the native metals and tellurides are formed, and the last to be deposited are antimony and mercury sulfides.

Sequence of mineral deposition

Related to paragenesis is the zoning of ore deposits (Figure 4). As mineralizing solutions move through channels in the rock they undergo changes in temperature, pressure, and chemical composition that may result in the deposition of different mineral concentrations at increasing distances along the course of deposition, and presumably from the magmatic source. This zoning is common but is not always present in ore deposits. In general, minerals of tin, tungsten, and bismuth lie closer to the magmatic sources of the mineralizing solutions than do those of the copper minerals. Lead and zinc minerals are more distant; native gold and silver, and minerals of gold and silver, are still farther away; and the most distant of all from their source are the minerals of antimony and mercury.

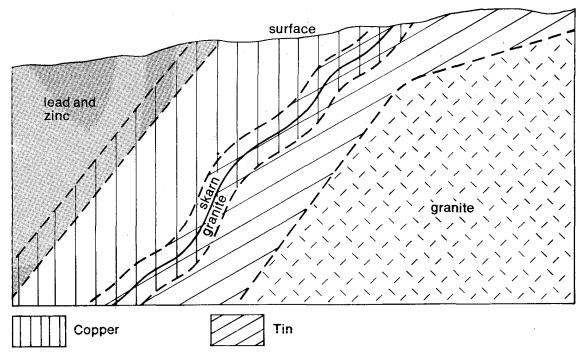


Figure 4: Relationship of mineralized zones to the granite from which the mineralizing solutions are believed to have emanated.

Experimental models. Some of the conditions under which minerals are formed can be simulated in the laboratory. Very high temperatures can be attained in electric furnaces and very high pressures in pressure bombs. But time, a factor in forming ore deposits that is measured in hundreds or thousands of years, cannot be duplicated. The physicochemical conditions required to form certain minerals have been determined, with some limitations, by means of synthesis. Much of the experimental work on synthesis of minerals has been carried out with dry components, under conditions that do not exist in nature. One method is to heat certain components in a closed container, under constant pressure. The lowest temperature at which a particular mineral will crystallize at a constant which minerals are formed can be simulated in the labo-pressure is taken as a control point. A series of control points determined at various pressures can then be shown graphically. These and similar approaches to geothermometry are limited and equivocal in their application to the understanding of the formation of ore deposits.

Geothermometry and mineral inclusions

The determination of melting points for magnetite and chromite may give an approximation of the temperature

at which certain magmatic ores were formed. Certain inversion points (conditions at phase changes), such as that of the silver sulfide system, are reliable indicators of the temperatures at which ore deposits were formed. Silver sulfide, for example, changes at about 180° C (365° F) from the isometric (possessing three equal and mutually perpendicular crystal axes) crystal form known as the mineral argentite, which is stable at high temperatures, to the monoclinic (one crystal axis is inclined) mineral acanthite, which is stable at low temperatures.

Some minerals formed at high temperatures contain impurities, which, as the mineral cools, will crystallize out as laminae and blebs (small, irregular masses) of other minerals within the original crystal. This process, known as exsolution, takes place at certain temperatures for each mineral; the resulting relationships can be observed in polished sections or thin sections under the microscope and are also useful indicators of the temperature range under which an ore deposit was formed.

Another method of determining the temperature of ore formation involves the study of fluid inclusions in vacuoles (enclosed cavities) within a crystallized mineral. It is inferred that at the time of growth of the crystal the vacuoles were filled with gas. At room temperature the vacuoles are filled with liquid and gas. Where the inclusion is more than half liquid, it is inferred that the mineralizing solution forming the ore was hydrothermal (a high-temperature liquid solution); and where it is mainly gas, that the solution was pneumatolytic (gaseous). On heating the crystal the fluid within a vacuole will disappear as a single gaseous phase is reached at a certain temperature. This is considered to be the minimum temperature at which the mineral could have formed.

These and other experimental methods have been applied to particular ore deposits with fairly consistent results and may be useful in solving some problems of hydrothermal versus syngenetic origin.

Genetic and other classifications. To be useful, a classification of ore deposits must be simple and usable in the field. Unfortunately, ore deposits cannot always be clearly defined by type. In many cases they have been formed by complex physicochemical processes, and in some cases their genesis is still in question. Classifications can be derived from the agents that form ore deposits, such as magmas, hydrothermal solutions, pneumatolytic solutions, syngenetic processes such as precipitation in a body of water, weathering, and mechanical processes. Other classifications can be built on temperature and depth relationships, on stratigraphic and structural relationships, and on mineralogical and chemical relationships. Each classification may satisfy certain categories of ore deposits, but any comprehensive classification of all ore deposits must of necessity be general and, consequently, of limited application.

Bases and limitations of classifications

There are three broad classifications of ore deposits in use. The Niggli classification has two main divisions, namely plutonic, or intrusive, and volcanic, or extrusive. The former division has three categories: orthomagmatic, pneumatolytic to pegmatitic, and hydrothermal. The Schneiderhöhn classification has four main divisions: intrusive and liquid-magmatic, pneumatolytic, hydrothermal, and exhalation. The Lindgren classification has two primary divisions: deposits produced by chemical processes and deposits produced by mechanical processes. The first division contains three categories: ore deposits formed by differentiation in magmas, ore deposits formed in bodies of rock, and ore deposits formed in bodies of water.

Whatever classification is used, the critical point is that it must serve a purpose. Each classification has its limitations and is subject to modifications.

EARTH PROCESSES INVOLVED IN THE FORMATION AND DISTRIBUTION OF ORE DEPOSITS

Magmatic concentration. Magma is molten rock at depth; when it rises to the surface through fissures and volcanic vents it is called lava. Lava cools quickly and is consequently fine grained to glassy, or frothy like pumice. Magma cools slowly and is medium to coarse grained. As the temperature of a magma is lowered the melting points of various minerals that can exist in equilibrium with the melt are reached in a known sequence. In a basic magma, namely, one high in ferromagnesian constituents, the accessory minerals apatite, magnetite, ilmenite, chromite, titanite, rutile, and zircon are the first to crystallize, followed closely by essential silicates such as olivine and the orthorhombic pyroxenes. Other silicate minerals such as the clinopyroxenes (those crystallizing in the monoclinic crystal system) and plagioclase crystallize later.

Where cooling proceeds very slowly the earliest crystals to form, particularly magnetite, chromite, and apatite, are able to settle by gravitation through the magma and so become concentrated. Stresses resulting in the flow of magma may cause further concentration by squeezing off the liquid fraction from the crystalline mush, a process known as filter-pressing.

Concentrations of magnetite and chromite formed as magmatic concentrations are typically within sill-like, dikelike, or lenticular bodies of basic igneous rock. But whether they originated as the result of differential crystallization and concentration of the early formed crystals of magnetite and chromite by means of gravitational settling or filter pressing, or whether they originated as the result of magmatic differentiation is not always clearly indicated by the ore deposits.

From A.M. Bateman, *Economic Mineral Deposits* (1954); John Wiley & Sons, Inc.

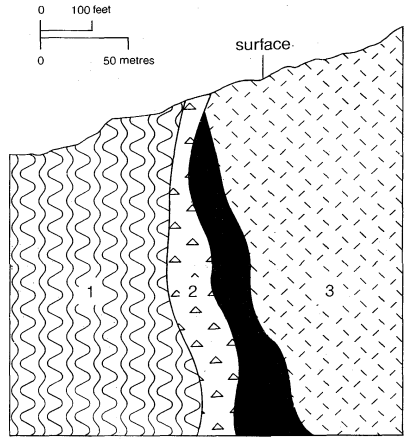


Figure 5: Hydrothermal deposit of massive pyritic copper ore (solid black) forming a lens in (3) sheared quartz porphyry that intrudes (1) slate and is in turn cut by (2) a dolerite dike, Río Tinto, Spain.

Differentiation within the magma itself can result in the separation of distinct melts of the oxides magnetite and chromite with apatite and olivine. The immiscibility of a magnetite–apatite fluid within a silicate melt has been demonstrated in the laboratory, and field investigations have led to the belief that the same segregation can take place in nature. What is not known is the extent to which magmatic differentiation can occur with reference to sulfide and silicate melts. Magmatic segregation processes and hydrothermal activity are to some extent overlapping, and this is of importance in the interpretation of certain sulfide deposits. Massive sulfide or arsenide bodies, such as those of the Río Tinto District, Spain, and the Sudbury District, Ontario, were for many years considered to be direct products of magmatic segregation. With regard to Sudbury, the original hypothesis was discarded in favour of a hydrothermal theory, because the ores did not seem to be simple fractions of magmatic differentiation. Studies of ore localization indicated that the sulfide ores were related to fractures and other structural features. Later investigations led to a compromise. The essential nature of the problem is that the ore deposited by magmatic segregation may subsequently be remobilized during metamorphism and igneous intrusions. Hence, the deposits actually have a dual origin. This may also be true of the copper deposits at Río Tinto (Figure 5), which may have in part originated as sulfide melts.

Magmatic differentiation and segregation

Contact metasomatism. Metasomatism means replacement, and commonly refers to the replacement of a volume of rock by an aggregate of minerals (Figure 6). This replacement can be on a microscopic or megascopic scale and may preserve the external shape and internal features of the original rock. Selective replacement of certain layers and laminae within a rock can occur, and some examples have given rise to controversy over the criteria by which such selective replacement and syngenetic deposition can be distinguished.

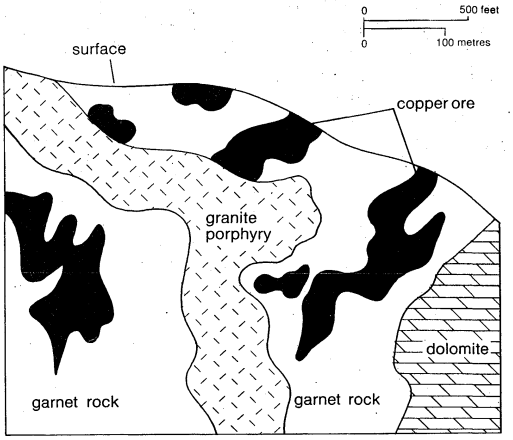


Figure 6: Contact-metasomatic deposit of copper ore (solid black) in garnet rock that has been formed as an alteration product of dolomite by the intrusion of granite porphyry.

The means by which such replacement takes place are not well understood but are thought to be caused by the diffusion of hydrothermal and pneumatolitic solutions that selectively dissolve the country rock and deposit an aggregate of minerals in its place. This mineralization can consist of gangue minerals only, such as carbonates and silicates, or of ore minerals associated with the gangue. The chemical reactions upon which deposition depends are in part controlled by the chemical composition of the rock replaced, in part by the chemical composition of the mineralizing fluids or gases, and in part by temperatures and pressure. When the temperature exceeds the critical temperature of water, the mineralizing solution is gaseous, regardless of pressure, and highly penetrating. Rocks that have very low permeability to water at surface temperatures are capable of absorbing water vapour like a sponge at high temperatures. The vapour penetrates the rock through intergranular interstices, minute fractures, and other openings that allow it to replace and mineralize the rock by diffusion.

The problem of volume-for-volume replacement

One problem that has always intrigued geologists is the question of how a large volume of rock can be removed by the same solutions that deposit the ore body. This replacement is generally volume for volume, although the replacement of one mineral by another usually results in a change of crystal volume. Where limestone is dolomitized by magnesia-rich solutions, for example, the volume of rock altered may remain the same, although replacement of the mineral calcite by the mineral dolomite, involving replacement of calcium by magnesium in the crystal structure of the carbonate mineral calcite, causes a shrinkage of crystal volume. This shrinkage is taken up within the volume of rock itself as an increase in pore or intergranular space.

The term contact metasomatism implies replacement of the country rock in contact with an intrusive body of magma. The mineralogic changes and the distance at which these changes are effected from the intrusive body depend on many factors, including the chemical composition, water-vapour content, and temperature of the magma, and the chemical composition and permeability of the intruded country rock. Intrusion of a quartz sandstone will result only in silicification, by infilling of intergranular space with quartz or recrystallization of the quartz grains, to a quartzite. Intrusion of a pure limestone may result only in recrystallization of the calcium carbonate to marble or in the diffusion of iron and magnesia-rich solutions that will react with the calcium carbonate to form an assemblage of calcium–magnesium silicates and oxides, including actinolite, tremolite, wollastonite, epidote, diopside, scapolite, grossularite, garnet, and magnetite. Some important iron-ore deposits of magnetite and hematite—such as those at Cornwall, Pennsylvania; Iron Mountain, New Mexico; and Iron Springs, Utah—were formed by the replacement of limestone intruded by basic igneous rocks.

Ore deposits of the platinum arsenide mineral sperrylite, associated with the copper and nickel sulfide minerals chalcopyrite and pentlandite, have been formed by contact metasomatism where norite has intruded the carbonate rock dolomite and banded ironstone, in the Bushveld Complex of South Africa.

The tungsten mineral scheelite is associated with grossularite garnet and epidote in contact metasomatic deposits formed by the replacement of limestone beds intruded by granodiorite at Mill City, Nevada. One of the largest deposits of scheelite in the world, on King Island, Australia, is also of contact-metasomatic origin and was formed by the replacement of dolomite, intruded by granite, with a skarn containing garnet, pyroxene, and scheelite.

It is not always possible to differentiate clearly between contact metasomatism and subsequent hydrothermal alteration. Some metalliferous deposits, including certain of the copper sulfide deposits at Bisbee, Arizona, and a deposit of the zinc sulfide mineral sphalerite at Hanover, New Mexico, are of contact-metasomatic origin. Other deposits of copper, lead, and zinc sulfides form ore bodies in skarn (a garnetiferous silicate rock) as the result of secondary mineralization and partial replacement of the skarn. These ore bodies may be the result of subsequent hydrothermal and pneumatolytic mineralization rather than primary contact metasomatism.

Hydrothermal processes. Hydrothermal processes are the physicochemical means by which hot aqueous solutions effect alterations and replacement of rock, such as sericitization (a form of feldspar alteration), silicification, and sulfide mineralization (Figures 7, 8), along

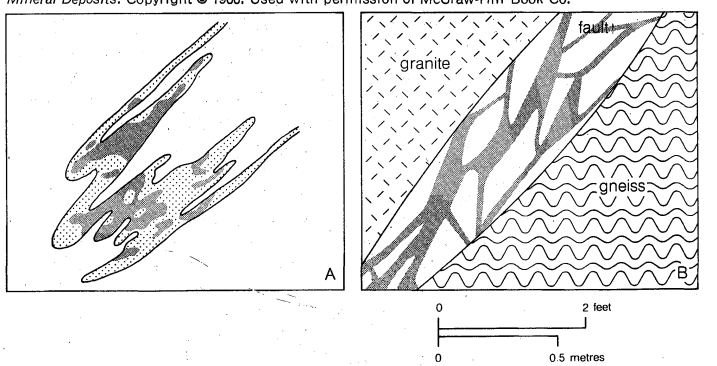


Figure 7: *Hydrothermal processes.*
(A) Plan view on 3,050-foot level of hydrothermal deposits of gold ore (shaded) in folded beds (stippled), Homestake Mine, South Dakota. (B) Hydrothermal deposit of zinc ore (shaded) forming a vein in crushed granite along a fault between gneiss and granite, Georgetown, Colorado.

the channels and zones through which the solutions move. Hydrothermal solutions have been classified as epithermal with an upper limit of 200° C, mesothermal with a limit of 300° C, and hypothermal in the range of 300°–500° C. These limits are arbitrary, and there is no clear distinction between the hypothermal range of hydrothermal solutions and pneumatolytic solutions.

Hot-water brines

Hot-water solutions, particularly heavy brines, are known to carry in solution exceptionally high concentrations of metals. Samples of sediments taken from highly saline depressions in the Red Sea floor contain metallic sulfides, and anomalous concentrations of heavy metals, produced at a depth of over 1,520 metres (5,000 feet) in

the Salton Sea area, California, represent a similar case. The brine, at a temperature exceeding 300° C (570° F), contains copper, silver, potassium, lithium, antimony, lead, arsenic, boron, beryllium, bismuth, gallium, and gold. On cooling through a discharge pipe, a dark, siliceous sludge was formed that contained 20 percent copper and 2 percent silver. During a period of three months an estimated five to eight tons of sludge was formed. Such a solution moving upward through fractures serving as conduits or pipes would deposit copper, silver, and other minerals.

Pure water is not a good solvent. With increasing alkalinity, commonly resulting from higher contents of sodium and potassium carbonates, the ability of a hot-water solution to dissolve quartz and other constituents in the rock is greatly increased, particularly under high pressure and temperature. The source of such hydrothermal solutions is not always easy to define, but probably all were originally meteoric waters that seeped down through crevices into hot parts of the earth's crust, to be subsequently forced upward by tectonic (vertical motions of the earth's crust) and magmatic forces, or else were connate (original or primeval) waters trapped with the original sediments and subsequently mobilized during metamorphism of the sedimentary rocks.

Of particular significance to ideas concerning the movement of hydrothermal solutions at depth is the realization in recent years, resulting from experimental work and field observations, that fluids under high pressure can fracture the enclosing rock and so inject themselves into zones of lower pressure. This process may account for some mineralized veins formed in fractures.

Sericitization and silicification are common processes in wall rock alteration associated with mineral deposits. Hydrothermal solutions permeate the adjacent country rock, altering or replacing it by removing certain constituents and depositing the minerals sericite, chlorite, and quartz. Sericite is more common in mesothermal deposits; the more abundant chlorite has a greater temperature range

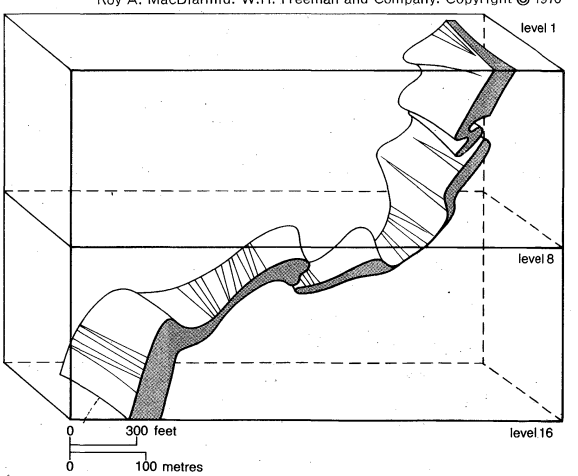


Figure 8: Hydrothermal deposit of massive lead–zinc–copper ore forming connected lenses along a pipelike body of aplitic rock intrusive into dolomite. Levels shown indicate principal zones of mining.

but is particularly common in epithermal deposits. Silicification of carbonate rocks and shales associated with ore deposits can form jasperoid, a cryptocrystalline silica that has the appearance of reddish chert, or a hard, fine-grained brittle rock.

Sublimation and evaporation. Sublimation is the volatilization of a solid by means of heat. On cooling, the vapour is deposited as an amorphous solid or as crystals.

Fumarole deposition

Sublimation of sulfur occurs in the fumaroles of volcanoes. Many nonvolatile substances can be dissolved and transferred by water vapour at high temperature, and it is not always possible to define clearly mineralization effected by sublimation and that effected by pneumatolytic or gaseous action. For example, incrustation of minerals in the fumaroles of the Valley of Ten Thousand Smokes, Alaska, contain magnetite in the early phase and galena and sphalerite in the later stages. Fluorides, borates, and other sulfides also are present, having been deposited by gaseous exhalations from hot, rhyolitic pyroclastics (rocks consisting of volcanic ejecta). These exhalations contained a high percentage of superheated steam. Other solid hydrocarbons, including gilsonite, wurtzilite, and grahamite are found as veins, the origin of which is now known. They may have been formed by some pneumatolytic process, the hydrocarbons being derived from a deep-seated source. Large veins of gilsonite are mined in Utah. The products include high-quality carbon used as anodes in the electrolytic production of aluminum ingots, gasoline, and other petroleum products used in a variety of industrial processes. Certain nonvolatile hydrocarbons, insoluble in organic solvents, and known as pyrobitumens because of their possible origin as sublimates, are found as disseminations in cavities and fissures in sedimentary rocks.

Evaporite deposition

Evaporation has been one important process in the formation of certain mineral deposits such as beds of calcium, sodium, magnesium, and potassium salts; but it is not a process that, by definition, forms ore deposits. A salt bed has no appreciable gangue, whereas an ore consists of valuable minerals that must be separated from the nonvaluable gangue. Salt beds, including common salt known as the mineral halite, are either mined or exploited by injecting hot water through bore holes and pumping up the resulting brine. Important mineral deposits of salts, including gypsum, halite, carnallite, kieserite, and sylvite, are found in many parts of the world.

Salt beds, referred to as evaporites are bedded deposits. Subsequent deformation of the enclosing sedimentary beds and solution and redeposition of the salts have in many places formed irregular or dome-shaped bodies of salt. The genesis of salt beds apparently requires conditions in which a shallow and restricted sea is subject to rapid evaporation together with continual replenishment of seawater. Deepwater origin is also possible (see EVAPORITES). Salt beds are commonly associated with primary dolomite formed by magnesia-rich seawater altering the calcium carbonate mineral aragonite, precipitated as carbonate ooze, to the calcium-magnesium carbonate mineral dolomite.

Precipitation. Precipitation means the deposition of a substance or mineral from a solution. This process has taken place in the formation of many hydrothermal deposits, particularly in the zones of secondary enrichment, and also in the formation of certain deposits on the sea floor, such as concentrations of phosphatic and manganiferous nodules. Precipitation of the sulfides of base metals is known to occur in localities on the floor of the Red Sea, where the salinity is abnormally high; and syngenetic deposits in sedimentary beds, believed to have possibly formed in this way, have been described in many parts of the world.

Precipitation in seawater

Precipitation does not imply crystal growth. The assemblage of minerals in many deposits has formed as crystals growing in a mineralizing solution. Precipitation does imply the deposition of minute crystals or colloidal particles. This process can involve rapid changes of temperature. Where the sea over shallow carbonate banks and reefs is warmed by the sun, for example, white patches caused by the sudden precipitation of minute crystals of aragonite have been observed to form suddenly. Warming of the water causes the release of dissolved carbon dioxide, which in turn lowers the solubility of calcium carbonate in seawater. As the saturation point is reached, precipitation takes place.

A well-known reaction is the precipitation of colloids on mixing with salt water. This may take place where river water mingles with the sea, or where there occur changes in salinity and temperature of the sea itself. Precipitation of colloids is a common reaction in the supergene zone of secondary enrichment and is indicated by the development of radiating and colloform structures, and rare dessication cracks in nodules and encrusting masses of minerals. Colloform structure has been observed in bauxite,

psilomelane, garnierite, chrysocolla, malachite, pitchblende, chalcedonic quartz and opal, marcasite, cassiterite, and other minerals. The precipitated colloidal particles coagulate to form a gel that subsequently hardens by loss of water. Reduction in volume may take place with the development of dessication cracks; but as these are not always apparent it is postulated that where hardening of a layer takes place before deposition of the overlying layer, no reduction in volume occurs. Among the classic examples of colloidal origin for ore deposits are the lead–zinc deposits of Upper Silesia in Poland, where the galena and sphalerite occur in dolomitized limestone; and the sulfide ores on the islands of Honshu and Hokkaido, Japan, which include chalcopyrite, sphalerite, galena, martite, and tetrahedrite in brecciated and altered volcanic and sedimentary rocks.

Sedimentation and mechanical concentration. Bedded mineral deposits of salts are products of sedimentation, as are diatomite, calcium-rich limestone, and some beds of phosphatic rock. These are mineral deposits, but they are distinguished from ore deposits on the basis of their lack of gangue. Diatomite is formed by the accumulation of siliceous skeletons of diatoms that rain down through the sea. Limestone and phosphatic rock are formed by depositional processes that include the accumulation of particles of sediment and the precipitation of ooze.

In some cases, precipitation of minerals has resulted in the formation of deposits that cannot be defined clearly as mineral deposits or ore deposits, such as the iron and manganese oxides formed in beds and restricted lakes. These have been formed by the precipitation of hydroxides, subsequently converted to oxides and carbonates of iron and manganese. A classic example, among the large manganese deposits of the world, is the Chiaturi Deposit of the Caucasus Mountains, U.S.S.R. This consists of oolitic pyrolusite and psilomelane, with the iron oxide mineral braunite, in a bed of marly sand of Oligocene age (about 26,000,000 to 38,000,000 years old).

Other syngenetic accumulations of minerals in sediments, such as those formed by the precipitation of sulfides of base metals, are defined as ore deposits because they require separation from the rock material. In many cases, later fracturing of the syngenetic deposit has resulted in partial solution of the sulfides and deposition within the fractures as secondary mineralization, which clearly places the deposit in the category of an ore deposit. In some syngenetic deposits—for example in the Gulf of Carpentaria area of Australia, where there are very large deposits of disseminated galena and sphalerite in metamorphosed sedimentary rocks—the sulfide particles are so fine that separation from the rock has proved extremely difficult in the milling process.

Placer deposits

Mechanical processes of sedimentation are also important agents in forming some types of ore deposits. Placer deposits of gold, the platinum metals, diamonds and other gemstones, cassiterite, rutile, and zircon are mined in many parts of the world. Some of these deposits are in modern or ancient river gravels, others are in beach sands. They have been formed by the mechanical action of moving sand and gravel transporting the mineral particles of higher density and winnowing them into concentrations.

Important placer gold deposits have been found worldwide, but those of platinum are mainly in the Ural Mountains, U.S.S.R. Large alluvial diamond deposits have been mined in Sierra Leone, Ghana, South Africa, South West Africa, and Zaire, and alluvial deposits of cassiterite in Malaysia. Among the large alluvial deposits of rutile and zircon are those found in the beach sands of Queensland, New South Wales, and Western Australia.

Residual deposits. Residual ore deposits are formed as concentrations of minerals at or near the surface by chemical decomposition and mechanical disintegration of the country rock or ore deposit in which they are contained. This definition includes laterite deposits, which, strictly speaking, are products formed from residue. Lateritic material and residue, whether formed by chemical or mechanical processes, are essentially concentrations formed *in situ*. Chemical decomposition of alumina-rich country rock, produced by weathering in a warm, humid environment where vegetation cover prevents mechanical disintegration and erosion, results in the formation of hydrated aluminum oxides known collectively as bauxite. The chemical reactions are not well understood, but they involve the removal of silica by downward percolating meteoric water. Underlying the bauxite deposits, and separating them from the decomposed bedrock, there is invariably present a layer of clay. It is probable that chemical decomposition of the rock forms clay that is subsequently converted to bauxite. Among the large bauxite deposits of the world are those of Australia, Guyana, and Jamaica.

Formation of bauxite and other deposits by weathering processes

Large residual deposits of manganese oxides formed by the chemical decomposition of manganiferous rocks are mined in India, Ghana, and Morocco. The ores of this type in India and Ghana are formed from the decomposition of metamorphic rocks called schists, which contain spessartite garnet, and other manganiferous silicates. The Moroccan deposits have been formed by decomposition of the sedimentary rock dolomite, which contains manganese carbonates. Other important residual ore deposits are those of nickel silicates, including garnierite, in New Caledonia, the Philippines, and Cuba. These have been derived from the weathering of nickeliferous serpentine, which is a soapstone type of rock formed as an alteration product of ultramafic igneous rocks such as dunite and gabbro.

Extremely rich deposits of gold ore have been mined in gossan deposits (oxidized cappings) formed at the surface by the weathering and consequent chemical decomposition of gold-bearing mineralized zones. The gold is resistant to chemical alteration and thus becomes concentrated downward *in situ* while other minerals, including the gangue, are removed by solution. Gossans are characterized by their vugular structure (cavities and cavity fillings) and rusty appearance caused by the abundance of limonite, a mixture of the hydrous iron oxide minerals goethite and lepidocrocite. Although such concentrations of gold result from chemical decomposition of the enclosing ore body, they are essentially formed by the physical movement downward of gold particles. Deposits of kyanite in India, forming surficial concentrations resulting from the weathering of a kyanite–quartz granulite, and deposits in the United States formed from the weathering of kyanite schist, have been mined.

Supergene or secondary enrichment. Supergene minerals are those that have formed during a period of secondary mineralization by the alteration of primary min-

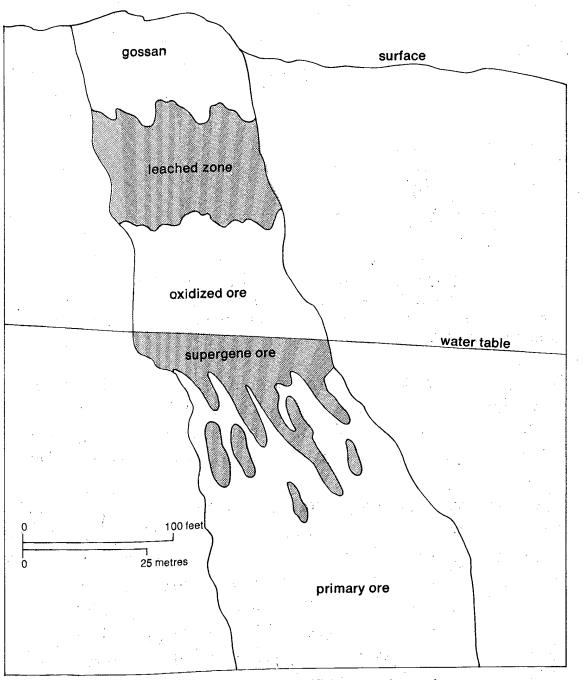


Figure 9: Supergene or secondary sulfide ore deposit (see text).

erals. They are alteration products that have commonly formed near the surface by the action of weathering and descending meteoric water (Figure 9). Where an ore body extends to the surface and weathers *in situ* a gossan is formed in the zone of oxidation above the water table. Meteoric water seeping downward through the zone of oxidation becomes a leaching solution that dissolves minerals in the upper zone and deposits minerals in lower zones. Above the water table and below the gossan and leached zone, carbonates, oxides, and silicates are commonly formed. Below the water table, and above the zone of primary mineralization in the ore body, enrichment takes place. Below the gossan or rusty cap rock the leached zone may be barren of minerals with the exception of gold, which, although not chemically inert, tends to remain unaltered.

Enrichment of sulfide deposits

Minerals can be dissolved and removed, and their constituents deposited elsewhere as other minerals, or they can be replaced. Chemical alteration of sulfide minerals within the zone of oxidation results in the formation of acids; these, in turn, enable the meteoric water to act as a solvent. For example, the iron sulfide pyrite, a common mineral in ore deposits, weathers to ferric sulfate and sulfuric acid. Ferric sulfate reacts with the sulfide minerals of copper, lead, and zinc to form sulfates of these metals. Sulfuric acid is a potent solvent of carbonates such as the gangue minerals calcite, dolomite, and siderite. The metals removed by solution are carried downward, and secondary enrichment takes place in a range extending from immediately below the water table to considerable depths in the zone of primary mineralization. It has been demonstrated in the laboratory that metals in solution are precipitated as sulfides in the presence of sulfides of other metals; and that, in the order of solubility, the sulfide of iron is higher than the sulfides of zinc and lead, which in turn exceed copper sulfide. In solutions of metal and sulfide ions, when the latter are not sufficient to unite with all the former, only the least soluble sulfides will crystallize. In the zone of secondary enrichment copper minerals such as chalcocite and covellite not only replace other copper minerals, including chalcopyrite, but also the lead and zinc minerals galena and sphalerite. The paragenesis of these relationships can be complex, and interpretation may depend not only on examination of polished sections of the ore under a microscope but on analysis by X-ray equipment such as the microprobe.

Supergene sulfide enrichment can be of considerable economic significance in a mining operation. In the case of large, low-grade, copper porphyry deposits that average less than 1 percent copper by weight, the grade has been increased by secondary enrichment to as much as 5 percent and is commonly over 2 percent.

Alteration of minerals, rocks, and ore bodies

Metamorphism and crustal movements. Metamorphism means a change in form and refers to the mineralogic reconstitution that occurs in a rock or mineral deposit that is subjected to variations in pressure, temperature, and hydrothermal agencies. Under increased pressure and temperature an alumina-rich shale can be converted to a schist characterized by the minerals kyanite, sillimanite, andalusite, or garnet, all of which are industrial minerals. Under similar conditions an ore deposit of metalliferous minerals such as galena, chalcopyrite, or stibnite can develop flow structures and become gneissic, that is to say streaked and banded in appearance. In such cases the origin of the ore is not easily determined, particularly as to whether it was syngenetic or a replacement of earlier foliation. Crustal movements producing intense folding can result in ore bodies that show highly contorted laminae of mineralization.

The physical properties, including the stability ranges, of sulfide minerals under dynamothermal metamorphism are not well known. Some, in particular galena, are readily deformed in a manner suggesting plastic flowage, but the deformation is caused by a mechanism known as translation, whereby movement takes place along crystallographic planes without rupturing the mineral grains. Galena deformed in shear zones can have the appearance of shiny, striated steel.

Metamorphism and crustal movements may also be the mechanisms whereby low concentrations of metalliferous sulfides in sedimentary beds can be mobilized and driven off, to be concentrated in zones of lower temperature and pressure. It is recognized that some metallogenic provinces may have had close genetic and spacial relationships to sedimentary beds containing disseminated copper, lead, zinc, and other sulfides. Much work remains to be done along these lines before the origin of many ore deposits that appear to have been altered or formed by metamorphism can be stated unequivocally.

METALLOGENIC EPOCHS AND PROVINCES

Ore deposits in Precambrian rocks

The world is more than 4,600,000,000 years old. Most of this long span of time, up to 570,000,000 years ago, is known as the Precambrian; the subsequent span is divided into the Paleozoic, Mesozoic, and Cenozoic eras. Many ore deposits are found in Precambrian rocks, including the iron deposits of the Lake Superior region in North America, the Kiruna district in Sweden, and Western Australia; the nickel–copper deposits of Sudbury, Ontario; the chromium deposits of South Africa; and numerous deposits of lead, zinc, manganese, tin, tungsten, asbestos, platinum, cobalt, silver, and gold. This does not mean that the ore deposits found in Precambrian rocks today are necessarily of Precambrian age. The iron ore deposits of Western Australia and Lake Superior are of recent origin, having formed as residual concentrations of the original iron content of carbonates, silicates, and oxides in the Precambrian sedimentary rocks. Many ore deposits of Precambrian age must have been transformed by metamorphism or mobilized in whole or in part by heat, pressure, and hydrothermal agents.

During the Paleozoic Era (225,000,000 to 570,000,000 years ago) there were several periods of igneous activity and orogenic movement (mountain building) of the earth's crust, accompanied by mineralization in various parts of the world, notably in Europe, Asia, and Australia. Many deposits of salts, gypsum, phosphatic rock, and sedimentary copper and iron ore in North America and Europe were formed during the Paleozoic and the succeeding Mesozoic Era (65,000,000 to 225,000,000 years ago). The Mesozoic was also a time of igneous activity and orogeny. In western North America these disturbances resulted in copper, lead, zinc, gold, and silver mineralization in the marginal zones of the intrusive granitic rocks. Similar mineralization occurred about the same time in Japan, the U.S.S.R., and other parts of the world.

The Tertiary Period of the Cenozoic Era (2,500,000 to 65,000,000 years ago) was also a time of widespread crustal disruption and igneous intrusion. In western North America several large deposits of porphyry copper–molybdenum ore lie in the peripheral zones of intrusive bodies of granodiorite and monzonite emplaced during the early Tertiary. Alpine intrusions of Tertiary age in Europe resulted in lead–zinc mineralization. In other parts of the world, mineralization associated with Tertiary intrusives include ores of gold, silver, copper, lead, zinc, antimony, tungsten, tin, and mercury.

The Precambrian shields of the world, and the orogenic belts that exposed bodies of intrusive rock, are the main prospective provinces for metalliferous ore deposits, particularly those of hydrothermal origin that are spatially controlled by magmatic and structural features. Other provinces, including those of syngenetic origin such as the Kupferschiefer copper deposits in Germany, and some of probable hydrothermal origin such as the lead–zinc deposits of the Tri-State district, Mississippi Valley, are controlled by stratigraphic factors. Within broad geographical limits certain areas are noted for their deposits of gold and copper; copper and molybdenum; lead, zinc, and silver; cobalt and silver; chromium and the platinum metals; tin and tungsten; diamonds, and other associated minerals. These associations are, from field observations, related to geological situations, but the reasons for the relationships are not always clear. Although many large ore deposits may yet be found, production from known deposits indicates that it is unlikely that ore deposits are distributed uniformly throughout the world.

BIBLIOGRAPHY. F.D. ADAMS, "Origin and Nature of Ore Deposits: An Historical Study," *Bull. Geol. Soc. Am.*, 45:375–424 (1934), a good review of the development of theories concerning ore genesis; A.M. BATEMAN, *Economic Mineral Deposits*, 2nd ed. (1950), a useful textbook and general reference, and (ed.), *Economic Geology* (1955), symposium papers on specific ore deposits; R.L. BATES, *Geology of the Industrial Rocks and Minerals* (1960), a comprehensive reference on nonmetallic minerals; P.T. FLAWN, *Mineral Resources* (1966), a useful reference for mineral economics; C.A. LAMEY, *Metallic and Industrial Mineral Deposits* (1966), a good reference to use in conjunction with Bates' text; W. LINDGREN, *Mineral Deposits*, 4th rev. ed. (1933), an older textbook regarded as a classic, and particularly useful to compare with Adams' paper; T.S. LOVERING, *Minerals in World Affairs* (1943), a reference work on the historical development of mineral resources and economics; H.E. MCKINSTRY, *Mining Geology* (1948), a good textbook on structural and other physical characteristics of ore deposits; W.H. NEWHOUSE (ed.), *Ore Deposits as Related to Structural Features* (1942), symposium papers on specific mineral deposits with particular reference to structural problems in mining; C.F. PARK and R.A. MACDIARMID, *Ore Deposits* (1964), an excellent textbook and general reference for students; F.G. SMITH, *Physical Geochemistry* (1963), a comprehensive textbook of particular use as a reference to chemical reactions involved in the genesis of ore deposits.

(C.E.B.C.)

Oregon

Admitted to the Union as the 33rd member in 1859, Oregon comprises a region of startling physical diversity, from the moist rain forests and mountains and the fertile valleys of its western third to the naturally arid and climatic harshness of its eastern deserts. Mountains, plateaus, plains, and valleys of different geological ages and materials are arrayed in countless combinations, including such natural wonders as the Columbia River Gorge, Oregon Caves National Monument, Crater Lake National Park, the majestic snow-covered peaks of the Cascade Range, and the "moon country" of central Oregon.

An overview of the state

Historically, Oregon comprised all of the United States' Pacific Northwest, a region that today includes the states of Oregon, Washington, and Idaho, and a small portion of Montana west of the Rocky Mountains. To the north of the state's 96,981 square miles (251,181 square kilometres) of land and inland water lies Washington, from which Oregon receives the waters of the Columbia River; to the east, Idaho, much of its border formed by the winding Snake River and its Hells Canyon, the deepest gorge on the North American continent; to the south, Nevada and California, with which Oregon shares its mountain and desert systems; and, to the west, the Pacific Ocean, to the beneficent influences of which Oregon owes the moderate climate of its western lands.

The forested mountains of western and northeastern Oregon have supplied the traditional core of the state's economy. Its many forest-product plants produce more than one-fifth of the nation's softwood lumber, one-half of its plywood, one-fourth of its hardboard, as well as large quantities of pulp and paper. In addition, the multipurpose development of the Columbia River System provides huge quantities of electricity, water for irrigation and industry, shipping channels, and water for recreation. The heartland of Oregon, however, is the Willamette Valley, containing the major cities of Portland, Eugene, and Salem (the capital) and a rich and diversified agriculture. (For information on related topics, see UNITED STATES; UNITED STATES, HISTORY OF THE; NORTH AMERICA; PACIFIC COAST RANGES; and COLUMBIA RIVER.)

THE HISTORY OF OREGON

The Indian cultures of Oregon

When the first white men arrived in the Oregon Country —a region vaguely defined at the time but roughly comparable to the present Pacific Northwest—about 125 Indian tribes with a population estimated at 100,000 to 180,000 lived in and around the area. In what became present-day Oregon, the leading tribes were the salmon-eating Chinook, along the lower Columbia River; the Tillamook, Yamel, Molaha, Clackamus, and Multnomah in the northwest; the Santiam and Coos in the southwest; the Cayuse, Northern Paiute, Umatilla, Nez Percé, and Bannock in the dry lands east of the Cascade Range and in the Blue–Wallowa Mountains; and the Klamath and Modoc in the south central area. Their mode of life was responsible for their small population. Since they had no form of agriculture and no domesticated animals other than the dog, they depended entirely upon the natural fauna and flora of the land and water, existing with crude implements by gathering, hunting, and fishing. The tribes along the Columbia River, known as the Canoe Indians, fashioned excellent canoes from logs.

The explorers. The first white men to see the Oregon coast were Spanish sailors searching for a northwest passage to facilitate trade with the Orient. In 1579 the English buccaneer Francis Drake, in quest of Spanish loot and a northwest passage in his "Golden Hind," anchored in an inlet north of the Golden Gate and with a brass plate "took possession" of the country for Queen Elizabeth I. Until the third quarter of the 18th century, when the Spanish renewed exploration along the coast, the Oregon Country remained unexplored. In 1778 the English sea captain James Cook visited Oregon. His men bought beaver and other skins, which they sold at huge profits in China.

In 1787 Boston merchants sent two ships to the Oregon Country under the command of Captains Robert Gray and John Kendrick. On his second voyage, Gray entered the harbour that bears his name (in Washington), and in May 1792 he sailed over the bar of the Columbia River and named it after his ship, the "Columbia." This was the first United States claim to the Pacific Northwest by right of discovery.

The Northwest also was approached by land. Two English fur companies, the Hudson's Bay Company and the North West Company, raced across the continent to open routes to the Pacific; the Americans were not far behind. Meriwether Lewis and William Clark reached the mouth of the Columbia in 1805, strengthening the United States claim to the region.

John Jacob Astor, at the head of the Pacific Fur Company, began the white settlement of the Oregon Country with the establishment of a trading post at Astoria in 1811. The Hudson's Bay Company established Ft. Vancouver in 1824. Dr. John McLoughlin was appointed to head this company's far-flung operations, and for the next 22 years he was the dominating figure in the region.

Permanent settlement. From 1830 onward, thousands of Americans from the Middle West migrated to the Pacific Northwest. Missionaries played a role in settlement. In 1834 the Methodists, headed by Jason Lee, established the first permanent settlement in the Willamette Valley. The migrations that carved deep wagon wheel ruts still visible in the Oregon Trail began in the early 1840s. These settlers pressed for a practical answer to the undetermined ownership of the Oregon Country. After 1838 United States claims and rights to the region were constantly before Congress. American settlers in the Willamette Valley made known their desire to become part of the United States. In 1843 representatives met at Champoeg to organize a provisional government; a set of laws patterned after those of Iowa was accepted. By 1844 the British government had concluded that the Columbia River as the boundary line would have to be abandoned, and the Hudson's Bay Company moved its chief Northwest depot to Ft. Victoria. In spite of the "Fifty-four Forty or Fight" slogan of the presidential campaign of 1844, the 49th parallel was accepted by both nations as the boundary, and the Oregon Country was added to the United States in 1846.

Disputes over ownership

The influx of population led to political agitation, and in 1853 the Washington Territory was given independent status; the Idaho Territory gained similar status in 1863.

Statehood and growth. By 1883, following several "wars," most of the Indians of Oregon were on reservations. The same year saw the beginning of the linkage of Oregon with the rest of the nation by railroad, vastly improving the opportunity for economic growth. Agriculture and forestry were especially stimulated, and by the turn of the 20th century, two-thirds of the people of Oregon lived in rural areas.

The 20th century has witnessed rapid growth of cities, and since 1940 there has been significant diversification of the economy. Today two-thirds of the people live in urban areas.

THE NATURAL AND HUMAN LANDSCAPE

The great diversity of landforms and climates in Oregon is reflected in the different patterns of human settlement and varying bases of economic activity throughout the state.

The natural environment. *Physical regions and vegetation.* Oregon has nine major landform regions: the Coast Range, the Klamath Mountains, the Willamette Valley, the Cascade Range, the North Central Oregon Plateau, the Blue–Wallowa Mountains, the High Lava Plains, the Basin and Range Province, and the Malheur–Owyhee Upland.

Elevations

The forest-blanketed Coast Range, which borders the Pacific Ocean from the Coquille River northward, is the lowest of Oregon's main mountain systems. Its elevations are usually below 2,000 feet, but Marys Peak, southwest of Corvallis, reaches 4,097 feet (1,249 metres), the highest point in the range.

The Klamath Mountains, which extend into Oregon from California, lie south of the Coast Range and west of the Cascades. Composed of ancient resistant rocks, they have had a complicated geologic history. They are higher and more rugged than the Coast Range and lack the north–south orientation. The famous Rogue River, bisecting the area, provides the major drainage. Thick forests grow on these mountains, which also contain the state's richest mineral deposits.

The Willamette Valley is essentially an alluvial plain that has been produced by burying stream-modified lowland with enormous quantities of sediments brought by tributary streams from the bordering mountains. The low hilly areas in the central and northern portions are composed of resistant rocks. This valley contains the prime land of the state and is its population centre. Its soils support intensive agriculture.

The Cascade Range in Oregon forms a broad lava plateau. The wider western section is deeply eroded by numerous streams fed by heavy precipitation. The eastern section, less dissected, is crowned with a chain of volcanic peaks. Mt. Hood, reaching 11,245 feet (3,427 metres) above sea level, and Mt. Jefferson, rising to 10,499 feet (3,200 metres), are the highest. The western slopes of the Cascades are mantled with Douglas fir forests; on the upper slopes western hemlock and true firs become dominant. The forests of the drier east side are largely of ponderosa pine.

In the North Central Oregon Plateau, a portion of the Columbia River Basin, streams are entrenched and provide some bold relief. The interstream areas are broad, little-dissected, smoothly rolling surfaces that provide the land for Oregon's great wheat ranches.

The Blue–Wallowa Mountains comprise two separate highland masses in the northeastern part of the state. The name Blue Mountains refers to the eroded plateaus and ranges extending westward from the agriculturally important Grand Ronde and Baker valleys. Basins and valleys, the headquarters for large cattle ranches, are scattered through the Blue Mountains. The Wallowa Mountains, beyond the Grande Ronde and Baker valleys and near the Idaho border, contain the highest elevations in northeastern Oregon. They were strongly glaciated and display some of the most spectacular scenery to be found in the American West.

The area of the flat High Lava Plains, or High Desert, is located south of the Blue Mountains and eastward from the Cascade Range. The smoothness of the surface, however, is broken by cinder cones, buttes, and craters. Immaturity of erosion and localized interior drainage are other features. Low precipitation, short and erratic growing seasons, and the absence of soil in many places result in an arid landscape of skimpy vegetation, with the details of the surface features commonly visible.

The Basin and Range Province to the south, which merges with the High Lava Plains, is a youthful high lava plain interrupted by mountains and fault troughs. Small volcanoes are numerous in the western portion, where an extensive sheet of pumice greatly modifies surface runoff, vegetation, and land use. Irrigation agriculture is practiced in the Upper Klamath Lake area, and hay is grown with irrigation in a number of other basins and valleys, but most of this region is used by range livestock.

The Malheur–Owyhee Upland of southeastern Oregon is for the most part a high, warped plateau. It contains older lava and has been more dissected than the High Lava Plains. The major drainage system, the Owyhee River, has incised several notable canyons in an area locally called the "Rimrock Country." Along the Snake River in the east central portion of the state there is highly productive irrigation agriculture, but most of this region is livestock-grazing country.

Climatic variation

Climatic regions. Oregon's climates range from equable, mild, marine conditions on the coast to continental conditions of dryness and extreme temperature in the interior. Location with respect to the ocean, prevailing wind and storm paths, and topography and elevation are the principal controls. Six climatic areas can be recognized.

The narrow coastal area and the bordering mountain slopes are marine influenced. Temperatures are mild and equable: July averages 55° to 60° F (13° to 16° C), January about 40° F (4° C). Summers are relatively dry but receive only half of the possible sunshine; other seasons are cloudy and wet. Annual precipitation ranges from 60 to 120 inches (1,500 to 3,000 millimetres) or more.

The lowlands of the Willamette, Umpqua, and the Middle Rogue rivers are warmer in summer, slightly cooler in winter, and have less precipitation than the coast. July averages 67° to 72° F (19° to 22° C) and receives 65 to 70 percent of the possible sunshine; January averages about 40° F (4° C). The rainy season extends from October through April, with precipitation averaging 35 to 40 inches (875 to 1,000 millimetres) except in the Middle Rogue Valley, where 20 to 25 inches (500 to 625 millimetres) are common.

The Cascade Range has copious winter precipitation, including phenomenal snow depth, and short, dry, sunny summers. Above 3,000 feet, January average temperatures are below 32° F (0° C). Snow begins to fall in October and remains through April, with large patches persisting until July. The higher peaks support snowfields and small glaciers throughout the year. July average temperatures range from 50° to 60° F (10° to 16° C) depending on elevation.

The North Central Oregon Plateau, stretching from northern Wasco County through northern Umatilla County, is sufficiently elevated and exposed to receive ten to 20 inches (250 to 500 millimetres) of precipitation. Distribution is fairly even, but winter has the majority of the rainy days. Summers are sunny, with July average temperatures 70° to 75° F (21° to 24° C). The brisk winters have considerable sunny weather, and January temperatures average 31° to 33° F (−1° to 1° C). The plateau area of central and southeastern Oregon has climatic characteristics similar to the north central plateau except for somewhat lower precipitation and lower temperatures at higher elevations.

The Blue–Wallowa Mountains have variety in climatic detail. The intermontane basins and valleys are similar to the north central plateau, with colder winters. The higher, exposed elevations receive comparatively heavy precipitation, much in the form of snow during winter.

Human imprints. At least five major patterns of human land use emerge from the tangle of Oregon's natural landscapes and climates. The forested mountains—the Coast Range, the Cascades, the Klamath, and the Blue–Wallowas—show relatively little evidence of human habitation or modification except for the harvest pattern of block cutting in the Douglas fir region, the logging and forest-management roads, and scattered roadside homesites at lower elevations. Most loggers—few in number because of technological efficiency—live in the valley towns.

Man's uses of the land

The western valleys, dominated by the Willamette, are

Oregon's main centres of population, industry, and transportation. Most persons live close to well-populated centres. The nearly 1,300 small sawmills that in 1947 were located in valley towns or up tributary valleys into the forested mountains have dwindled to fewer than 250; but these are large, integrated operations producing a multiplicity of forest products.

In the rolling, sparsely populated wheat country of north central Oregon, ranches commonly exceed 1,500 acres in the eastern portion and double that size to the west, where wheat-fallow rotation is practiced. In regions of natural erosion, alternate bands of crop and fallow occur. Farmsteads are widely separated, and owners often live in towns.

The growth of natural feed in wide-open range country is relatively poor, and cattle scatter over enormous areas; seldom do more than a few cluster. Appurtenances of the area include fences and occasional watering places with a metal tank. Ranchsteads are few and far between, and ranchers travel about in four-wheel-drive pickup trucks.

Most of the eastern Oregon towns, except Pendleton, lie in the area of irrigated agriculture, on the eastern slopes of the Cascades or near the Idaho border. Farming is highly mechanized.

THE PEOPLE OF OREGON

Composition. Oregonians are predominantly United States-born. The less than 4 percent of foreign birth comprise mainly older persons who immigrated from the Scandinavian countries, Finland, and Canada. Roman Catholics form the largest single religious denomination in Oregon, but about 77 out of 100 church members are of the Protestant faiths. Methodists, Baptists, Presbyterians, Disciples of Christ, and Lutherans are the major Protestant groups.

Contemporary demography. The 1970 census reported that Oregon had 2,091,385 inhabitants, an 18.2 percent increase since 1960. The people are unevenly distributed, 87 percent living west of the crest of the Cascade Range and 69 percent in the Willamette Valley. Average densities in eastern Oregon are, for the most part, low: Harney County had a 1970 average of 0.7 person per square mile, Lake County, 0.8, and Wheeler County, 1.1.

Rural and urban Oregon

About two-thirds of the Oregonians lived in urban areas; and rural population declined in 22 of the 36 counties, while only seven showed losses of urban population.

Approximately two-thirds of all Oregonians live in the three Standard Metropolitan Statistical Areas of the state, Portland, Eugene, and Salem. Portland, near the confluence of the Willamette and Columbia rivers and the largest city in the state, is a leading West Coast port and the major commercial, industrial, service, and cultural centre of the state. Eugene and Salem, the second and third cities, are important for trade and processing. Salem, the state capital, is among the nation's leading food-processing centres. The major cities outside the Willamette Valley are Medford, in the Rogue Valley; Klamath Falls, in south central Oregon; and Pendleton, in the north central plateau.

THE STATE'S ECONOMY

Traditionally, Oregon has had a resource-oriented economy, strongly dependent upon its forests and farms. In recent years diversification has occurred as various new industries have been established and trade and service activities have grown.

Components of the economy. *Manufacturing.* Forest-products manufacturing still ranks as Oregon's leading industry. It accounts for about one-half of the state's 5,000 manufacturing establishments, of its manufacturing employment, and of the value added by manufacturing. About one-half the land area of the state is forested. Public agencies control about 85 percent of Oregon's commercial forest, private owners about 14.9 percent. Another 4,100,000 additional acres of forest are reserved for recreation and watershed use. The forests are capable of permanently sustaining a wood-based sector in the economy of the present size.

Forestry as the economic base

The forest industry began as a producer of lumber; since 1938 Oregon has been the leading state in softwood lumber. In recent years the products have been changing radically, and now only 40 percent of the forest income is from lumber. About one-quarter of the logs harvested go into plywood, which accounts for about one-third of the value of forest products. Pulp and paper plants and hardboard and particle-board plants contribute most of the remainder.

Food processing adds about $325,000,000 to the value of agriculture and fishery commodities and employs about 25,000 workers. The development of sources of electricity, the availability of natural gas via pipeline, abundant water, and the growth of population are the assets upon which new industries have been based. The metals-related group of industries—including primary metals, fabricated metals, electrical and other machinery, and transportation equipment—has been the pacesetter. In the 1960s, employment in these industries doubled, and the value added now exceeds that of food processing. The greatest concentration of metals-related industries is in the Portland metropolitan area, but an aluminum smelter is located at The Dalles, and Albany has two metal-processing plants.

Agriculture and fishing. The agricultural land base of Oregon includes about 5,400,000 acres of cropland, 11,500,000 acres of farm pastures and ranges, and 20,900,000 acres of public range; about 1,700,000 acres are irrigated. Livestock products contribute nearly one-half of the total commodity value, led by cattle and calves; dairy and poultry products are also significant. Wheat is the leading crop, with vegetables and fruits among the other major crops.

Chinook, silver, blueback, and pink salmon are the most valuable fishery products, contributing about 12 percent of the fishery volume and 25 percent of the value. Shellfish amount to about 25 percent of both volume and value; other fish include flounder, tuna, ocean perch, and rockfish.

Mining. Oregon mines $60,000,000 to $70,000,000 worth of minerals annually; sand and gravel make up the bulk of the value. Quarrying occurs in every county, but the greatest quantities are taken near the growing urban areas. The only nickel mine in the nation is located near Riddle. Mercury is also mined.

Transportation. In 1970 Oregon had more than 41,000 miles of highways and roads under the jurisdiction of the State Highway Commission, the Federal-Aid Secondary Highway System, and counties and municipalities. In addition, more than 52,000 miles of forest development roads, national park roads, and military and Indian reservation roads are controlled by federal agencies and various local governments. Some 5,000 miles of railroads provide north–south and east–west routes. The largest airport is Portland International Airport; other significant commercial airfields are at Eugene, Medford, Pendleton, and Corvallis.

Waterways and shipping

Through its entire history, water transportation has been important to Oregon, and today the state has 23 port districts. Six are located on the Columbia above the head of deep navigation, where barge traffic is composed principally of grain and petroleum downstream and cement and structural steel upstream. Portland, open to oceangoing vessels, is by far the most important port. The other districts stretch along the Oregon coast and up the Columbia on the deep-draft channel. Astoria, Newport, and Coos Bay, in addition to Portland, have regular shipments to and from foreign countries.

ADMINISTRATION AND SOCIAL CONDITIONS

The "Oregon System": initiative, referendum, and recall

Structure of government. Oregon has been in the vanguard of several innovative movements in American government collectively known as the "Oregon System." In 1902 the concepts of initiative and referendum were introduced, by which voters were able to initiate and vote upon statutes or constitutional revisions; these were supplemented in 1908 by the system of recall, under which the removal of elected officials could be initiated by the voters. The state was also one of the earliest to impose a state income tax, in 1923.

The state level. State government follows the pattern of most states, though the governor is perhaps stronger than in many. Limited to two four-year terms within any 12-year period, he supervises the state budget, agency heads, boards, and commissions, and coordinates their activities, initiates future planning, and is the focus of federal–state interaction. He may also veto individual items in appropriation bills.

Legislative power is shared by the people of Oregon, through the system of initiative and referendum, and their elected legislators. The legislature comprises the Senate, with 30 members serving four-year terms, and the House of Representatives, with 60 members serving for two-year terms.

The court system is headed by the seven-justice Supreme Court, which has general administrative authority over all other courts. The justices, elected for six-year terms, elect one of their members as chief justice.

Local government. Oregon gives its towns and cities home rule, the right to choose their own form of government. Most cities with populations greater than 5,000 have the council–manager form of government, whereas smaller cities are governed by a city council and a mayor. Portland is governed by four commissioners and a mayor.

In 1958 home rule was extended to counties, but to date few have taken action under this privilege. In most counties, a county judge and two commissioners or a board of three commissioners exercise the powers of government. These officials usually are elected for terms of three years.

Political life. The Republican Party has dominated Oregon's politics through much of its history. Although with industrial growth in recent decades Democrats have come to outnumber Republicans in registration, Republicans continue to control the governorship and most major elective offices. An unusual number of Oregonians have made their mark in the U.S. Congress by their independent stances, perhaps a reflection of continuing frontier attitudes.

The social milieu. Although crime and related problems have increased, Oregon in the early 1970s was far less beleagured than many other states by issues related to such social ills as deteriorating inner cities and inadequate tax bases to pay for rising costs in education, welfare, and health care. Racial problems have been few, and only Portland, since World War II, has had a significant black community.

Areas of state involvement. Oregon's biennial budget consists of segments supported by General Fund and Other Fund revenues. The General Fund is derived from personal and corporate income taxes, excise, inheritance, and insurance taxes, and liquor sales. Other Fund revenue comes from federal grants, use taxes, trust funds, licenses, and the sale of services and commodities.

Funding of state and social activities

Health, education, public welfare, corrective institutions, legislative and judicial functions, and general government administrative functions are supported out of the General Fund. Activities substantially supported by the Other Fund include transportation programs, employee-protection programs, regulatory activities such as public utilities, banking, and corporations, and some natural resources functions. In 1971 the legislature passed a far-reaching program to deal with air and water pollution.

Education. French Prairie, present-day Wheatland, was the site of Oregon's first school, in 1834; 15 years later the first free public school system was created by the territorial legislature. Today a board of education, appointed by the governor, and an elected superintendent of public instruction oversee the system.

Opportunities for education after high school are provided by 12 community colleges, a state system of higher education comprised of three universities, four regional colleges, and 19 independent colleges. The community colleges (operating under state law and guidelines established by the State Board of Education) are administered by lay boards, locally tax supported and especially responsive to local needs in their curricula.

Reed College in Portland is a private liberal arts institution with a relatively short (founded 1911) but distinguished history. It has an extraordinary record in the proportion of its graduates who go on to advanced degree elsewhere. Willamette University, in Salem, granted its first degree, "Mistress of English Literature," in 1859.

CULTURAL LIFE AND INSTITUTIONS

As a relatively young region of the United States, and one in which the imprints of man are scarcely visible over vast stretches of land, Oregon has not developed a cultural identity equivalent to those of the longer-settled or more heavily populated regions. Its people, however, no less in the sparsely settled areas of the east than in the Willamette Valley centres, take full part in the increasingly homogeneous character of American life. Television, radio, and newspapers are available in all corners of the state. Theatrical and musical groups are found in the cities and larger towns, and the Oregon Shakespearean Festival in Ashland draws thousands of viewers each summer to its productions. University and college communities have available many public offerings in the arts or other cultural activities.

The cultural milieu

Popular culture. In addition to the ubiquitous sporting events, both spectator and participatory, Oregon offers a number of attractions related to its history and its location. These include the Pendleton Round-Up (held ironically in wheat country), which attracts participants from across the West and spectators from around the Northwest. Albany's World Championship Timber Carnival, which takes place each July 4, features logger events, carnivals, a parade, and the like. Portland's Rose Festival in early June is perhaps the most famous of the state's communal celebrations.

Libraries and museums. Oregon has about 100 free public library systems, including about 20 county libraries and several travelling libraries, or bookmobiles. The Multnomah County Library, in Portland, was the first to serve the public on a large scale; it began membership service in 1864 and free service in 1902. The Oregon State Library in Salem maintains a general reference service and loan collection for use by the public either directly or through local libraries.

The Oregon Historical Society in Portland and the Horner Museum at Oregon State University own large collections of items of pioneer days in the Oregon Country. The Oregon Museum of Science and Industry in Portland features demonstrations of science at work in Oregon industries. The Portland Art Museum features Northwest Indian art and pre-Columbian Mexican art in its collection. The Murray Warner Collection of Oriental Art at the University of Oregon is one of the largest collections of its type in the United States.

PROSPECTS

A retrospective look over Oregon's development suggests a romantic early period, with a succession of voyagers by sea, overland explorers, traders and trappers, and pioneer settlers. In the recent period, the pattern is more prosaic. Economic growth, reflected in national patterns of employment and income, has been similar to that of the nation as a whole, and this industrial development has been characterized by intensification in forest products other than lumber and by utilization of electric power and skilled labour. Growth in trade and services has responded to population growth, improved transportation, and urbanization. The outdoor recreation attractions have been increasingly discovered by Americans, popularizing the state as a vacation land. With the state's increased concern with its environment, it is likely that Oregon will continue to offer the appurtenances of amenable modern living and a close relation to its natural wonders.

BIBLIOGRAPHY. EWART M. BALDWIN, *The Geology of Oregon*, 2nd ed. (1964), a handy summary reference; CHARLES W. BOOTH, *The Northwestern United States* (1971), a description and analysis of the geographical character of the region; PHIL F. BROGAN, *East of the Cascades* (1964), a popularization of central Oregon's geology; SAMUEL N. DICKEN, *Pioneer Trails of the Oregon Coast* (1971), a regional geography; RICHARD M. HIGHSMITH (ed.), *Atlas of the Pacific Northwest*, 4th ed. (1968), a regularly revised sourcebook of maps of the Pacific Northwest; EDWIN R. JACKMAN and R.A. LONG, *The Oregon Desert* (1964), a definitive work on the

Oregon ranch country; RICHARD L. NEWBERGER, *The Lewis and Clark Expedition* (1951), an outstanding historical contribution; *The Oregon Blue Book* (biennial), regularly revised information on all aspects of Oregon's political organization, population, settlement, and economy.

(R.M.Hi.)

Organ

A keyboard instrument in which the sound is produced by pipes to which wind is supplied through a mechanism under the control of an operator is an organ. Its technical classification is aerophone, or wind instrument. The word organ derives from Greek *organon* and Latin *organum*, an instrument. By common usage, organ has come to embrace any keyboard instrument capable of producing indefinitely sustained sounds, but these should be particularized as reed organ (harmonium) or electronic organ. Organ, alone, implies an organ with pipes, and the term pipe organ is tautologous. This article describes the mechanical aspects, tone production elements, and historical development of the organ as an instrument and relates national schools of composers and their works.

PARTS, MECHANISM, AND PRODUCTION OF SOUND

Basic parts. An organ is divided into three main parts. At one end of the instrument are the keyboards, or manuals, and other controls that collectively are called the console. At the other end are the pipes that produce the tone. Between these two is the mechanism, or action, that accounts for a large part of the bulk and cost of any organ. The simplest type of organ has one keyboard and one pipe to each note. The pipes stand in a row on an airtight box or chest that is supplied with wind, through a trunk, from bellows. Under each pipe is a valve, or pallet, connected by a system of cranks and levers to its respective key of the keyboard. A reservoir is interposed between the bellows and the wind-chest, appropriately weighted to keep the supply of wind at a constant pressure. This reservoir has a blowoff valve that comes into operation when the reservoir is full. Although the bellows may resemble basically the familiar domestic type that is operated by hand or feet, wind is normally supplied from an electrically driven rotary blower.

Elements of simple organs

The pitch of each note is determined by the length of the pipe; the longest pipe makes the deepest note, the shortest pipe the highest note. If two comparable pipes sound an octave apart, the effective length of the higher-pitched pipe is exactly half that of the lower-pitched.

Since the tone of a pipe sounding on a constant pressure of wind is immutable, both as to quality and quantity, the uses of an organ with only one pipe to each note are strictly limited. Even the smallest organs, therefore, have at least three pipes to each note, and organs of cathedral size commonly have as many as 100 to each note. These sets, or ranks, of pipes are arranged in parallel rows on the wind-chest. The pallet controlled from each note admits wind to all the pipes belonging to that note; but, in order that the organist may be able to use at will all, none, or any of the sets of pipes, an intermediate mechanism is provided by which he may stop off any set or sets of pipes. From this function the control at the console by whose operation the pipes are stopped off has come to be known in English as a stop, a term also used loosely for each rank of pipes.

Mechanism. *Tracker (mechanical) action.* The operative part of the stop mechanism lies between the pallet and the footholes of the pipes. It normally consists of a strip of wood or plastic running the full length of the set of pipes, or stop. In it is drilled a series of holes. One hole registers exactly with each pipe. The strip of wood is placed in a close-fitting guide in which it may be moved; when it is moved longitudinally a short distance, so that its holes no longer register with the pipes, wind will no longer reach that set of pipes, even when the organist opens the pallets. These strips are therefore called sliders, and wind-chests in which the stops are operated in this way are called slider chests. There are other ways of working the stops, both ancient and modern, which will be referred to later; but the slider chest was in almost universal use before the 20th century, and many modern organ builders consider it the best. The slider is connected to the console by a system of levers and cranks, and it terminates in a knob that the organist pulls toward him to bring the stop into play or pushes in to silence it.

The importance of two manuals

It often happens that the organist needs either to play polyphonic music (*i.e.*, with interweaving of several voices) in two or more contrasted parts, to give prominence to a melody against a softer accompaniment, or to play loud and soft passages in rapid succession. None of these effects can be achieved on an organ with one keyboard, or manual, as so far described. Loud and soft passages can be played to some extent, but to change the stops between each alternation takes time, which is not always available. For this reason, organs of more than about seven or eight stops usually have two manuals, each controlling its separate wind-chest and stops. Each manual department is self-contained, so that the organ is really a composite instrument. By pre-arranging the stops on the manuals, the organist may perform in any of the three ways mentioned above. The organist, therefore, may vary the sounds he produces in one or both of two ways: by changing the stops on the manuals he is playing or by leaving the stops as they are and changing from one manual to another.

Since the 18th century he has had yet a third way of controlling the volume of sound. The pipes of one or more manuals are usually placed in a box, one side of which consists of hinged and movable shutters (similar to vertical Venetian blinds) that are connected to a pedal at the console. By opening and closing the shutters, the sound from the stops of the manual concerned is made louder or softer. Such boxes are called swell boxes.

Since the 14th century, one department of the organ has commonly been played from a keyboard, or more properly a pedalboard, controlled by the organist's feet. The pedal department is basically like the manual departments but controls the longer pipes.

The organist sometimes wishes to combine the stops of two different manuals or to couple one or more of the manuals to the pedals. This is effected by a simple mechanism, called a coupler, that is controlled by a stop knob at the console (stops that control a set of pipes are called speaking stops).

Certain combinations of stops on each manual are more commonly needed than others; in order that these combinations may be readily available, the console is provided with a number of short pedals disposed above the pedalboard. Each of these pedals is connected to one commonly needed combination of stops. When a pedal is depressed, the stops connected to it are drawn on, and any others that are already drawn are pushed off. These pedals are called combination (composition) pedals.

In the simplest mechanical action, the connection from key to pallet is by a series of cranks and levers. The overall distance may be considerable, and the main distance is bridged by trackers, slender strips of wood, metal, or plastic, that always work in tension.

The trackers

The mechanism of the organ as described so far is entirely mechanical, and such organs are said to have tracker action. Tracker action is used in many modern organs, especially in Germany, The Netherlands, Scandinavia, and increasingly in the United States and Canada; many organists prefer it to all other forms because it is so direct and sensitive in response. Organs may, however, have pneumatic, direct electric, or electropneumatic action, although these actions result in a loss of touch and responsiveness. In very large organs with tracker action, considerable strength may be necessary to depress the keys. Also, where the layout of the building is inconvenient and the departments of the organ have to be widely separated, tracker action is not practicable. To overcome these difficulties, especially with the object of lightening the touch, other forms of action were devised.

The Barker lever

The first effective system was developed (after a device invented by David Hamilton of Edinburgh in 1833) by Charles Spackman Barker, an Englishman. It consisted of a series of small, high-pressure pneumatic bellows, or motors, one attached to each note of the main manual at

the console. When a note was depressed, compressed air was admitted to the motor, which, in turn, operated the tracker action. Lacking encouragement at home, Barker went to France, where the great French builder Aristide Cavaillé-Coll employed the Barker lever almost exclusively from 1840 on.

Tubular pneumatic action. Later, the trackers were supplanted by lead tubes, and the connection from key to pallet was solely by compressed air travelling through these tubes. This system was called tubular pneumatic action. At its best, it was remarkably effective, being reliable, long-lived, reasonably silent in action, and perfectly prompt in operation. At anything but its best, it was none of these things, and its worst fault usually lay in sluggish operation. Tubular pneumatic action is almost never used in modern times.

Electric action. As early as 1860, electric action was used experimentally, and in 1888 it was employed by the English builder Henry Willis at Canterbury Cathedral. His action remained in satisfactory use there for 50 years before it needed to be replaced. The modern type of electric action was pioneered by Robert Hope-Jones in Britain at the end of the 19th century. Direct electric action may be used, but a combination of electric and pneumatic mechanism is more general. In this system the depression of a key completes an electrical circuit, which energizes an electromagnet, allowing wind to enter a pneumatic motor attached to the wind-chest, and this motor opens the pallet. The stops may be operated in exactly the same way, but, where they are operated electrically, the sliders are often replaced by a series of valves, one to each pipe. The organ is then said to have a sliderless chest, and the most usual type is the pitman chest, so called because it contains a type of floating valve called a pitman. This action is commonly known as electropneumatic.

The pitman chest

The combination pedals can also be operated electropneumatically. They are usually supplemented by a series of buttons, or pistons, placed in the keyslips on each manual, where they are conveniently operated by the organist's thumbs. The pistons may easily be made adjustable so that the organist can quickly alter the combination of stops controlled by each one.

No electric action has yet lasted more than 50 years without needing a comprehensive rebuilding, and many have lasted for much shorter periods. But, with improvements in design and standardization of parts, it may be anticipated that rebuilding will become less frequent and expensive. On the other hand, there are small tracker-action organs working satisfactorily after 300 years, and even large ones have continued to operate for more than a century, despite almost total neglect.

Electric auxiliaries to tracker action

A compromise has been used successfully with tracker action for each department, with the coupler action operated electrically. This arrangement has considerable merit, since the coupling together of three or four manuals with tracker action results in a very heavy touch. Electric stop action may also be combined with tracker key action, enabling the use of electric (including solid-state) combinations—an invaluable aid, especially in larger instruments.

Tone production. The pipes are the most important part of an organ. There are two main categories: flue pipes and reed pipes.

Flue pipes. Flue pipes (made either of wood or metal; their construction is basically similar in principle) account for about four-fifths of the stops of an average organ. Figure 1 shows a front view and a vertical section of the most typical sort of metal flue pipe. The pipe consists of three main parts: the foot, the mouth, and the speaking length.

The pipe stands vertically on the wind-chest, and wind enters at the foothole. The foot is divided from the speaking length by the languid, a flat plate; the only airway connection between the foot and the speaking length is a narrow slit called the flue. The wind emerges through the flue and strikes the upper lip, producing an audible frequency, the pitch of which is determined by and amplified in resonance by the speaking length of the pipe. A pipe of this kind is, in fact, identical in principle with a recorder or a tin whistle; but, whereas they have holes along the speaking length, which the player covers and uncovers with his fingers to secure the notes of the musical scale, in an organ there is a separate pipe for each note.

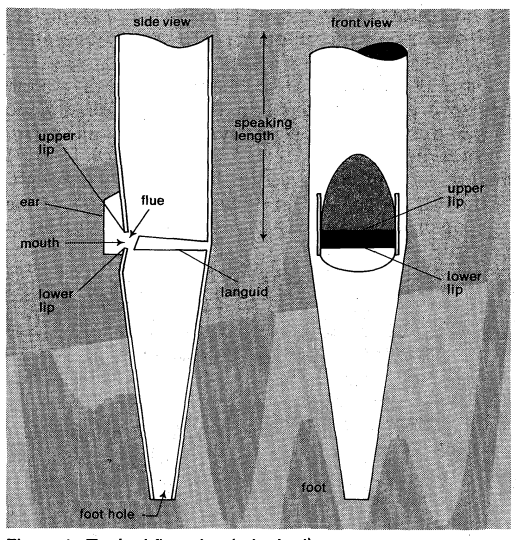


Figure 1: Typical flue pipe (principal).

The tone of a pipe is determined by many factors, including the pressure of the wind supply, size of foothole, width of flue, height and width of mouth, and the scale, or diameter, of the pipe relative to its speaking length. The material of which the pipe is made also exerts an influence; it may be metal (*i.e.*, an alloy of lead and tin), wood, or, more rarely, pure tin or copper, and for the bass pipes zinc. The pipes may also vary in shape, a common variant being an upward taper in which the pipe is smaller in diameter at the top than at the mouth. Or, the top of the pipe may be completely closed by a stopper. Such a pipe is said to be stopped; a stopped pipe sounds an octave lower in pitch than an open pipe of the same speaking length.

Open pipes of large diameter are said to be of "large scale," and open pipes of small diameter are said to be of "small scale." Large-scale pipes produce a dull or foundational quality of tone that is free from the higher harmonics (the numbered series of partials, or component tones). Small-scale pipes produce a bright quality of tone that is rich in harmonics. Stopped pipes can be particularly foundational in tone, and they favour the odd-numbered at the expense of the even-numbered partials. Tapered pipes are somewhere between stopped and open pipes in tone quality.

Tuning and voicing

Flue pipes are tuned by increasing or decreasing the speaking length. In the past, several methods of tuning were employed, but in modern times this is often done by fitting a cylindrical slide over the free end of the speaking length and sliding it up and down, lengthening or shortening the pipe as required. In stopped pipes the stopper is pushed farther down to sharpen the pitch or is pulled upward to lower it.

The pipe maker thus broadly fixes the type of tone that a pipe will produce; but this is further controlled within fairly wide limits by the wind pressure and, finally, by the voicer, who adjusts the tone of each pipe by manipulating the foothole, flue, and upper and lower lips. The attack of the note may also be greatly influenced by cutting a series of small nicks in the edge of the languid. Heavy nicking, so commonly practiced in the early 20th century, produces a smooth and sluggish attack. Light nicking or no nicking, as used up to the 18th century and in more advanced modern organs, produces a vigorous attack, or chiff, somewhat like tonguing in a woodwind instrument. This enhances the vitality and clarity of an organ. The voicer is the artist upon whom the ultimate success of any organ depends, although the tonal designer or architect is hardly less important. It is he who decides upon the choice of stops, their disposition in the organ, and the

scales to be followed by the pipe maker. A completely successful organ depends upon the effective cooperation of designer and voicer.

Reed pipes. Reed stops have beating reeds of a kind that finds several counterparts in the orchestra, and no doubt organ reeds were originally copied from instrumental prototypes.

The shallot, seen in cross section in Figure 2, is roughly cylindrical in shape, with its lower end closed and the upper end open. A section of the wall of the cylinder is

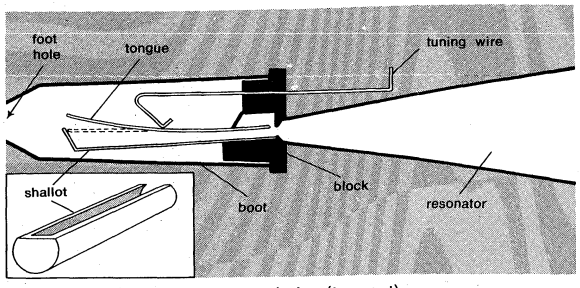


Figure 2: Section through a reed pipe (trumpet).

cut away and finished off to a flat surface, as shown in the inset to Figure 2. The slit, or shallot opening, thus formed is covered by a thin brass tongue that is fixed to the upper end of the shallot. The tongue is curved and normally only partially covers the shallot opening. But, when wind enters the boot, the pressure of the wind momentarily forces the tongue against the shallot, completely closing the opening. Immediately, the elasticity of the brass asserts itself, and the tongue reverts to its curved shape, thus uncovering the opening. This process is repeated rapidly. The frequency of the pulsations of air that enter the shallot is determined by the effective length of the reed and, in turn, determines the pitch of the note. Thence, the pulsations pass out into the tube, or resonator, which further stabilizes the pitch and decides the quality of the note. Most reed resonators have a flared shape, as shown in Figure 2. As in flue pipes, a wide scale favours a fundamental tone, and a narrow scale favours a bright tone. Cylindrical resonators produce an effect similar to that of stopped flue pipes, the note being an octave lower than the equivalent flared pipe and the tone favouring the odd partials. Some reed pipes, such as the Vox humana, have very short resonators of quarter or eighth length. Pipes the resonators of which have no mathematical relationship to the pitch are known as regals; regal stops were very popular in the 17th century, particularly with the North German school, and their use has been revived in modern times. Their short resonators have varying and peculiar shapes, which produce a highly characteristic snarling tone; they can be difficult to keep in tune.

Reed length and pitch

Reed pipes are tuned by moving the tuning wire, thus shortening or lengthening the tongue (Figure 2). As in flue pipes, the scale and shape of the resonator largely determine the quality of tone to be produced; but the wind pressure, shape and size of the shallot, and thickness and curvature of the tongue also have important influence. The tongues may also be weighted with brass or felt; this weighting produces a smoother quality of tone, especially in the bass notes.

Organ reeds have been referred to as beating reeds because the tongue is larger than the shallot opening and therefore beats against it. In a free reed, on the other hand, the tongue is smaller than the opening and so vibrates through rather than against it. Harmoniums and harmonicas have free reeds, which are almost never used in organs.

Choruses. It has already been explained that the pitch of any pipe is proportional to its length. Most modern organs have a manual compass of five octaves, from the second C below middle C to the third C above; an open pipe sounding the low C is about eight feet (2.5 metres) in speaking length (64 vibrations per second). The shortest pipe in the same rank is thus about three inches (eight centimetres) long (2,048 vibrations per second).

The most characteristic tone of the organ is produced by its diapason, or principal, stops. These are of medium scale (usually about 6-in. scale at the 8-ft open pipe) and moderate harmonic development—*i.e.*, neither particularly dull nor bright. Such a tone quality becomes boring if heard for a long time. Also, when greater power is required, there is a distinct limit to what can be done by adding more stops of unison pitch. From the earliest times, stops, especially the principals, were arranged in choruses, and the principal chorus is the very backbone of any organ; without a complete principal chorus, an organ is hardly worthy of the name.

Diapasons, or principals

A chorus consists of stops of roughly similar quality and power but at a great variety of pitches. A unison principal is known as Principal 8 ft because of its longest (8-ft) pipe, and the figure 8 appears on the stop knob or tablet (rocking tablets are often used in place of knobs with electric action) at the console to give an indication of its pitch. The first step toward a chorus is to add a stop sounding an octave above 8-ft ranks (*i.e.*, at octave pitch), the largest pipe of which is therefore four feet long. Next comes a 2-ft stop, while in the other direction the suboctave pitch may be represented by a 16-ft stop. The top pipe of a 2-ft stop has a speaking length of only three-quarters of an inch, and this is about the practical upper limit. Nevertheless, an organ with nothing higher in pitch than a 2-ft stop would be lacking in brilliance, especially in the lower parts of the compass.

From the earliest times, organs have, therefore, been supplied with what are known generically as mixture stops, which have several high-pitched pipes to each note. But, since, for example, a 1-ft rank could not be carried right up to the top note, it breaks back an octave at some convenient point in the compass. Ranks pitched even higher will break back more than once. Thus, in the bass, a mixture adds definition to the slow-speaking, low-pitched pipes; in the treble, where the small pipes tend to be lacking in power, it duplicates the unison and octave ranks. A mixture, therefore, helps to maintain a balance of power between bass and treble, while adding harmonious power of a kind that is completely peculiar to the organ and can be produced in no other way.

Mixtures and mutations

Mixture stops also contain ranks sounding at pitches other than in octaves with the 8-ft principal. In chorus mixtures these sound at a fifth above the unison (*e.g.*, G above C), although ranks sounding at a third above and even at a flat seventh (*e.g.*, E and B♭ above C) and their respective octaves are also found; but these are best restricted to mixtures intended for somewhat special effects. The theoretical justification for these quint- (fifth) and third-sounding ranks is that they reinforce the natural upper partials of the harmonic series, but they were included in organs long before this was understood. The fact is that they were found to sound well, and any attempts to build organs without mixtures and off-unison ranks have been completely unsuccessful. The colourfulness and vitality of any organ depend largely upon copious, artistically voiced mixtures.

Off-unison ranks are also available as separate stops, mostly sounding at an interval of a 12th (an octave and a fifth; 2⅔ ft), 17th (two octaves and a third; 1⅗ ft), or 19th (two octaves and a fifth; 1⅓ ft) above the unison. These are used melodically to colour the unison and octave stops, and they may be wide or narrow in scale. Such stops are known as mutation stops, as opposed to the mixtures, or chorus stops. Their use is essential for the historically (and therefore artistically) correct performance of organ music written before 1800 and of much modern music as well. After a period of disuse throughout the 19th century, they are again included in all modern organs that have any pretensions to being artistically competent.

HISTORY OF THE ORGAN TO 1800

Early history. The earliest history of the organ is so buried in antiquity as to be mere speculation. The earliest surviving record is of the Greek engineer Ctesibius, who lived in Alexandria in the 3rd century BC. He is credited with the invention of an organ very much on the lines of

The hydraulis

the single-manual, slider-chest organ already described, except for its wind supply, which made use of a principle that was most ingenious, though applicable only to a very small instrument. A piston pump supplied air through an ordinary clack valve to a reservoir; at its upper end, this reservoir communicated directly with the wind-chest. The reservoir, cylindrical in shape and with no bottom, was placed in a large drum-shaped container that was partly filled with water. As the reservoir became filled with air, the air would escape around its lower edge. In this way a more or less equal pressure of air was maintained inside the reservoir. Because of this arrangement the instrument was known as a hydraulis. A clay model of a hydraulis was discovered in 1885 in the ruins of Carthage (near modern Tunis, Tunisia), and the remains of an actual instrument were found in 1931 at Aquincum, near Budapest, Hungary.

The development of the organ during the early Middle Ages is obscure, but by the 8th or 9th century it was being used in Christian churches. In the 10th century the famous instrument in the cathedral at Winchester, England, was constructed, of which the monk Wulfstan left a much quoted but manifestly garbled description ending: "the music of the pipes is heard throughout the town and the flying fame thereof is gone out over the whole country."

The artistic history of the organ begins with the development of the chromatic keyboard (*i.e.*, having 12 keys per octave) in the late 12th and early 13th centuries. By 1361 the cathedral organ at Halberstadt, Germany, had three chromatic keyboards and pedals; the keys, however, were much wider than those of the modern keyboard. The modern size of keys was fairly generally established by the end of the 15th century. Although the Halberstadt organ had three manuals, it had no stop mechanism. The main keyboard controlled a huge mixture stop, and the other keyboards controlled reduced groups of stops.

Ctesibius' slider arrangement was probably rediscovered some time in the early 15th century, and it became common soon after 1450. Reed stops began to appear at the same time, and by 1500 the organ had reached a stage in northern Germany in which all the important features of the modern organ were present. Each department had separate choruses; stopped, tapered, and open flue pipes; mutation stops; and reeds. The North German organ builders continued to be pre-eminent until about 1700, when the southern German builders took the lead.

Portable organs

During the Middle Ages and the Renaissance, three diminutive forms of the organ were widely used. These were, first, the positive (in which category are included most chamber organs of the period), a small organ capable of being moved, usually by two men, either on carrying poles or on a cart. The second type, the portative, was smaller still, with only one set of pipes and a manual of very short compass. It was carried by the player and was supported by a strap around his neck. He worked the bellows with one hand and played the keys with the other. Such instruments were used in processions and possibly in concerted instrumental ensembles. In between the last two in size was the third type, the regal, which usually had only one reed stop, a regal, as previously described.

Since national styles of organ building vary widely and it is necessary to know something about them before the music of each nation can be performed intelligently, the more important styles must next be considered briefly. Of the basic medieval organ, prior to the development of national styles, little if any material survives, except in the old cathedral at Sion in Switzerland, where a large proportion of the seven-stop organ appears to date from about 1400. Although voiced on very low wind pressure, the tone of the chorus is brilliant, colourful, and amazingly powerful. Not much is known about the precise uses of church organs in the Middle Ages. The organ hardly began to possess a literature of its own before the last portion of the 15th century.

Italy. Italy is mentioned first because its organs developed to their maturity soon after 1500 and remained relatively unaltered until about 1800. The Italian organ had one manual and usually only an octave of pedal keys, which had no pipes of its own (except an occasional independent 16-ft *contrabasso*) but was coupled permanently to the manual. The manual chorus (*ripieno*) had the peculiarity that there was no collective mixture; all the ranks were drawn by separate stops. Each rank broke back an octave as it reached the 1½-in. pipe. In addition, there were flute stops of 4-ft, 2 ⅔-ft, and 2-ft pitch and a register called the *fiffaro* or *voce umana* (not to be confused with the French *voix humaine* or German *Vox humana*, which are regals), a principal rank found only in the treble and tuned sharp so that when it is played together with the *principale* one hears an audible beat. It was the forerunner of the similarly constructed *voix céleste* stop popular in the 19th-century romantic organ. The scale of the classic Italian *principale* was not much different from its counterpart in the north, but its mouth was narrower, its voicing more delicate, and there was a notable lack of chiff. Reeds were not found until late in the 16th century and were never considered essential. There are well-preserved 16th-century instruments surviving, especially in Brescia and Bologna.

These simple resources were adequate for the performance of the keyboard works of Andrea (*c.* 1520–86) and Giovanni (*c.* 1556–1612) Gabrieli and Girolamo Frescobaldi (1583–1643). Organ music of these men and their contemporaries was not clearly differentiated from that for harpsichord, as indicated by the collections *"per organo o cembalo."* The organ enjoyed some popularity: in Rome, according to Baini, no less than 30,000 people flocked to the square of St. Peter's hoping to find entrance to the cathedral to hear Frescobaldi's magic organ playing.

Spain and Portugal. The Iberian organ followed the Italian tradition, but, later, many reeds were added, most notably the *trompetas reales* ("royal trumpets") and other horizontal (*en chamada*) reeds arrayed in fanlike projections from highly ornamental cases. These reeds were on extremely low wind pressure and achieved amazingly full sounds that filled the huge edifices.

The divided manuals of Iberian organs

Like their Italian counterparts, Spanish and Portuguese organs had only a few rudimentary pedals. The manuals, however, were divided, with notes up to middle C controlled by a draw knob to the left and notes up from C sharp by a draw knob to the right. This enabled the playing of a solo voice against an accompaniment on the same manual.

A unique feature of Iberian churches was the presence of several separate and distinct instruments. The 13th-century Toledo Cathedral, for example, houses three organs: one over the Lion's Gate (south transept), known as the Emperor's organ (*c.* 1543), in an elaborate stone case that now houses a later 18th-century instrument; on both sides of the choir are separate organs—the Epistle organ from 1758 and the Gospel organ completed by the builder José Verdalonga in 1791. These two instruments have horizontal reeds projecting from both sides: into the choir and into the side aisles, enabling interesting uses of antiphony, or contrasting masses of sound. Toledo's three are surpassed in number at Mafra, Portugal, by the six organs in the monastic church, located in a sumptuous building considered Portugal's national monument. The reasons for multiple instruments have never been entirely clear. The use in services of different altars may have dictated the course, or perhaps when one instrument fell into disrepair it may have seemed simpler to add another; at least numerous organs have not been in playable condition for many years.

The leading composers of this era wrote for the instruments at hand. Their music is exciting on these instruments but is seldom effective elsewhere. Likewise, northern European literature is not satisfactory on the Iberian instrument.

In Spanish cathedrals the custom prevailed of alternating verses between organ and singers at the liturgical offices (Matins, Lauds, and Vespers) as well as for the ordinary (Kyrie, Gloria, Sanctus, and Agnus Dei) of the mass. The organ verses were invariably improvised, and the organist could change his registration during the choir verses between. Because improvisation was the

norm, comparatively little written music exists from this period.

Antonio de Cabezón (1510–66) is the most illustrious of the historic Spanish composers. Others include Sebastián Aguilera de Heredia (1570–?) and Juan Cabanilles (1644–1712). Carlos Seixas (1704–42) is known as the Portuguese Bach.

Germany. From 1500 to 1800 Germany led the world in organ building and the composition of organ music. The organ builders reached the peak of their achievement about 1700 in the work of Arp Schnitger. His was the organ of the high Baroque; but his countrymen Andreas and Gottfried Silbermann were equally the masters of the slightly later, more sophisticated style of the mid-18th century.

The Schnitger organs

Schnitger made organs with four manuals, pedals, and as many as 60 speaking stops, but he made some instruments with less than 30 speaking stops that are capable of dealing with the whole pre-Romantic repertory. The finest surviving examples in this size are at Steinkirchen, near Hamburg, and at Cappel, near Cuxhaven. Two great larger examples are at Zwolle and Alkmaar, The Netherlands, both restored to excellent condition in the mid-20th century.

Seventeenth- and 18th-century German organs were usually constructed on *Werk*-principle lines: each department of the instrument, or *Werk,* was separately cased, the Hauptwerk (main manual) in front of and above the player, with the Pedals at each side and the Positiv (auxiliary manual) behind on the gallery railing. Each department, including the Pedal, had its own principal chorus, complete up to at least one mixture. All departments were roughly equal in power but varied in pitch, having, respectively, a 16-ft, 8-ft, and 4-ft preponderance (and 32-ft and 2-ft as well in larger instruments). Each manual department had a set of flutes and mutations that could be combined in a variety of ways to provide accompaniment and melody or the balanced but contrasting tone qualities essential for duet and trio passages. Although the pedal department consisted mainly of its principal chorus, it could be coloured for solo and obbligato passages by 2-ft flute and reed stops. The reeds were not much louder than the flue stops, and the Pedal 16-ft and 8-ft reeds were frequently drawn with the principal chorus for improved definition. When used in this way, they by no means caused the Pedal to overwhelm the Hauptwerk. Such an instrument could deal with the requirements of all 15th- through 18th-century organ music, although its limited supply of manual reeds placed it at some disadvantage in French music of the period.

The earliest organ music consisted of simple arrangements of vocal and instrumental music. A significant development in ornamented versions of such compositions for other mediums is the *Fundamentum organisandi* (*Fundamentals of Organ Playing*; Nürnberg, 1452) by Konrad Paumann, organist at the court of Bavaria. By the end of the 15th century, the polyphonic style was establishing itself in an independent form for the organ as a separate and distinct instrument. Pedal notes were ideally suited for a long-note cantus firmus ("fixed song," a basic melody chosen, for example, from a hymn) against manual counterpoint. The great Amsterdam organist Jan Pieterszoon Sweelinck (1562–1621), having studied in Venice (a tradition already established from northern Europe), developed a notable school of organ playing and composition that dominated a significant part of the musical world for almost two centuries, culminating in the genius of J.S. Bach (1685–1750). Among its composers were one of Sweelinck's greatest pupils, Samuel Scheidt of Halle; in the direct line from Sweelinck to Bach, Johann Adam Reinken of Hamburg; and the greatest composer of the high Baroque and one of the chief influences on Bach, Dietrich Buxtehude of Lübeck. In Austria and south Germany, a pupil of Frescobaldi, Johann Jakob Froberger of Vienna, was noted for his improvisational style. Johann Pachelbel of Nürnberg was the agent by which Italian and south German influences were again transported to northern Germany. By the time of Bach, the written *Choral-Vorspiel,* or chorale prelude (as opposed to the improvised prelude to the chorale within the Lutheran service), prelude and fugue, toccata, passacaglia, chaconne, and trio were firmly established forms for the organ—probably never again to be surpassed. A musical critic said of Bach: "He is the spectator of all musical time and existence, to whom it is not of the smallest importance whether a thing be new or old, so long as it is true."

Sweelinck and his successors

France. As far as the manual departments are concerned, French organs differed little from the German type, but the principal choruses were generally larger in scale. The separate, large-scaled Tierce (1⅗ ft) was also universal, and there were many cornet stops. These mixture stops consisted of five pipes to each note: a stopped unison (8 ft) and large-scale open 4 ft, 2⅔ ft, 2 ft, and 1⅗ ft. They extended only from middle C upward and were largely melodic in use. They were never drawn with the principal chorus (Plein Jeu) but generally were used with the reed chorus (Grand Jeu). Apart from this, the Plein Jeu, Grand Jeu, and Jeux de Mutation were seldom or never intermixed in French music.

The cantus firmus pedal

The pedal department of the French organ prior to 1700 was regarded largely as a sort of solo cantus firmus section that consisted usually of only 8- and 4-ft flutes and 8- and 4-ft trumpets. Only in the largest 18th-century organs were 16-ft stops included, although there were often as many as three on the Grand Orgue (the manual analogous to the German Hauptwerk and the English Great Organ). When French organs had more than two manuals (Grand Orgue and Positif), the others (Récit and Écho) were usually of short compass; but if, as sometimes, there was a fifth manual, it was a Clavier de Bombardes, consisting of 16-, 8-, and 4-ft trumpets and a cornet. Unlike its German counterpart, the main case housed all divisions except the Positif, which was in its usual location on the gallery railing.

French organs were notable for their reeds, and the highly stylized French music of the 17th and 18th centuries calls for their frequent use. Surviving specimens in good order are rare; but unaltered, late 18th-century, four-manual organs survive at Poitiers Cathedral (by the noted builder François-Henri Clicquot) and at Saint-Maximin, Provence (by Jean-Esprit Isnard).

Jean Titelouze (1563–1633) of Rouen may be considered the father of organ music in France. Nearly four centuries later the musicologist André Pirro and the composer Alexandre Guilmant say of his *Hymnes de l'église* (*Hymns of the Church*; 1623) in their preface to Guilmant's edition of the works of Titelouze, "He writes continued dissonances in the modes' natural scales and his use of sevenths gives, sometimes, a quite modern character to his modulations." A great tradition of playing and composing developed with a single dynasty presiding at a single church, Saint-Gervais in Paris, for more than 150 years: the Couperin family, including Louis, François, Charles, and culminating in François le Grand (1668–1733), whose two famous masses are repertory staples even today. Contemporaries included the Paris organist Nicolas Lebègue and his pupil Nicolas de Grigny of Reims, whose mass and hymn verses Bach copied in his own hand; Louis Clérambault, known for two modal suites (*i.e.*, using the tonal structure of the old church modes rather than that of major and minor keys); and Louis-Claude Daquin, a great favourite in Paris who played his popular *Noëls* at the royal chapel.

Great Britain. British organs before the Commonwealth (1649–60) seem to have been very immature. Only a very few had two manuals, and none had pedals. Mixtures and reeds seem to have been unknown, and mutations were restricted to a single twelfth.

After 1660 a new school rapidly grew up, and, although the two principal builders had both been abroad during the Commonwealth (Bernard Smith in Germany or Holland and Renatus Harris in France), their British work owed little to foreign influence. Only the Great Organ had a complete diapason chorus, and the Choir, or Chayre, organ usually extended upward only to a single 2 ft. Almost every organ had a Cornet, and the reeds in common use were Trumpet, Vox humana, and Cremona,

or Krummhorn, with half-length, cylindrical resonators. There were no pedals, but the manual compass almost invariably extended to the third G below middle C. If there was a third manual, it consisted of a short-compass Echo department in which all the pipes were shut up in a box to produce the echo effect. In 1712 the builder Abraham Jordan first fitted the echo box with shutters that were controlled by a pedal at the console; this arrangement produced what Jordan described as the swelling organ, but it was not to reach its full development until 150 years later; no 18th-century organ music demands a swell box. There are hardly any surviving examples of British instruments of this period in original condition, and the only one of any size is the 14-stop, two-manual organ at Adlington Hall, near Macclesfield, dating from the last quarter of the 17th century. It is possibly the work of Bernard Smith. It is entirely original and was restored to perfect order in 1959.

Such instruments were adequate for the music of John Lugge, John Blow, Henry Purcell, George Frideric Handel, John Stanley, and even early 19th-century composers, such as Samuel Wesley.

DEVELOPMENTS AFTER 1800

The romantic organ. Because of the increasing interest in orchestral and operatic music, the organ fell out of favour during the 18th century, and by 1800 it survived only as an ecclesiastical drudge. From the middle of the 19th century, however, a revival took place under the leadership of two great builders, Aristide Cavaillé-Coll of France and Henry ("Father") Willis of England. The German Edmund Schulze, who brought to England an organ built by his father's firm in central Germany, was also influential, especially in his flue choruses. In Britain during the first half of the 19th century, the introduction of pedals under the influence of Henry John Gauntlett made it possible for the first time to play the organ music of Bach and his German contemporaries and predecessors. While retaining respectable vestiges of the classical chorus, Cavaillé-Coll and Willis developed the solo stops, especially reeds, and Willis, in particular, provided new aids to registration.

Influence of Cavaillé-Coll

France. The work of Cavaillé-Coll was directly responsible for a new school of organ composition in France: César Franck (1822–90) composed chromatic and grandly Romantic works, culminating in his *Trois Chorals* (*Three Chorales*); also notable, although of less exalted stature, were Camille Saint-Saëns (*Three Rhapsodies on Breton Themes* and an important organ part in his *Symphony No. 3 in C Minor*), Alexandre Guilmant (numerous sonatas and small "service" pieces), and Charles-Marie Widor and Louis Vierne, both of whom wrote a number of large-scale "sonata suites" they entitled *Symphonies.* A successor of Franck at his post at Sainte-Clotilde in Paris was the great improviser Charles Tournemire, who left a highly individual collection of improvisatory-like modal pieces entitled *L'Orgue mystique.* Jean Langlais (born 1907) continued the modal tradition in many descriptive pieces for liturgical use, and Olivier Messiaen (born 1908) has perpetuated the mystical qualities in suites of great originality (*La Nativité du Seigneur, L'Ascension,* and others). Another prominent organist-improviser-composer-teacher in this largely Romantic mold was Marcel Dupré. Francis Poulenc contributed a popular *Concerto in G Minor for Organ, String Orchestra, and Timpani.*

Germany and Great Britain. Parallel to the French movement, German-born Felix Mendelssohn (1809–47), though working primarily in England, revived an interest in the works of Bach and composed sonatas and preludes and fugues for the organ. Franz Liszt and the German Julius Reubke added to the repertory several flamboyant, pianistic pieces. Johannes Brahms wrote chorale preludes and several preludes and fugues. A composer of monumental chromatic fantasies and fugues, as well as simple chorale preludes, was Max Reger (1873–1916).

Organists found that they could play effective arrangements of orchestral music on the new romantic style organ. Since orchestral music was popular and respectable orchestras very rare and other forms of public entertainment even more so, the organ suddenly regained an immense popularity hardly rivalled by that of the 17th and 18th centuries, when it was the acknowledged "king of instruments." Organ builders naturally responded by making their instruments increasingly orchestral in character, culminating at the end of the 19th century in the work of the English builder Robert Hope-Jones, who entirely abandoned the chorus and mutation stops and relied instead upon diapasons of vast scale on heavy-pressure wind, with reeds to match, backed up by huge-scaled flutes, tiny-scaled string stops (with keen-sounding flue pipes), and powerful stops of his own invention called diaphones. Hope-Jones emigrated to the United States, and, although a semblance of classical design returned to England soon after 1900, his influence continued to be felt throughout the first half of the 20th century. This discredited the organ as a musical instrument in the eyes of serious musicians and composers.

First instruments in the Colonies

The United States and Canada. The first organs in America had been imported from England beginning about 1700. This was the period of the English Commonwealth, and the Puritan view of the "unsuitability" of an organ in church was inherited by the colonies. Only parishes of the Church of England (later known as Protestant Episcopal Church) and Lutheran and Moravian churches in Pennsylvania would admit instruments. Another century elapsed before the New England Puritans did likewise. The only builder of note was the German-American David Tannenberg. A U.S. school of builders began to emerge in the early 1800s with such names as Henry Erben, Elias and George Hook, George Jardine, William A. Johnson, J.H. and C.S. Odell, and Hilborne and Frank Roosevelt. Perhaps the inevitable end of the U.S. "romantic" era was reached in Ernest M. Skinner, who lived until the middle of the 20th century. In Canada, Joseph Casavant built his first organ in Quebec province in 1837. Two of his sons visited France in 1878–79 and brought back to North America the Cavaillé-Coll tradition.

Although no significant school of organ composers has emerged in North America, two names, or at least two pieces, stand out: although neglected for years, Charles Ives in 1891 improvised (and later wrote down) *Variations on America*; this piece was remarkably ahead of its time and suits the U.S. organ at the turn of the century most admirably. Healey Willan, though British born, was for many years Canada's leading composer; his greatest work for the organ is probably *Introduction, Passacaglia and Fugue* (published 1919).

Organ revival. *The neoclassical movement.* Albert Schweitzer, organist, philosopher, and later medical missionary, wrote a booklet, *Deutsche und französische Orgelbaukunst und Orgelkunst* ("The Art of German and French Organ Builders and Players"), in 1906 outlining the inadequacies of the 19th-century organ for the performance of Bach and his contemporaries. It was not until 1926, however, with Karl Straube, that the revival began. Straube, organist at Bach's Thomaskirche in Leipzig and noted recitalist, teacher, editor of Baroque organ works, and leading exponent of the very Romantic works of Reger, renounced his whole approach to the organ and called for a return to the instrument of Schnitger and the high Baroque. Since then, the movement has spread among such organ builders as Karl Kemper, Rudolph von Beckerath, and Johannes Klais in Germany; Victor Gonzalez in France; Sybrand J. Zachariassen and Th. Frobenius in Denmark; Dirk A. Flentrop in The Netherlands; Th. Kuhn in Switzerland; and Walter Holtkamp, G. Donald Harrison, and Herman Schlicker in the United States. Many organists interested in fine phrasing and articulation feel that these qualities are only realizable through the medium of tracker action, where there is a direct connection between player and pallet. This neoclassical movement has inspired a few, but significant, composers to write for the organ: Willem Vogel of The Netherlands; Finn Viderø of Denmark; the Germans Paul Hindemith (three sonatas and two concerti), Hugo Distler, Hermann Schroeder, and Helmut Walcha; and the Americans Daniel Pinkham and Alan Stout.

Exemplary modern builders

The eclectic movement. The revival in France, the United States, and Britain strove to produce an instrument that could do equal justice to all legitimate organ music of whatever period. This was not easy, but it was possible. Undoubtedly, the most successful exponent up to about 1950 was G. Donald Harrison (British born but associated with the Aeolian-Skinner Organ Co. in Boston)—*e.g.*, the instrument in the Mormon Tabernacle in Salt Lake City, Utah. In England, successful examples are the London organs of the Royal Festival Hall and St. Giles Cripplegate.

BIBLIOGRAPHY. AUSTIN NILAND, *Introduction to the Organ* (1968), is, as its name implies, a good introduction to the subject from a contemporary British point of view; in a similar vein, very practical, but from a continental approach, is HANS KLOTZ, *Das Buch von der Orgel* (1938; 7th ed., 1965; Eng. trans., *The Organ Handbook*, 1969). The classic work on the subject in English is EDWARD J. HOPKINS and EDWARD F. RIMBAULT, *The Organ*, 3rd ed. (1877, reprinted 1965); less complete, but general, more contemporary coverage is given in WILLIAM LESLIE SUMNER, *The Organ*, 3rd ed. (1962). Primarily historical is an extremely detailed work by JEAN PERROT: *L'Orgue de ses origines hellenistiques à la fin du XIII^e siècle* (1965; Eng. trans., *The Organ from Its Invention in the Hellenistic Period to the End of the Thirteenth Century*, 1971). PETER WILLIAMS, *The European Organ, 1450–1850* (1966), is a thorough history of continental organs. POUL-GERHARD ANDERSEN, *Orgelbogen* (1956; Eng. trans., *Organ Building and Design*, 1969), emphasizes architecture and the organ's relation to it. National schools are discussed in CECIL CLUTTON and AUSTIN NILAND, *The British Organ* (1963), post-revival; NOEL A. BONAVIA-HUNT, *The Modern British Organ* (1947), pre-revival; FENNER DOUGLAS, *The Language of the Classical French Organ* (1969); and WILLIAM HARRISON BARNES, *The Contemporary American Organ*, 8th ed. (1965), heavy on mechanics with drawings. JOSEPH EDWIN BLANTON, *The Organ in Church Design* (1957), very extensive and aimed primarily at architects; and *The Revival of the Organ Case* (1965), are primarily pictorial works on organ cases. See also the classic pen and ink drawings of ARTHUR G. HILL, *The Organ Cases and Organs of the Middle Ages and Renaissance* (1883, reprinted 1966; German trans., *Vierzig Orgelgehäuse-Zeichnungen*, 2nd ed., 1964)—many of these drawings appear in the other books listed. A fascinating book on portatives, positives, and regals is MICHAEL WILSON, *The English Chamber Organ: History and Development, 1650–1850* (1968).

(C.Cl.)

Organic Halogen Compounds

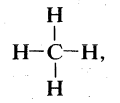
Organic halogen compounds are substances containing atoms of one or more of the so-called halogen elements (fluorine, chlorine, bromine, and iodine) joined to atoms of carbon. In their physical properties and chemical behaviour they are much like other organic compounds (compounds of carbon). Since very few organic halogen compounds occur in nature, most are products of the organic chemist's laboratory. Many organic halogen compounds are used as solvents; carbon tetrachloride, for example, is employed as a cleaning fluid. Other members of the family are used as anesthetics, refrigerants, and propellants for aerosols. The plastic material polyvinyl chloride is an organic halogen compound; so are the insecticide DDT (dichlorodiphenyltrichloroethane) and the herbicide 2,4-D (2,4-dichlorophenoxyacetic acid). Many organic halogen compounds are used in industry or in the laboratory for conversion to other compounds with useful properties, such as dyes and medicinal agents.

GENERAL CONSIDERATIONS

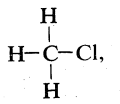
Hydrocarbon derivatives

Chemically, organic halogen compounds may be considered to be derivatives of hydrocarbons (compounds composed exclusively of carbon and hydrogen), in which certain of the hydrogen atoms have been replaced by halogen atoms. In the molecules of hydrocarbons, the carbon atoms form a framework, or skeleton, to which the hydrogen atoms are attached. When the halogen atoms replace hydrogen atoms, they too occupy peripheral or terminal positions, leaving the central framework positions in the molecule to the carbon atoms.

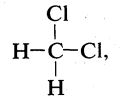
In theory, successive replacement of hydrogen atoms with halogen atoms is possible in any hydrocarbon molecule, until all hydrogen atoms have been replaced. The simplest hydrocarbon, for example, is methane, the molecules of which are composed of one carbon atom and four hydrogen atoms. (A carbon atom usually is able to combine with four other atoms; or, in chemical terminology, it has a valence of four. A hydrogen atom, however, can combine only with one other atom, or has a valence of one.) The structural formula of methane is



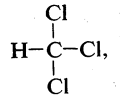
in which the letter C represents the carbon atom, the H's represent hydrogen atoms, and the lines stand for the bonds that hold the atoms together in the molecule. (The methane molecule also may be represented by the simpler, molecular formula CH_4, which simply notes the kind and numbers of the atoms.) Replacement of one of the hydrogen atoms in methane with a chlorine atom (Cl)—all halogen atoms also having valences of one—gives the substance methyl chloride



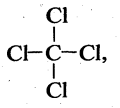
or CH_3Cl. Further replacements of hydrogen with chlorine give methylene dichloride



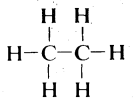
or CH_2Cl_2; chloroform



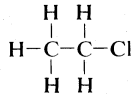
or $CHCl_3$; and carbon tetrachloride



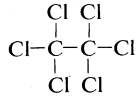
or CCl_4. Similar replacements can be carried out in molecules of the hydrocarbon ethane



(C_2H_6) to give a series of compounds ranging from ethyl chloride



(C_2H_5Cl) to hexachloroethane (C_2Cl_6).



Atoms of each of the other halogen elements—fluorine (F), bromine (Br), and iodine (I)—can similarly be substituted for hydrogen atoms in hydrocarbons (giving compounds containing atoms of more than one halogen element, as well as compounds with multiple atoms of any one element). When the precise nature of the halogen atom is not important, it is customary to use the symbol X, which can stand for an atom of any halogen element. Thus CH_3X could mean methyl fluoride (CH_3F), methyl chloride (CH_3Cl), methyl bromide (CH_3Br), or methyl iodine (CH_3I).

Although organic halogen compounds can be consid-

ered to be derived formally from hydrocarbons by replacement of hydrogen atoms with halogen atoms, this often is not the procedure by which organic halogen compounds are actually prepared in the laboratory. As will be shown, a variety of synthetic procedures is available for preparing organic halogen compounds from other classes of organic compounds.

Nomenclature and classification. Organic halogen compounds, generally, are named according to one of two principles, substitution or the use of organic radicals. In substitution names, the combining form of the name of the halogen element is used as a prefix to the name of the parent hydrocarbon. In this system the compound CH_3Cl is called chloromethane. The analogous combining forms of the other halogen elements are fluoro-, bromo-, and iodo-; halo- is the combining form of the generic term halogen.

Systems of naming

In the "radical" system of nomenclature the organic portion of the molecule (everything but the halogen atom) is designated by a combining form, or radical, which is followed by the name of the halogen in its -ide form (*e.g.*, chloride for chlorine). The organic radicals are derived from the names of the hydrocarbons by substituting the ending -yl for a portion of the hydrocarbon name. Thus methane becomes methyl, and the substance designated chloromethane above is also termed methyl chloride. Often, when it is desired not to specify one particular radical, the letter R— is used. In this way R—Cl represents any organic chlorine compound, and the designation R—X stands for the entire class of organic monohalogen compounds.

Almost any class of organic chemical compound—including alcohols, aldehydes, ketones, and carboxylic acids—can carry halogen atoms in its molecules. The phrase "organic halogen compounds," however, is usually taken to refer simply to halogen derivatives of the various hydrocarbons.

The hydrocarbons themselves fall into various categories depending on the nature of the bonds or linkages between the carbon atoms, and these categories are reflected in the classification of their halogen derivatives—*i.e.*, the organic halogen compounds. Generally hydrocarbons are classified as saturated, or paraffinic, if they contain only single bonds between carbon atoms, and as unsaturated if they contain multiple bonds between carbon atoms (a situation that arises when the molecule does not contain enough hydrogen atoms to satisfy all the valences of the carbon atoms, so that these must form multiple bonds between one another). These multiple bonds may be double or triple bonds, depending upon whether they are the equivalent of two or three single bonds, respectively. A further category of hydrocarbon is called "aromatic"; in their simplest form these are compounds with cyclic systems of six carbon atoms joined by alternating double and single bonds. Saturated hydrocarbons are called alkanes, and the organic halogen compounds derived from them by replacing their hydrogen atoms with halogen atoms are classified as alkyl halides. Hydrocarbons with double bonds are referred to as alkenes, and the halogen compounds derived from them are called alkenyl halides. Triply bonded hydrocarbons are called alkynes, and organic halogen compounds with triple bonds are alkynyl halides; and finally, aromatic hydrocarbons are arenes, and their halogen derivatives are aryl halides.

General chemical properties. The chemical behaviour of chemical compounds is largely determined by the electronic configuration (arrangement of electrons) of their molecules, and this in turn depends upon the electronic configurations of the constituent atoms. The electronic configuration of halogen atoms is quite different from those of carbon and hydrogen atoms, with the result that the incorporation of halogen atoms into hydrocarbon molecules has a marked effect on their chemical behaviour.

Atoms contain varying numbers of negatively charged electrons in what may be defined crudely as concentric shells about the positively charged nucleus. The atoms of the halogen elements, as a group, are characterized by the fact that their outermost electron shells include seven electrons. The halogen elements are differentiated from one another by the number of electron shells, and consequently by the total number of electrons they contain. These numbers increase with the atomic weights of the halogen elements, in the following order: fluorine, chlorine, bromine, and iodine. The seven electrons in the outermost shells, the so-called valence electrons, are farther and farther from the nucleus as the number of shells increases (that is, in the order given above). (The atomic radius of the iodine atom, for example, is rather more than twice that of the fluorine atom.) Consequently the valence electrons are most firmly held by the fluorine nucleus and least firmly held by the iodine nucleus. Because of its small size, fluorine is the most electronegative (electron-attracting) of the halogens, and iodine is the least. In general, less strongly held electrons, such as those of iodine, are more polarizable; *i.e.*, more easily distorted by attractive or repulsive forces near them. Polarizability is an important property that greatly influences the behaviour of halogen atoms with regard to organic molecules, as for example, when free halide ions (halide atoms containing an extra electron, and hence a net negative charge) attack positive centres in organic compounds (as occurs in the synthesis of many organic halogen compounds). In such "nucleophilic" (or positive-charge-seeking) attack, the halide ions attack preferentially in the decreasing order: iodide, bromide, chloride, and fluoride. This explains why hydriodic acid reacts more readily with alcohols to give alkyl halides than do the other halogen acids.

Because carbon is less electronegative than any of the halogen elements, the halogen atom attracts the two electrons that comprise a carbon–halogen bond more strongly than does the carbon atom. As a result, the electrons of the bond are displaced toward the halogen atom, and the bond is polarized. The carbon atom concerned will thus bear a partial positive charge, and will be susceptible to attack by nucleophiles—reagents that seek positive centres. Typical nucleophiles are negatively charged groups, such as the hydroxide ion, a negatively charged group comprised of a linked hydrogen and oxygen atom, OH^-, and molecules having atoms with lone pairs of electrons, such as ammonia, $:NH_3$ (in which the two dots before the N represent a lone pair of electrons). The result of nucleophilic attack on an organic halide is substitution (or replacement) of the halide by the nucleophile. The equation below represents a nucleophilic substitution reaction in which the iodine atom in methyl iodide is replaced by the hydroxide ion acting as a nucleophile:

Nucleophilic substitution

$$HO^- + \overset{\delta+}{C}H_3 - \overset{\delta-}{I} \longrightarrow HO - CH_3 + I^-$$

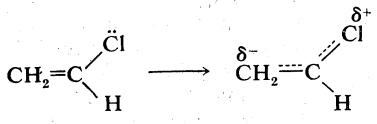
As is usually the case in chemical equations the starting materials are written at the left of the arrow and the products of the reaction are written at the right. In this equation, the polarization of the methyl iodide molecule is indicated by the symbols $\delta+$ and $\delta-$, in which δ (delta) means that partial charges (with signs as shown) appear on the atoms in question.

The readiness with which such nucleophilic substitutions occur depends on two factors: (1) the ease of dissociation of the carbon–halogen bond and (2) the nature of the leaving group (the halide ion). The ease of dissociation of a bond is inversely proportional to the bond energy, which for a carbon–halogen bond increases in the order: iodine, bromine, chlorine, and fluorine. A good leaving group, in general, is the anion of a strong acid, and the strengths of the halogen acids decrease in the order hydrogen iodide, hydrogen bromide, hydrogen chloride, and hydrogen fluoride. The bond dissociation energies and the nature of the leaving groups thus act in the same direction for the halides, with the result that the susceptibility of the various alkyl halides to nucleophilic attack decreases in the order: iodide, bromide, chloride, and fluoride.

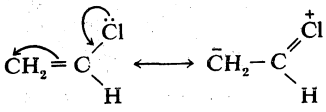
Organic halogen compounds also undergo free-radical reactions, reactions that lead to splitting of a bond, one electron going with each fragment. In such reactions,

only the bond dissociation energies are significant, and, in this case, the order of reactivity of the alkyl halides is the same as it is in nucleophilic attack.

When a halogen atom is attached to an unsaturated carbon atom (one in which not all the valence bonds are saturated), as in certain alkenyl halides, such as vinyl chloride, $CH_2{=}CHCl$, the halogen atom still attracts electrons to it, because of its electronegativity, but another factor must also be considered. In vinyl chloride, for example, the orbital (position in space) of one of the lone pairs of electrons of the chlorine atom overlaps with the orbital of certain of the electrons associated with the double bond. These are the so-called pi electrons, two of which (one from each carbon atom) occupy an orbital covering both carbon atoms and form one of the bonds (called a pi bond) uniting the two carbon atoms. These pi bonds occur only in multiple bonds. As a result of the interaction of the chlorine electrons with the pi electrons of the double bond, carbon–chlorine bonds of this particular type themselves have some double-bond character. Since this involves a partial loss of electrons from chlorine and a gain by the double bond, the molecule becomes polarized as shown in the equation below:



in which the dotted lines indicate partial double bond character. This phenomenon also can be described in a different way as a movement of the lone pair of electrons associated with the halogen atom, as shown in the diagram below by a curved arrow. This shift of electrons causes a polarization of the double bond. The actual state of the molecule is between the two extreme forms shown below, connected by a double-headed arrow:

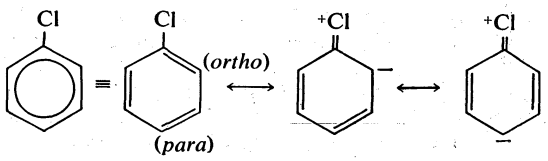


As a result of its interaction with the double bond, the halogen atom of a vinyl halide is much less susceptible to nucleophilic substitution than is an ordinary alkyl halide.

The presence of halogen atoms adjacent to a double bond—the situation in vinyl halides—also has an effect on the chemical reactivity of the double bond. Generally, one of the principal reactions of double bonds is electrophilic addition—a reaction that is initiated by an electron-seeking (electrophilic) reagent and leads to saturation of the double bonds, that is, to production of a molecule with nothing but single bonds. Electrophilic addition is strongly influenced by the polarization of the double bond caused by the halogen atom. As indicated above, a halogen atom on one carbon atom of a double bond leads to a partial negative charge on the other atom of the double bond. This negative charge is the target of an electrophilic reagent, so that the direction of electrophilic addition is dictated by the halogen atom.

When there are several halogen atoms attached to the double bond, their electron-attracting powers become more evident; in fact, they may reduce the negative character of the double bond to such an extent that it adds electrophilic reagents very much less readily than does a double bond without halogen substituents. At the same time, however, it may become more reactive toward nucleophilic reagents (which attack positive centres).

The same situation arises in the aryl halides. In this case, however, there is a more extensive unsaturated system, which can polarize in several different ways, as shown below:



In these diagrams, the hexagon represents a ring of six carbons (carrying single hydrogen atoms except where the chlorine atom is shown); the aromatic nature of the ring system is indicated either by a circle within the hexagon or by three double bonds (the two forms being connected by an identity sign). Forms in which separation of charges occur (as shown by the plus and minus signs) are connected by double headed arrows indicating that they contribute to the overall structure of the molecule. Because this polarization produces partial negative charges at the positions in the ring designated *ortho* and *para* (from Greek roots chosen somewhat arbitrarily before the correct positions were known), electrophilic substitution (*i.e.*, attack by positively charged groups) occurs preferentially at these positions. In aromatic compounds of most varieties, substitution reactions occur more readily than addition reactions because addition destroys the aromaticity of the ring, whereas substitution does not. In polyhalogenated aromatic derivatives, the electronegative halogen atoms reduce the susceptibility of the ring to electrophilic substitution by withdrawing electrons generally from the ring, but, at the same time, they activate the molecule toward nucleophilic attack, of such nature that the halogen atoms themselves are replaced by the attacking nucleophiles.

Electrophilic substitution

General physical properties. Generally in organic halogen compounds, the boiling points increase as the atomic weights of the halogen elements increase and as the molecular weights of the compounds themselves increase (see Table 1). On the other hand, the more highly branched the chain of a given number of carbon atoms is, the lower the boiling point of the compound is.

Table 1: Boiling Points of Alkyl Halides
(°C)

radical	formula	fluoride	chloride	bromide	iodide
Methyl	CH_3—	−78.4	−23.76	3.45	42.5
Ethyl	CH_3CH_2—	−38.0	+12.5	38.4	72.4
n-Propyl	$CH_3CH_2CH_2$—	− 2.5	46.6	71.0	102.5
Isopropyl	$(CH_3)_2CH$—	− 9.4	34.8	59.4	89.5
n-Butyl	$CH_3(CH_2)_3$—	32.5	78.5	101.5	130.4
Isobutyl	$(CH_3)_2CHCH_2$—	...	68.8	91.4	121.0
sec-Butyl	$C_2H_5CH(CH_3)$—	25.1	68.2	91.2	120.0
tert-Butyl	$(CH_3)_3C$—	12.1	50.7	73.2	103.3

Boiling points (all temperatures given in degrees Celsius unless noted otherwise) usually increase with the number of halogen atoms; *e.g.*, methyl chloride (CH_3Cl), −24°; methylene chloride (CH_2Cl_2), 40°; chloroform ($CHCl_3$), 62°; carbon tetrachloride (CCl_4), 77°.

Densities increase with increase in the atomic weight of the halogen and also with the number of halogen atoms in the molecule; *e.g.*, ethyl fluoride (C_2H_5F), 0.818; ethyl chloride (C_2H_5Cl), 0.923; ethyl bromide (C_2H_5Br), 1.47; ethyl iodide (C_2H_5I), 1.98; ethylidene chloride (CH_3CHCl_2), 1.18; ethylene dichloride (CH_2ClCH_2Cl), 1.25; 1,1,1-trichloroethane CH_3CCl_3, 1.35; 1,1,2-trichloroethane $CH_2ClCHCl_2$, 1.44 (the numbers in the names designating the carbon atoms to which the halogens are attached).

Melting and boiling points

Melting points and boiling points of organic halogen compounds, however, do not depend exclusively on the molecular weights of the molecules concerned; they also are influenced by the intermolecular forces involved and hence on the polar (unsymmetrically charged) nature of the molecule (polar molecules being more attracted to one another than nonpolar molecules and hence more resistant to both melting and boiling). For instance, geometrical isomers—that is, compounds with the same structural framework but different spatial arrangements of the substituent groups of atoms—have different melting and boiling points (which also differ from those of any other isomers of the same compounds) because the halogen atoms are differently oriented with respect to one another. An example of such a circumstance is found in the difluoroethylenes: 1,1-difluoroethylene, $CH_2{=}CF_2$, boils at −84° C; whereas the two geometrical isomers of 1,2-difluoroethylene, CHF=CHF, which

Table 2: Melting Points and Boiling Points of Aryl Halides (°C)

compound	fluoro compound		chloro compound		bromo compound		iodo compound	
	melting point	boiling point	melting point	boiling point	melting point	boiling point	melting point	boiling point
Monohalobenzene	−41.9	85	−45.2	132.0	−30.6	156.2	31	188.5
o-Dihalobenzene	−34	91	−17.2	179.2	+ 6.7	225	26.7	286
m-Dihalobenzene	−59.3	83	−26.3	172	− 7	220	35	285
p-Dihalobenzene	−13	89	53.0	174.5	+87.3	220	129	285
1,2,3-Trihalobenzene	−13.5	95	52.4	218	88		116	
1,2,4-Trihalobenzene	−35	88	16.6	213	44		91.5	
1,3,5-Trihalobenzene	− 5.5	75.5	63	203	119	271	184	
1,2,3,4-Tetrahalobenzene	−42	95	47.5	254	47.5		136	
1,2,3,5-Tetrahalobenzene	−48	83	51	246	98		148	
1,2,4,5-Tetrahalobenzene	4	90	138	244	173		254	
Pentahalobenzene	−48	85	87	276	159		172	
Hexahalobenzene	5	80	229	326	326		340–350	

are known as *cis*- and *trans*-1,2-difluoroethylene, boil at −26° C (−15° F) and −53° C (−43° F) respectively. There is a considerable variation in physical properties among polyhalogenated compounds also because of differences in polarity. For example, a nonpolar fully fluorinated compound often has a lower boiling point than less highly fluorinated compounds with the same carbon skeleton, chiefly because the latter are more polar (see Table 2). Boiling points of representative alkyl halides are given in Table 1, and melting and boiling points of aryl halides are presented in Table 2.

The weak intermolecular forces in fully fluorinated compounds (resulting from their low polarity) also are responsible for the low viscosities these substances exhibit—they are used as oils and greases under carefully controlled conditions. (Their viscosities decreased rapidly with increase in temperature, however, and this fact prevents their more general use.)

The bond dissociation energies of carbon-to-halogen bonds, given in the standard units of kilocalories per mole (a mole being equal to the molecular weight expressed in grams), are carbon–fluorine, 108; carbon–chlorine, 81; carbon–bromine, 67; and carbon–iodine, 53. For purposes of comparison, it may be noted that the carbon–carbon and carbon–hydrogen bond dissociation energies are 84 and 102 kilocalories per mole, respectively. The high dissociation energy of the carbon–fluorine bond and the low dissociation energy of the carbon–iodine bond help to explain why the fluorocarbons are thermally stable whereas organic iodine compounds dissociate readily on heating (with liberation of iodine).

ALKYL HALIDES

Monohalides. *Preparation.* Hydrocarbons bearing one halogen substituent are most commonly prepared by two methods: (1) addition of halogen acids to olefins (alkenes) and (2) replacement of hydroxyl (—OH) groups in alcohols. In its simplest form addition of halogen acids to olefins is an electrophilic addition reaction in which the positively charged end of the polarized halogen acid is attracted to the olefinic double bond. The acid then dissociates producing a halide ion and a proton (hydrogen ion) which, with the pi electrons of the double bond, forms a positively charged intermediate known as a pi complex. Finally, the halide ion reacts with the pi complex to give the ultimate product. A typical reaction of this type is shown in chemical formulas below:

$$CH_3CH{=}CH_2 + \overset{\delta^+\ \delta^-}{H{-}X} \longrightarrow H_3C{-}\overset{H}{\overset{|}{C}}{-}CH_2 \quad (H \text{ bridging to } C^+)$$

propylene — halogen acid — pi complex

$$+ X^- \longrightarrow CH_3CHXCH_3 .$$

halide ion — isopropyl halide

The direction of addition to the double bond can be predicted by a consideration of the effects of the alkyl groups (in this case, only methyl, CH_3-) attached to the double bond carbons. The various halogen acids undergo addition to double bonds, in decreasing order of reactivity, as follows: hydrogen iodide, hydrogen bromide, hydrogen chloride, and hydrogen fluoride.

In the presence of traces of peroxides (which can be produced from olefins and traces of oxygen by sunlight), hydrogen bromide reacts with olefins in a different fashion, often to give products in which the addition occurs in the opposite direction to that in the above reaction (that is, with bromine atom appearing on the other double bond carbon atom).

Alkyl halides are produced from alcohols by means of nucleophilic substitution reactions with hydrogen halides (halogen acids), phosphorus halides, or thionyl halides. The reactivities of the halogen acids in this reaction are, in decreasing order: hydrogen iodide, hydrogen bromide, and hydrogen chloride. (Hydrogen fluoride cannot be used.) Tertiary alcohols (alcohols with three carbon atoms attached to the hydroxyl carbon) undergo this reaction readily. Secondary and primary alcohols (alcohols with two and one carbon atom joined to the hydroxyl carbon, respectively) react with less facility. For example, tertiary butyl alcohol reacts with aqueous hydrogen iodide at room temperature, according to the following equation:

Conversion of alcohols to halides

$$H_3C{-}\overset{CH_3}{\overset{|}{C}}{-}OH + HI \longrightarrow H_3C{-}\overset{CH_3}{\overset{|}{C}}{-}I + H_2O .$$
$$\quad\ CH_3 \qquad\qquad\qquad\qquad\quad CH_3$$

tert-butyl alcohol — hydrogen iodide — *tert*-butyl iodide — water

Secondary bromides and iodides (corresponding in structure to secondary alcohols) usually can be prepared in the same way. For primary alkyl chlorides, however, it is necessary to use a catalyst and warm the solution. (Long heating may cause isomerization—rearrangement to another product.)

Phosphorus pentachloride, phosphorus tribromide, red phosphorus and bromine, phosphorus triiodide, and red phosphorus and iodine all react readily with the appropriate alcohols to give primary, secondary, and tertiary halides. Thionyl chloride and thionyl bromide react with primary or secondary alcohols in the presence of an organic base, such as pyridine. Tertiary alcohols need no base.

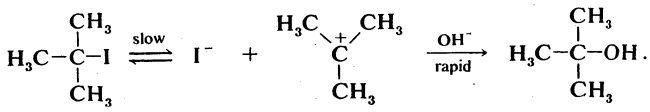
The various alkyl halides can be interconverted, and they also can be prepared from other alkyl compounds. Either process is one of nucleophilic substitution by halide ions, and the choice of solvent for the reaction is important.

Reactions. In most of their reactions, the alkyl iodides are more reactive than bromides, which are more reactive than the chlorides, and these, in turn, are more reactive than the fluorides. This order of reactivity reflects the relative ease with which the respective carbon–halogen bonds are broken. Also, generally speaking, the tertiary halides react most readily, the secondary less so, and the primary least of all, reflecting the fact that alkyl substituents tend to donate electrons to the carbon atom that carries the halogen, causing it to release the halide ion,

and the more alkyl substituents there are (three in tertiary compounds), the more easily is the halide ion lost. The chemical reactions undergone by alkyl halides can be classified under three headings: (1) nucleophilic substitutions, (2) elimination reactions, and (3) free-radical processes.

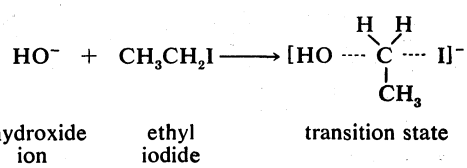
Nucleophilic substitutions

Nucleophilic substitution reactions are classified as unimolecular or bimolecular according to whether the reaction involves one or two molecules. Unimolecular nucleophilic substitutions proceed by way of ionization of the alkyl halide—that is, by cleavage of the molecule to give a positively charged alkyl, or carbonium, ion and a negatively charged halide ion. This type of reaction is undergone most readily by a tertiary alkyl halide in ionizing solvents (aqueous alcohol, for example). Reactions of this type are known to chemists as S_N1 reactions (meaning substitution, nucleophilic, unimolecular). A good example of an S_N1 reaction is the hydrolysis of *tert*-butyl iodide, as shown below:

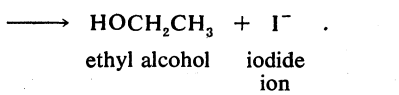


tert-butyl iodide | iodide ion | carbonium ion | *tert*-butyl alcohol

Bimolecular nucleophilic substitutions—S_N2 reactions—proceed through a transition state that includes a nucleophile—for example a hydroxide ion—as well as a molecule of the alkyl halide. They are most common for primary alkyl halides in nonionizing solvents. An example is the reaction of ethyl iodide with a hydroxide ion to give ethyl alcohol and the iodide ion:



hydroxide ion | ethyl iodide | transition state

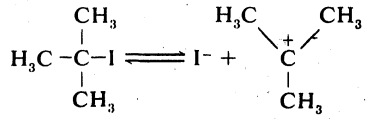


ethyl alcohol | iodide ion

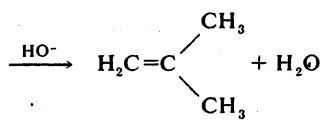
Nucleophiles other than the hydroxide ion that react in this fashion are ammonia and trimethylamine, and the ethoxide, acetate, cyanide, azide, acid sulfide, and hydride ions.

Elimination reactions

Whenever alkyl halides are hydrolyzed, a certain amount of olefin is formed by an elimination process. Elimination sometimes becomes the main reaction, as in the hydrolysis of tertiary halides, in which the carbonium ion formed can easily eliminate a proton. An example is the hydrolysis of *tert*-butyl iodide to give isobutene:



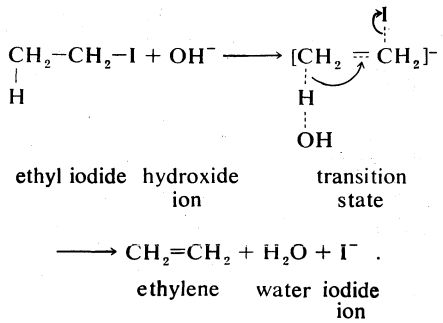
tert-butyl iodide | iodide ion | carbonium ion



isobutene | water

This is a unimolecular elimination—E_1 reaction—so classified because the rate at which it occurs is governed by the rate of formation of the carbonium ion from the alkyl halide.

Biomolecular eliminations—E_2 reactions—occur mainly with primary halides. They occur by removal of a hydrogen ion from the carbon atom adjacent to that carrying the halogen atom. For example, the elimination reaction involving ethyl iodide proceeds as shown (with the movements of the electrons comprising the various bonds indicated by curved arrows):



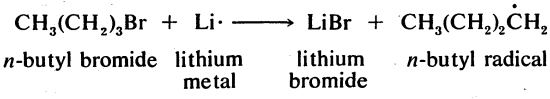
ethyl iodide | hydroxide ion | transition state

ethylene | water | iodide ion

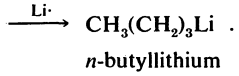
By suitable choice of the reaction conditions (that is, by selecting the right solvent, temperature, and concentration of nucleophile) it is often possible to influence whether a particular reaction proceeds by bimolecular or unimolecular substitution or by an elimination reaction. Because the products of these reactions often are different, the distinction may be vitally important in a synthetic process.

Free-radical reactions

Alkyl halides usually react with metals by a process that generates free radicals—that is, substances with single (unpaired) electrons. The free radical may then react with a second metal atom, resulting, as shown in the example below, in formation of an organometallic compound (that is, an organic compound with a metal substituent):

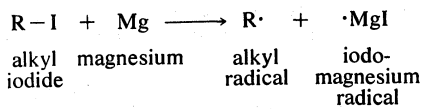


n-butyl bromide | lithium metal | lithium bromide | *n*-butyl radical

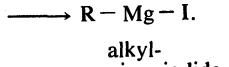


n-butyllithium

(In the names of these compounds *n*, for normal, indicates that the hydrocarbon chain is straight, or unbranched.) With magnesium, the reaction is carried out in an ether as solvent to give an organomagnesium halide:

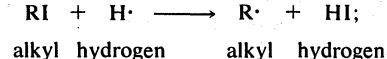


alkyl iodide | magnesium | alkyl radical | iodomagnesium radical

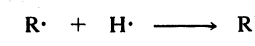


alkylmagnesium iodide

In these reactions iodides and bromides are generally used because chlorides and fluorides react much less readily. The reduction of alkyl halides by sodium and alcohol, zinc and dilute acid, or similar reducing agents may also be a free-radical process:



alkyl iodide | hydrogen atom | alkyl radical | hydrogen iodide



alkyl radical | hydrogen atom | alkane

Polyhalides. *Preparation.* Alkylidene halides—that is, halides that have two halogen atoms attached to the same carbon atom—can be made from aldehydes and ketones. The chlorides, bromides, and iodides are prepared using phosphorus halides. For example, isopropylidene chloride is produced from acetone and phosphorus pentachloride, which may be illustrated according to the following equation:

$$PCl_5 + CH_3COCH_3$$

phosphorus pentachloride; acetone

$$\longrightarrow POCl_3 + CH_3CCl_2CH_3.$$

phosphorus oxychloride; isopropylidene chloride

To prepare the fluorides, sulfur tetrafluoride or phenyl sulfur trifluoride is used with an aldehyde or ketone. Another general method for preparing alkylidene halides is to add a halogen acid to a vinyl halide or an alkyne. An example is the addition of hydrogen chloride to acetylene, vinyl chloride being produced first:

$$HC \equiv CH \xrightarrow{HCl} CH_2{=}CHCl \xrightarrow{HCl} CH_3CHCl_2.$$

acetylene; vinyl chloride; ethylidene chloride

Alkylene dihalides, in which two halogen atoms are attached to different carbon atoms, are formed from halogens and alkenes by electrophilic addition, or from glycols (compounds with two alcohol groups) by nucleophilic substitution exactly as the alkyl halides are prepared from ordinary alcohols.

The best known trihaloalkanes are those in which all three halogen atoms are on the same carbon atom; these are the haloforms, of which chloroform, $CHCl_3$, is the most familiar. Haloforms generally are made by the halogenation and subsequent alkaline hydrolysis of compounds containing the acetyl group (CH_3CO-) or structures capable of producing it on oxidation. Fluorination of acetophenone, for instance, followed by hydrolysis of the product, gives fluoroform, as shown:

$$C_6H_5COCH_3 \longrightarrow C_6H_5COCF_3$$

acetophenone; α,α,α-trifluoroacetophenone

$$\xrightarrow{H_2O} C_6H_5COOH + HCF_3.$$

benzoic acid; fluoroform

Larger molecules can be made by the free-radical addition of chloroform to olefins, as follows:

$$CH_2{=}CH_2 + \cdot CCl_3 \longrightarrow \cdot CH_2CH_2CCl_3$$

ethylene; trichloromethyl radical (from $CHCl_3$); intermediate radical

$$\cdot CH_2CH_2CCl_3 + HCCl_3$$

intermediate radical; chloroform

$$\longrightarrow CH_3CH_2CCl_3 + \cdot CCl_3.$$

1,1,1-trichloropropane; trichloromethyl radical, regenerated

If the proportion of chloroform present in the reaction mixture is reduced, polymerization (formation of long, chainlike molecules with repeating units) occurs because the intermediate radical reacts with more olefin, giving a series of products of varying chain length, all of which have trichloromethyl end groups. Such products are formed as in the following equations:

$$\cdot CH_2CH_2CCl_3 + nCH_2{=}CH_2 \longrightarrow \cdot CH_2CH_2(CH_2CH_2)_nCCl_3$$

intermediate radical; ethylene; polymer radical

$$\cdot CH_2CH_2(CH_2CH_2)_nCCl_3 + HCCl_3$$

polymer radical; chloroform

$$\longrightarrow CH_3CH_2(CH_2CH_2)_nCCl_3 + \cdot CCl_3,$$

polymer product; trichloromethyl radical

in which n is a variable number that depends on the number of times the first step is repeated.

Compounds more highly halogenated than the above classes of polyhalides result from the direct halogenation of alkanes by elemental fluorine, chlorine, or bromine. This process, too, is considered to involve free radicals. Because the energy liberated in replacing hydrogen atoms by fluorine atoms is more than enough to break a carbon-to-carbon bond, the vapour-phase fluorination of alkanes occurs violently unless the reaction is moderated by dilution of the reactants with gaseous nitrogen and unless the heat produced by the reaction is conducted away by a metal surface. As an alternative, metal fluorides can be used instead of elemental fluorine, or an electrochemical process can be employed. In most instances, but not in the electrochemical method, the amount of partly fluorinated material can be adjusted by varying the reaction conditions, but mixtures are always obtained. Chlorination of alkanes with elemental chlorine in the liquid phase gives mainly polychloro compounds but, in the gas phase, it is possible to arrange conditions so as to give more lightly chlorinated material. Bromination is carried out like chlorination; direct iodination, however, is not feasible. Fluorination is such a vigorous process that there is no discrimination in the attack on the different types of hydrogen atoms; in chlorination and, to a greater extent in bromination, tertiary hydrogen atoms are substituted more readily, followed by secondary and primary hydrogens, in order.

Direct halogenation

Reactions. The reactions of alkylene dihalides are very similar to the reactions of alkyl halides, but, when there are two or more halogen atoms on the same carbon atom, there is a sharp decrease in the reactivity of the halogen atoms. In general, as the atomic weight of the halogen atoms increases the stability of the polyhaloalkanes decreases, not only because the carbon–halogen bonds are weaker, but also because of the crowding caused by the packing together of the larger atoms. Stabilities decrease enormously, therefore, in the series from carbon tetrafluoride, through the chloride and bromide to carbon tetraiodide. Furthermore, 1,1,1-trifluoroethane is resistant to hydrolysis by acids, whereas the trichloro compound is easily hydrolyzed to acetic acid. Because the fluorine atom is not much bigger than the hydrogen atom, it is possible to prepare series of fluorocarbon derivatives entirely analogous to the related hydrocarbons. The sizes of the other halogen atoms, however, prevent the formation of many fully halogenated compounds. Fully fluorinated alkanes (called perfluoroalkanes) are particularly stable—both to heat and to chemical reagents—because of the very strong carbon—fluorine bonds and because the fluorine atoms are of just such size that they give good protection to the carbon chains.

Influences on reactivity

Polyhalogenated compounds mainly undergo elimination reactions. Halogen atoms are eliminated either as molecules of the halogen acid (HX) or as molecules of the free halogen itself (X_2). In both cases iodides are most reactive, and fluorides least. Elimination reactions are widely used in the preparation of halogenated alkenes and alkynes (see below *Alkenyl halides* and *Alkynyl halides*). Exchange of one kind of halogen atom for another occurs when polyhalo-compounds are heated with aluminum halides.

ALKENYL HALIDES

Preparation. Alkenyl halides, halogen derivatives of olefins, differ from one another in their chemical behaviour chiefly because of differences in the relationship of the halogen atom to the carbon–carbon double bond. Vinyl halides are those alkenyl halides in which a halogen atom is attached directly to one of the double-bond carbons; allyl halides are those in which the halogen is attached to one of the carbons adjacent to the double-bond carbons. Only in these classes of alkenyl halides is there interaction between the electrons of the halogen atoms and those of the double bond. Therefore, only these alkenyl halides show unique chemical properties; others of the general class resemble ordinary alkyl halides in their properties and methods of preparation.

Varieties of alkenyl halides

One method that is used to prepare vinyl halides of many different types is the addition of an appropriate simple molecule to an alkyne (triply bonded hydrocarbon). These reactions may occur spontaneously, when a simple molecule like a halogen acid is added, or they may need heating, when more complex molecules are to be added. Examples of both are shown below:

$$CH{\equiv}CH + HCl \longrightarrow CH_2{=}CHCl;$$

acetylene — hydrogen chloride — vinyl chloride

$$CF_3C{\equiv}CH + CCl_3I$$

trifluoromethyl-acetylene — trichloro-iodomethane

$$\xrightarrow{\text{heat}} CF_3Cl{=}CHCCl_3$$

1,1,1-trifluoro-2-iodo-4,4,4-trichloro-2-butylene

The other common methods of preparing vinyl halides are dehydrohalogenation (removal of hydrogen halides) and dehalogenation (removal of halogens) of polyhaloalkanes. Dehydrohalogenation can, in simple cases, be performed by heating with bases, for example:

$$CH_2ClCH_2Cl + NaOH$$

ethylene dichloride — sodium hydroxide

$$\longrightarrow CH_2{=}CHCl + NaCl + H_2O.$$

vinyl chloride — sodium chloride — water

The most common method of dehalogenation is to heat the haloalkane in ethanol solution with zinc dust, as in the following example:

$$CH_2BrCHBrF \xrightarrow[C_2H_5OH]{Zn} CH_2{=}CHF.$$

1-fluoro-1,2-dibromoethane — vinyl fluoride

Allyl halides

Allyl halides are prepared by the action of the appropriate halogen acid on allyl alcohols. Allyl bromide also is made by treating propene with *N*-bromosuccinimide. The bromide can be converted into other halides by halogen exchange.

$$CH_2{=}CHCH_3 + BrN\begin{<}COCH_2\\COCH_2\end{>} \longrightarrow CH_2{=}CHCH_2Br + HN\begin{<}COCH_2\\COCH_2\end{>}$$

propene — *N*-bromo-succinimide — allyl bromide — succinimide

Reactions. In the vinyl halides, because of the interaction of the electrons of the halogen atom with those of the double bond, the carbon to halogen bond is stronger than it is in ordinary alkyl halides. Moreover, the carbon atom to which the halogen is attached does not carry a positive charge, as does the corresponding carbon atom in typical alkyl halides, and as a result nucleophilic replacement of the halogen atom in vinyl halides does not occur. This lack of positive charge, however, does not prevent the formation of organomagnesium halides, nor does it interfere with normal electrophilic and free radical additions to the double bond.

As the number of halogen atoms in a vinyl halide molecule increases, the tendency to undergo electrophilic additions decreases and that to undergo nucleophilic additions increases. Both effects are due to the decrease in electron density of the double bond caused by the electronegative substituents. For polyhalo vinyl compounds, then, the most commonly observed reaction is the addition of nucleophiles, such as alcohols, amines, and thiols. For example, the addition of ethanol to 1,1,2-trichloroethylene gives an ether as an initial reaction product; this is subsequently dehydrohalogenated to dichlorovinyl ethyl ether.

Polyhalo vinyl compounds

$$CCl_2{=}CHCl + C_2H_5OH \longrightarrow C_2H_5OCCl_2CH_2Cl$$

1,1,2-trichloro-ethylene — ethanol — initial product

$$\xrightarrow{NaOC_2H_5} C_2H_5OCCl{=}CHCl$$

dichlorovinyl ethyl ether

Allyl halides, which conform to the structural pattern $CH_2{=}CHCH_2X$, are very much more susceptible to nucleophilic substitution of the halogen atom than are alkyl halides. When a carbonium ion intermediate is formed, (S_N1 reaction) it may be represented by the two formulas:

$$\underset{\gamma}{CH_2}{=}\underset{\alpha}{CH}{-}\overset{+}{\underset{\alpha}{CH_2}} \longleftrightarrow \overset{+}{\underset{\gamma}{CH_2}}{-}\underset{\alpha}{CH}{=}\underset{\gamma}{CH_2}.$$

Attack by the nucleophile occurs at either the alpha (α) or gamma (γ) carbon atoms (as shown). This choice becomes significant for unsymmetrical allyl carbonium ions as the product is different in each case. Bimolecular (S_N2) substitution usually involves direct replacement of the allyl halogen atom, but this reaction also can occur with attack at the gamma position to give an alternate product.

Halogenated alkenes of only two carbon atoms usually polymerize readily, but halogenated alkenes bearing large, halogenated alkyl groups attached to the unsaturated carbon atoms may not do so. Fluorinated olefins easily form cyclic dimers when heated above 200° C (392° F). Halogenated alkenes often can be dehydrohalogenated or dehalogenated to alkynes.

ALKYNYL HALIDES

Preparation. In many cases, compounds containing halogen atoms directly joined to triply bonded carbon atoms are formed by reaction between molecular halogen and a metallic derivative of the alkyne. Various sodium acetylides, for example, react with free iodine to give the corresponding iodo compounds.

$$RC{\equiv}CNa + I_2 \longrightarrow RC{\equiv}CI + NaI$$

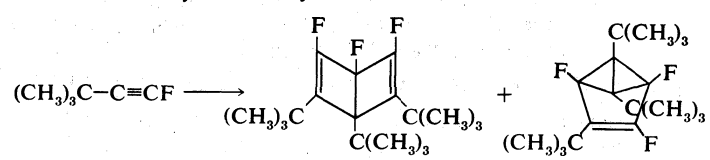
In other instances, alkynyl halides are more easily prepared by dehalogenation or dehydrohalogenation reactions of halogenated alkanes or alkenes. Examples of both reactions, using 1,1,2-trichloroethylene as starting material, are shown below:

$$CCl{\equiv}CH \xleftarrow[\text{alcohol}]{Zn} CCl_2{=}CHCl \xrightarrow[NaOH]{\text{aq alc}} CCl{\equiv}CCl.$$

chloro-acetylene — 1,1,2-trichloro-ethylene — dichloro-acetylene

Reactions. Halogenated acetylenes (alkynes), in which the halogen atom is directly bonded to one of the unsaturated carbon atoms, generally are unstable; the lower the molecular weight, the more likely they are to explode. This instability may be due to spontaneous exothermic (heat-producing) polymerization, which, in some cases, can be regulated so that trimerization to halogenated benzenes occurs. Sometimes isomers of benzene derivatives are formed, as for example, in trimerization of *tert*-butylfluoroacetylene:

Instability of halo-alkynes



tert-butylfluoro-acetylene — isomeric benzene derivatives

In these structures the polygons represent rings of carbon atoms (the carbon atoms at the angles not being shown), with the lines indicating the bonds between atoms.

A halogen atom attached to an unsaturated carbon atom in a halogenated acetylene is resistant to nucleophilic

substitution. Most nucleophiles, therefore, react with haloalkynes first by addition and then by elimination of hydrogen halide. Ammonia, for instance, adds to dichloroacetylene; the initial product then loses hydrogen chloride to give an alkyne, and this in turn isomerizes to a nitrile.

$$ClC{\equiv}CCl + NH_3 \longrightarrow ClCH{=}C(NH_2)Cl$$

dichloroacetylene — ammonia — 1,2-dichloroaminoethylene

$$\longrightarrow HCl + ClC{\equiv}CNH_2 \xrightarrow{\text{isomerization}} CH_2ClC{\equiv}N$$

hydrogen chloride — 1-amino 2-chloroacetylene — chloroacetonitrile

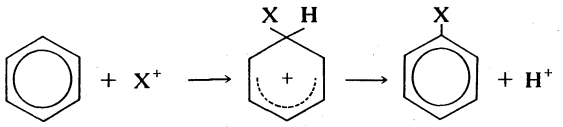
In most solvents, such as carbon tetrachloride, electrophilic additions of halogens and halogen acids to alkynyl halides take place, but the reaction is vigorous and must be controlled carefully. The hydrogen in monohaloacetylenes is acidic, as in acetylene itself, and can be replaced by metals. Haloacetylenes in which the halogen is remote from the triple bond are generally stable and show the separate reactivities of alkynes and alkyl halides. The perfluoroalkylalkynes are less unstable than most other alkynes, and they retain the reactivity of the triple bond, readily undergoing free-radical, electrophilic, and nucleophilic addition reactions.

ARYL HALIDES

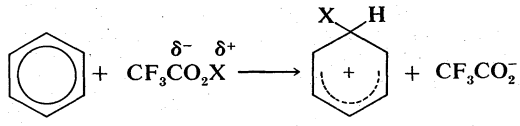
Nuclear-halogenated aromatic compounds. *Preparation.* There are three main methods for producing aromatic compounds that carry halogen atoms on the aromatic ring (or aromatic nucleus, as it is often called). These are (1) replacement of hydrogen by halogen, (2) replacement of an amino group by halogen, and (3) dehydrohalogenation or dehalogenation of polyhalocyclohexanes or polyhalocyclohexenes.

Replacement of hydrogen

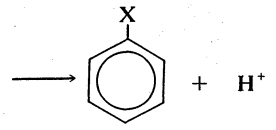
The mechanism of hydrogen replacement is one of electrophilic substitution. It involves, first, the production of either a positively charged halogen ion or a polarized molecule in which the halogen has a partial positive charge (the electrophile). This positive centre then reacts with benzene to form an intermediate from which a proton is eliminated.



benzene — intermediate — halobenzene — proton



benzene — polarized halogen compound — intermediate



halobenzene — proton

Ferric chloride or aluminum chloride are used as catalysts with chlorine or bromine to produce the positively charged halogen ion.

$$Cl_2 + AlCl_3 \rightleftharpoons Cl^+ + AlCl_4^-$$

Iodination requires the presence of an oxidizing agent, such as nitric acid, to remove hydrogen iodide from the reaction as soon as it is produced, as well as to encourage the formation of positively charged iodine. Hypohalous acids, *N*-chloro- or *N*-bromoamides, and acyl hypohalites are useful halogenating agents because they are polarized in such a way that the halogen atom carries a positive charge.

A chlorine, bromine, or iodine atom in the benzene nucleus deactivates it and causes further substitution to occur in the *ortho* and *para* positions (see above *General chemical properties*). Fluorine also causes *ortho* and *para* orientation, but does not deactivate the nucleus. Excellent yields of nuclear halogenated compounds are formed by chlorinating, brominating, or iodinating substituted benzenes.

Replacement of amino groups

Any of the four halogens can be introduced into the nucleus by diazotizing an aromatic amine and then allowing the resulting diazonium salt to decompose in the presence of a suitable salt of the halogen. Chlorine and bromine, for example, are introduced by heating a diazonium salt in dilute acid solution with cuprous chloride or bromide:

$$ArNH_2 \longrightarrow Ar\overset{+}{N}{\equiv}N\ X^- \xrightarrow[HX]{Cu_2X_2} ArX + N_2.$$

aromatic amine — diazonium salt — aryl halide — nitrogen

Ar = any aromatic nucleus

Alternatively, the appropriate diazonium salt can be heated in dilute acid with copper powder:

$$Ar\overset{+}{N}{\equiv}N\ X^- \xrightarrow[HX]{Cu} ArX + N_2.$$

diazonium salt — aryl halide — nitrogen

Iodine is introduced simply by warming a diazonium salt in the presence of aqueous potassium iodide, for example:

$$C_6H_5\overset{+}{N}{\equiv}N\ Cl^- + KI \longrightarrow C_6H_5I + N_2 + KCl.$$

benzenediazonium chloride — potassium iodide — iodobenzene — nitrogen — potassium chloride

The most common method used to introduce fluorine is the dry distillation of a diazonium fluoroborate:

$$Ar\overset{+}{N}{\equiv}NBF_4^- \longrightarrow ArF + N_2 + BF_3.$$

diazonium fluoroborate — aryl fluoride — nitrogen — boron trifluoride

The photochemical (*i.e.*, free-radical) chlorination of benzene causes the addition of chlorine to the double bonds and gives a mixture of hexachlorocyclohexanes. Dehydrohalogenation of this mixture by heating with aqueous alcoholic alkali gives a mixture of trichlorobenzenes.

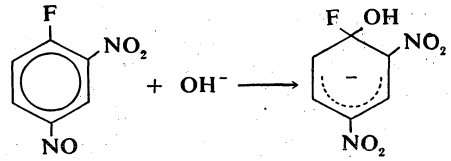
$$C_6H_6 \xrightarrow[\text{ultraviolet light}]{Cl_2} C_6H_6Cl_6 \xrightarrow{\text{alkali}} C_6H_3Cl_3 + 3HCl$$

benzene — hexachlorocyclohexanes — trichlorobenzenes — hydrogen chloride

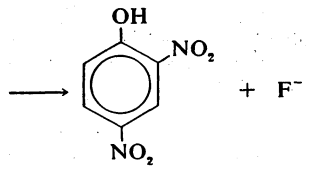
From the mixture obtained by fluorinating benzene with cobalt trifluoride, octa- and nonafluorocyclohexanes can be isolated. Dehydrofluorination of these substances by heating with concentrated aqueous potash gives penta- and hexafluorobenzenes. Deca-, nona-, and octafluorocyclohexanes can be dehydrofluorinated to polyfluorocyclohexenes and -dienes, and these compounds in turn can be defluorinated giving hexa-, penta-, and tetrafluorobenzenes.

Reactions. The main difference between aryl halides and alkyl halides is that the former are much more resistant to nucleophilic attack. If, however, the halogen in an aromatic compound is activated by the presence of other groups in the molecule, or if especially vigorous conditions (such as elevated temperatures and the presence of strong bases) are used, then nucleophilic substitutions or elimination–addition reactions (see below), which give the same end result, can take place. Thus, 2,4-dinitrofluorobenzene on treatment with a nucleophile

such as hydroxide ion undergoes a bimolecular substitution to give a phenol. When the nucleus is not activated either stronger bases or much more vigorous conditions

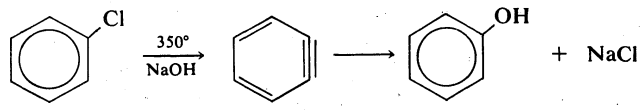


2,4-dinitrofluorobenzene | hydroxide ion | intermediate



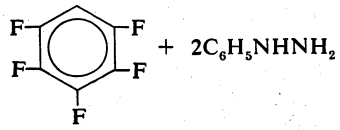
2,4-dinitrophenol | fluoride ion

are needed for nucleophilic attack; in such cases elimination of hydrogen halide gives a benzyne structure, to which the nucleophile then adds.

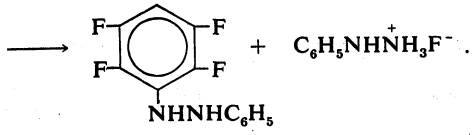


chlorobenzene | benzyne | phenol | sodium chloride

In hexachlorobenzene, the halogen atoms activate the nucleus in the same way as do the nitro groups in dinitrofluorobenzene; and, on heating with potassium fluoride, hexafluoro- and mixed chlorofluorobenzenes are formed by nucleophilic reactions. Pentafluoro- and chlorofluoropyridines are similarly made. Polychloro- and polyfluorobenzenes are highly susceptible to this form of attack. Reagents such as ammonia, sodium methoxide, lithium aluminum hydride, and phenylhydrazine often are used as nucleophiles. An example is given below:



pentafluorobenzene | phenylhydrazine

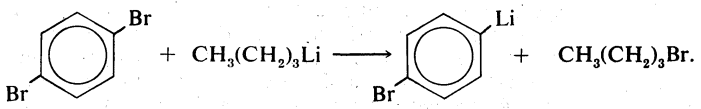


2,3,5,6-tetrafluorohydrazobenzene | phenylhydrazine hydrofluoride

The nuclear hydrogen atoms in aryl halides generally are subject to electrophilic attack and these compounds, therefore, undergo such typical aromatic reactions as halogenation, nitration, and sulfonation. Substitution occurs chiefly in the *para* position, but also in the *ortho* position.

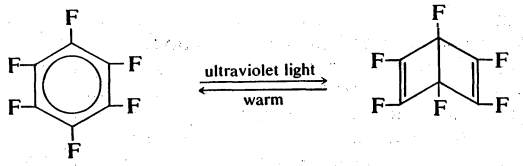
Formation of organometallic compounds

Reactions of aryl halides with metals are similar to those of alkyl halides. Aryl fluorides do not form organomagnesium halides, and chlorides react sluggishly; bromides and iodides, however, react readily. In polyhalo compounds, any hydrogen atoms left in the molecule are acidic because of the presence of the electronegative halogen atoms, and as a result these compounds regularly form organometallic compounds by hydrogen–metal exchange. Organolithium compounds often are formed by halogen–metal exchange in the less halogenated compounds. For example:



p-dibromobenzene | butyllithium | *p*-bromophenyllithium | butyl bromide

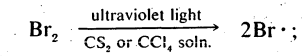
In nuclear halogenated compounds, chlorine, bromine, and iodine atoms are replaced by hydrogen atoms when the compounds are heated with hydrogen gas in the presence of a catalyst such as finely divided nickel, platinum, or palladium. Hexafluorobenzene, on ultraviolet irradiation, is converted into the so-called Dewar isomer (an abnormal form of benzene), which reverts to the normal form on gentle heating, but which may explode.



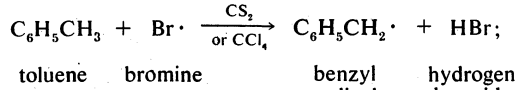
hexafluorobenzene | Dewar isomer

Side-chain-halogenated aromatic compounds. *Preparation.* Direct halogenation of an alkyl group (side chain) attached to an aromatic ring is a free-radical process and, as such, it is helped by exposure to visible or ultraviolet light and by boiling the reaction mixture, which aids dissociation of the halogen molecules to free atoms. After substitution is complete, the reagents effecting halogenation also cause saturation of the nuclear double bonds. In side-chain halogenation it is essential to avoid any trace of catalysts that promote nuclear substitution. Fluorine reacts so vigorously that it cannot be introduced preferentially into the side chain in this way, and iodine requires the presence of an oxidizing agent. The method is ideal for chlorine and bromine.

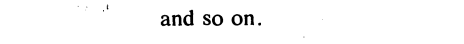
Side-chain halogenation



molecular bromine | atomic bromine



toluene | bromine atom | benzyl radical | hydrogen bromide

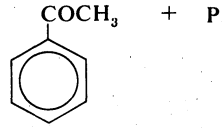


and so on.

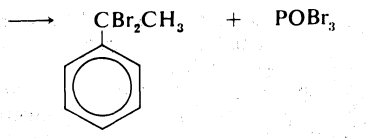
Sulfuryl chloride, with a trace of benzoyl peroxide as a free-radical initiator, can be used as a source of chlorine atoms for side-chain chlorination. Such chlorinations proceed rapidly in the dark. Other reagents that produce halogen atoms and therefore start chain reactions are tertiary butyl hypochlorite and hypoiodite and *N*-bromosuccinimide. The degree of halogenation can be controlled by regulating the amount of reagent used, but mixtures are always produced.

Other methods for preparing compounds halogenated in the side chain are similar to those used to prepare alkyl halides, either from the corresponding alcohol, such as benzyl alcohol ($C_6H_5CH_2OH$), or by the electrophilic addition of hydrogen halides or halogens to aryl olefins.

Aromatic aldehydes and ketones react with phosphorus halides or with sulfur tetrafluoride to give dihalo compounds.



acetophenone | phosphorus pentabromide

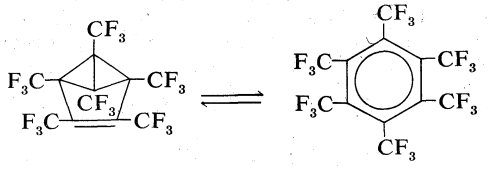


α,α-dibromoethylbenzene — phosphorus oxybromide

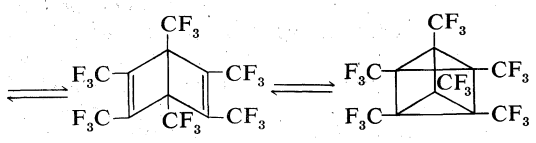
Aromatic acids react with sulfur tetrafluoride to give trifluoromethyl derivatives.

Reactions. Typical examples of aromatic compounds with halogen atoms in their side chains are benzyl chloride ($C_6H_5CH_2Cl$), benzylidene chloride ($C_6H_5CHCl_2$), and benzotrichloride ($C_6H_5CCl_3$). The reactivities of these compounds are intermediate between those of the alkyl and allyl halides with similar halogen atoms. Thus, benzyl halides react readily with nucleophiles; hydrolysis of benzylidene halides gives benzaldehyde, and that of benzotrichloride gives benzoic acid. When halogen atoms are not attached to the carbon atom adjacent to the benzene nucleus, the compounds react almost like similar alkyl halides.

Hexakis(trifluoromethyl)benzene undergoes ready photochemical valence bond isomerization to the benzvalene, "Dewar," and prismane isomers, which are surprisingly stable for compounds of this type.



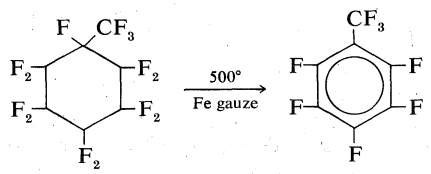
benzvalene isomer — hexakis(trifluoromethyl) benzene



Dewar isomer — prismane isomer

Nuclear- and side-chain-halogenated aromatic compounds. *Preparations.* Compounds that are lightly halogenated in both nucleus and side chain can obviously be prepared by methods already described. Of those that are highly or completely halogenated, only the polychloro and polyfluoro compounds are known, presumably because bromine and iodine atoms are too large. The photochemical chlorination of the side chains of benzene homologues fully chlorinated in the nucleus proceeds, for example, to the dichloromethyl or the trichloroethyl stage, where it stops. It also is difficult to complete the nuclear chlorination of benzene homologues with fully chlorinated side chains. There is one catalyst, aluminum trichloride and sulfur monochloride with sulfuryl chloride, sufficiently active to give perchloroalkaryl compounds (the prefix per- indicating complete substitution) from starting materials with fully halogenated side chains and nonhalogenated nuclei. Of the simple benzene homologues, only *m*-xylene and mesitylene resist complete chlorination by this method. Perfluoro homologues of benzene are prepared by first converting the hydrocarbons (using cobalt trifluoride) to perfluorocyclohexanes. The latter, then, are defluorinated to the fully fluorinated homologue of benzene.

Fully halogenated compounds



perfluoromethyl cyclohexane — octafluorotoluene

Nuclear fluorinated aromatic compounds can be fluoroalkylated by nucleophilic attack with anions generated from fluoride ion and perfluoroalkenes, as shown below:

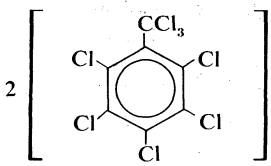
$$CF_3CF{=}CF_2 + F^- \longrightarrow CF_3\bar{C}FCF_3;$$

perfluoropropene — fluoride ion — intermediate anion

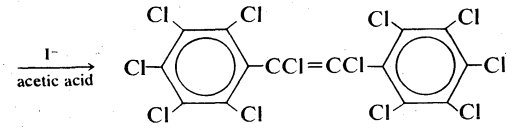
$$C_6F_6 + CF_3\bar{C}FCF_3 \longrightarrow C_6F_5CF(CF_3)_2 + F^-.$$

hexafluorobenzene — anion — perfluorocumene — fluoride ion

The behaviour of alkylbenzenes lightly halogenated in the nucleus and side chain is predictable on the basis of the separate effects of the two classes of halogen atoms. With perhalogenated compounds, new properties appear. For example, prolonged treatment of perchloro compounds with chlorine, in the presence of a catalyst or ultraviolet light, causes chlorinolysis of the side chain to give hexachlorobenzene and the fully chlorinated alkane. Treatment of perchloro compounds with iodide ion in acetic acid causes dechlorination and sometimes coupling (union of two molecules); for example,



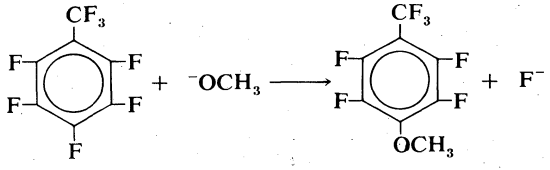
perchlorotoluene



perchlorostilbene

Chlorinated groups adjacent to the chlorinated benzene nucleus are easily hydrolyzed by heating with concentrated sulfuric acid: $C_6Cl_5CCl_2CCl_3 \rightarrow C_6Cl_5COCCl_3$. Trifluoromethyl groups in perfluoro- or chlorofluorobenzene homologues are similarly hydrolyzed to carboxyl groups. In all of these polyhalo compounds, but particularly in the fluorinated ones, the halogen atoms in the nucleus are susceptible to substitution by nucleophiles, such as lithium aluminum hydride, hydrazine, ammonia, methyllithium, and sodium alkoxides. In the perfluoro compounds, substitution occurs *para* to the perfluoroalkyl group (that is, in the position directly across the benzene ring).

Hydrolysis of side-chain halogens



perfluorotoluene — methoxide ion — *p*-methoxyperfluorotoluene — fluoride ion

Polyvalent iodine compounds (preparation and reactions). *Dichloroiodoarenes.* The iodine atom in an aryl iodide can exist in higher states of oxidation; *i.e.*, it can show valencies greater than one. Thus, when chlorine is bubbled into a solution of aryl iodide in cold dry chloroform, a dichloroiodoarene usually crystallizes out as a yellow solid. The formation of dichloroiodobenzene from iodobenzene occurs as follows:

$$C_6H_5I + Cl_2 \longrightarrow C_6H_5ICl_2.$$

These compounds, in which the iodine has a valence of three, are relatively unstable and decompose on heating to about 110° C (230° F). In the case of dichloroiodobenzene the product is *p*-chloroiodobenzene. Dichloro- and difluoroiodoarenes have been used as chlorinating and fluorinating agents.

Iodosyl compounds. When the dichloroiodoarenes are treated with aqueous alkali or aqueous pyridine, compounds containing the iodosyl group (—IO) are formed.

$$C_6H_5ICl_2 + 2NaOH \longrightarrow 2NaCl + H_2O + C_6H_5IO$$

dichloroiodobenzene; sodium hydroxide; sodium chloride; water; iodosylbenzene

Alternatively, these iodosyl compounds can be prepared directly from aryl iodides by oxidation. They behave as though they were anhydrides of the unknown bases, $ArI(OH)_2$, the salts of which can be made by dissolving the neutral iodosyl compounds in the appropriate acids (nitric, hydrochloric, hydrofluoric, and acetic).

Iodosyl compounds are oxidizing agents and liberate iodine quantitatively from acidified potassium iodide solution as follows:

$$C_6H_5IO + 2HI \longrightarrow C_6H_5I + I_2 + H_2O.$$

Iodyl compounds. Disproportionation (spontaneous oxidation and reduction) of iodosyl compounds on boiling with water or, more slowly, on standing, gives iodyl compounds and simple aryl iodides:

$$2C_6H_5IO \longrightarrow C_6H_5I + C_6H_5IO_2.$$

Iodyl compounds also can be made by oxidizing aryl iodides or iodosyl compounds.

Like iodosyl compounds, iodyl compounds are oxidizing agents that are quantitatively reduced by aqueous hydrogen iodide. They must be handled carefully because they explode on impact or on heating.

Iodonium compounds. A mixture of iodyl- and iodosylbenzene, on heating with silver oxide in aqueous suspension, gives a strongly alkaline solution of diphenyliodonium hydroxide:

$$2C_6H_5IO + 2C_6H_5IO_2 + Ag_2O + H_2O \longrightarrow 2(C_6H_5)_2I^+ OH^- + 2AgIO_3 .$$

The addition of a soluble iodide to a solution of a diphenyliodonium hydroxide yields a precipitate of a diphenyliodonium iodide. Stable salts of iodonium hydroxides can be formed by neutralization with acids. Such salts react with a great variety of nucleophiles, to give nuclear substituted products and aryl iodides:

$$(Ar-\overset{+}{I}-Ar)X^- + OH^- \longrightarrow ArOH + ArI + X^-.$$

They also decompose to substituted arenes on heating. For example:

$$(p\text{-}CH_3OC_6H_4-\overset{+}{I}-C_6H_5)Br^- \xrightarrow{180°} p\text{-}CH_3OC_6H_4I + C_6H_5Br.$$

p-methoxyphenylphenyliodonium bromide; *p*-iodomethoxybenzene; bromobenzene

TECHNICAL AND ANALYTICAL ASPECTS OF ORGANIC HALOGEN COMPOUNDS

Commercial uses. The polychloroalkanes and -alkenes are commonly used as solvents for drycleaning, chiefly because of their nonflammability. The most common are carbon tetrachloride and the trichloro- and tetrachloroethylenes. Chloroform is used as a solvent under controlled conditions, but its anesthetic properties make it dangerous. Because of their heavy vapours and their nonflammability some organic halogen compounds (carbon tetrachloride and dibromodifluoromethane, especially) have been used as fire extinguishers. Chloro compounds can be dangerous because under these conditions they give phosgene ($COCl_2$), a poisonous gas. The simple chlorofluoromethanes and -ethanes are used as refrigerants (Freons) and also, because of their great chemical stability, as propellants in aerosols, such as air fresheners and whipped cream.

Use as solvents

Chloroform was used for many years (mixed with acetone and ether) as an anesthetic. It is dangerous, however, because of the low margin between the lethal dose and the anesthetic dose and because the substance can damage the liver. It has now been completely superseded by fluothane ($CF_3CHBrCl$). Ethyl chloride is used as a local anesthetic.

One of the most common polymers is polyvinyl chloride. It is used mainly for electrical insulation and artificial fabrics for rainwear and upholstery. Tetrafluoroethylene and chlorotrifluoroethylene give Teflon (Fluon) and Kel-F polymers, respectively; these have very high thermal stability and electrical resistance. Because of the low intermolecular forces between highly fluorinated compounds, the lower molecular weight Kel-F polymers are used as lubricants, and Teflon is used to pack bearings in motors. Copolymers of hexafluoropropene and vinylidene fluoride (Vitons) are among the most thermally stable elastomers. Electronic apparatus and switchgear are sometimes immersed in fully fluorinated liquid alkanes (Flutec PP5) and cycloalkanes (Flutec PP3), which conduct away the heat developed during operation and act as insulators.

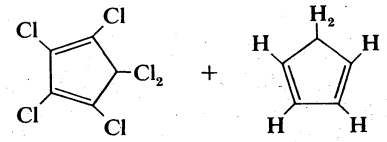
Use as polymers

Polychloro compounds are most important as insecticides. One of the earliest was DDT (dichlorodiphenyltrichloroethane), which is made by condensing together chlorobenzene and chloral:

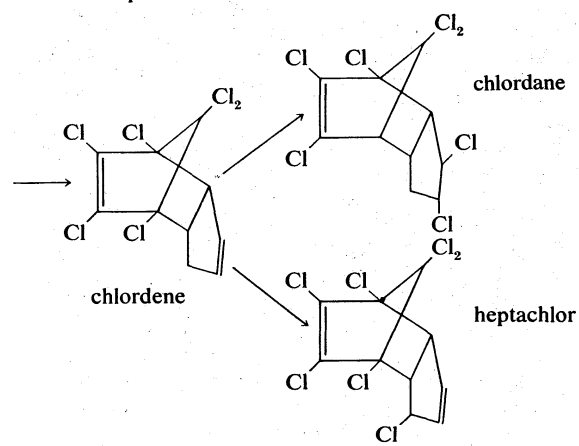
$$2C_6H_5Cl + CCl_3CHO \xrightarrow{H_2SO_4} (p\text{-}ClC_6H_4)_2CHCCl_3 + H_2O.$$

Use as insecticides

Gammexane, the gamma isomer of hexachlorocyclohexane, has properties and uses similar to those of DDT. It is made by the free radical addition of three molecules of chlorine to benzene. Of the nine possible isomers, eight are formed, and—of these—only the gamma form is active. Another important group of insecticides is derived from hexachlorocyclopentadiene. Addition of cyclopentadiene to the hexachloro compound gives chlordene. This substance can be converted either to chlordane by addition of chlorine or to heptachlor by substitution of chlorine for a hydrogen atom:



hexachlorocyclopentadiene; cyclopentadiene



Use of norbornadiene rather than cyclopentadiene in the above reaction gives aldrin, which can be converted to its epoxide dieldrin. Two similar insecticides, prepared by a slightly different route, are isodrin and endrin. The use of these highly chlorinated compounds has been criti-

cized because their stability causes them to accumulate in fish, birds, and animals. Monofluoroacetamide has been used as a systemic insecticide to kill black flies on plants such as roses; sodium monofluoroacetate is used as a rodenticide. Both are dangerous to human beings. Their volatile derivatives (*e.g.*, fluoroacetic acid and ethyl fluoroacetate) were among the first nerve gases.

Use as herbicides

Chlorinated and fluorinated benzenoid and heterocyclic compounds, and also urea derivatives, are used as herbicides. The reaction of chloral and urea gives dichloralurea, which is used as a soil sterilant to prevent the growth of weeds. The heterocyclic compound diquat (1,1′-dimethylene-2,2′-bipyridylium dihalide) and paraquat (1,1′-dimethyl-4,4′-bipyridylium dihalide) are used to destroy grasses and broad-leaved weeds. The substance 3,5-dichloro-4-hydroxydifluoropyridine (Teklon) is used for the control of wild oats and couch grass.

The well-known herbicides 2,4-D and 2,4,5-T (respectively, 2,4-dichloro- and 2,4,5-trichlorophenoxyacetic acids) are used to kill broad-leaved plants and smaller shrubs, respectively.

The compound 5-fluorouracil is thought to replace uracil in nucleic acids and hence interfere with the biochemical operation of living cells. It is used in the treatment of some forms of cancer. The introduction of one or two fluorine atoms into steroid molecules either by replacing hydrogen or by addition across a double bond often enhances the hormonal activity of the steroid. Certain halogenated steroids also are used to reduce skin inflammation.

Separation, identification, and analysis. The introduction of a bromine or iodine atom into a hydrocarbon molecule increases the boiling point of the substance by about 100° to 150° C (180° to 270° F) and creates bonds that are less stable than carbon–hydrogen bonds. In the main, however, alkyl halides, aryl halides, and most polyhalo compounds are stable. The most common are gases or liquids, and the most usual way to purify them is by fractional distillation. Unfortunately, however, polyfluoro compounds have a pronounced tendency to form constant-boiling mixtures, and—as a result—they are often not separable in this way. The technique known as vapour phase chromatography has proved useful in this case, and in fact this method is now used extensively for the separation of mixtures of reasonably volatile halo compounds (with boiling points up to 200° C, 392° F) of all kinds. Fundamentally, vapour phase chromatography is very similar to fractional distillation, the vapour of the mixture in a stream of nitrogen being adsorbed on and desorbed from a column of nonvolatile solvent supported on a substance like kieselguhr. By varying this stationary phase (dinonyl phthalate, silicone oil, and fluorolube are common materials used) it is usually possible to separate any mixture.

Spectroscopic identification

Apart from the use of melting and boiling points, pure organic halogen compounds are identified most often spectroscopically. There are strong characteristic absorptions in the infrared arising from stretching vibrations of the carbon-to-halogen bonds. Because the frequencies at which these absorptions occur are influenced by neighbouring groups, they give valuable information about the molecular structure, revealing, for example, whether the halide is an alkyl or aryl derivative. In the mass spectrometer, which separates materials on the basis of their relative masses, organic halides are usually sufficiently stable to show an ion derived from the intact molecule. As a result, an accurate molecular weight can easily be found. The breakdown pattern of the ions often gives strong clues to the structure of the parent compound.

Nuclear magnetic resonance spectroscopy, which measures the energy required to change the alignment of magnetic nuclei in a magnetic field, is used to determine the structures of compounds containing hydrogen and fluorine atoms. The resonance peaks associated with these atoms appear at different locations in the spectra of various compounds depending on the atoms adjacent to them. By finding the number of atoms in each different environment, the structure of a compound often can be determined.

By measuring another atomic property, the nuclear quadrupole moment, the structures of compounds containing chlorine and bromine can often be deduced. It is probable, however, that in the future the structures of chlorine-, bromine-, and iodine-containing compounds will best be determined by X-ray induced electron emission spectroscopy. This technique is based on the magnetic or electrostatic analysis of the electrons that are emitted from a substance on irradiation with X-rays. The energies of these electrons are characteristic of the atom from which they are emitted and of the environment of that atom. Hence, the results can be used both for elemental analysis and for structure determination. This method, however, was in its infancy in the early 1970s.

Elemental analysis

Conventional elemental analysis of halogens in organic compounds involves, first, fusion of the compound with sodium or potassium, and then estimation of the alkali metal halides. Fluoride can be determined by weight as lead chlorofluoride, and chloride by titrating the alkali liberated by its action on mercuric oxycyanide. Bromide is quantitatively oxidized to bromate by sodium hypochlorite, and this, in turn, is used to liberate iodine from hydrogen iodide. In a similar fashion, iodide is converted to iodate by hypobromite, and this, again, is used to liberate iodine from hydrogen iodide. In both cases the free iodine is titrated with standard sodium thiosulfate.

BIBLIOGRAPHY. W.K.R. MUSGRAVE, "Halogen Derivatives of the Aliphatic Hydrocarbons," in S. COFFEY (ed.), *Rodd's Chemistry of Carbon Compounds*, 2nd rev. ed., vol. 1A, ch. 3 (1964); W.J. FEAST, "Halogen Derivatives of the Aliphatic Hydrocarbons," *ibid.*, vol. 1A, ch. 3 suppl. (1971); W.J. FEAST and W.K.R. MUSGRAVE, "Halogen Derivatives of Benzene and Its Homologues," *ibid.*, vol. 3A, ch. 3 (1971)—together these chapters describe the preparation and reactions of halogenated aliphatic and benzenoid hydrocarbons, including some theoretical interpretation of the chemistry involved. W.A. SHEPPARD and C.M. SHARTS, *Organic Fluorine Chemistry* (1969), describes the chemistry of aliphatic, aromatic, and heterocyclic fluorine-containing compounds and their commercial application; R.O.C. NORMAN and R. TAYLOR, *Electrophilic Substitution in Benzenoid Compounds* (1965), and L.M. STOCK, *Aromatic Substitution Reactions* (1968), describe in detail theories concerning nucleophilic and electrophilic substitution in aromatic compounds.

(W.K.R.M.)

Organic Nitrogen Compounds

Since, by definition, organic compounds are compounds containing carbon, organic nitrogen compounds are substances the molecules of which contain at least one carbon and one nitrogen atom. Several hundred thousand such compounds are known, and new ones are being prepared or discovered continually. Organic nitrogen compounds are present in all known living organisms, as essential and often major components. Flesh, hair, horn, milk solids, and blood, for example, are composed largely of organic nitrogen compounds. Many drugs and medicinal agents, both natural and synthetic, including most narcotics, local anesthetics, the sulfa drugs, penicillin; most explosives, such as nitroglycerine and TNT (2,4,6-trinitrotoluene); some rocket fuels; many dyes; and some synthetic polymers (substances of high molecular weight made up of a number of identical or similar groups of atoms bonded together), such as nylons and the melamine resins, belong to this class. For detailed background information see CHEMICAL BONDING; MOLECULAR STRUCTURE; CHEMICAL COMPOUNDS, ORGANIC; CHEMICAL REACTIONS.

No single characterization can be made to apply to the properties of organic nitrogen compounds. Some are gases—*e.g.*, hydrogen cyanide (formula HCN) and methylamine (CH_3NH_2); some liquids—*e.g.*, nitroglycerin

$$[O_2NOCH_2CH(ONO_2)CH_2ONO_2]$$

and aniline ($C_6H_5NH_2$); but most are solids. Some mix completely with water; others are quite insoluble. They may be acidic, basic, or neutral, coloured or colourless, volatile or nonvolatile, intensely poisonous or essential to life.

BONDING IN NITROGEN AND ITS COMPOUNDS

Bonding characteristics of nitrogen. The wide variation in properties of members of the group is directly related to the ability of nitrogen to take part with other elements in a large number and variety of molecular structures. This ability results from the capacity of the nitrogen atom to have a large number of so-called oxidation states (*i.e.*, extents of being joined to more electronegative atoms; see below).

Chemical bonds result from a rearrangement of the electron structures of two or more atoms when they combine to form a molecule or a solid metal or a salt composed of ions (charged atoms). Every atom has a nucleus carrying a number of positive charges (called the atomic number, the largest number known being 105), surrounded by an equal number of negatively charged electrons. The electrons are arranged in orderly fashion in concentric shells and subshells; and their positions are defined by quantum mechanics in terms of energy requirements for the structure. An electron may change position as a result of acquiring or losing energy, and it is then defined by the new position it occupies. Certain configurations of the electron structure are more stable than others, and in the most stable the outermost shell has eight electrons, called the octet. The only neutral atoms with eight outermost electrons are the noble gases. Other atoms, with more or less than eight, lose or acquire one or more electrons, thereby becoming positively or negatively charged ions. Oppositely charged ions attract one another to form electrovalent, or ionic, bonds (as seen, for example, in crystals in which the bonding electrons involved are completely associated with the negative ions). Atoms may also form covalent bonds by sharing one or more pairs of electrons, which hold them together in discrete molecules. In still another form of bonding, the atoms release the electrons of their incomplete outermost shells to the aggregate of atoms, a condition found in metals. Ions may have as many as four positive or negative charges. Bonds of the covalent type may involve one, two, or three pairs of electrons and are termed single, double, or triple bonds; the chemical formula for a compound may show the bonds as a single line for each pair of electrons, thus: H—H, N=O, and N≡N, respectively. In complex molecules, both covalent and ionic bonds may be present.

Nitrogen has atomic number 7; its nucleus thus has a charge of +7 and is surrounded by seven electrons, each having a charge of −1. Two of the electrons fill an inner shell, and the remaining five occupy a second shell, which has a total capacity of eight. Because stable structures generally result when shells are filled, the incomplete second shell of nitrogen is responsible for its chemical combining power. The octet may be filled by electrons that are either shared or acquired.

Oxidation state

The electron pairs of covalent bonds are not shared equally when the bonded atoms are different, a fact that gives rise to the concept of oxidation state (or oxidation number), which is defined as the number of bonds shared with electronegative elements (*i.e.*, those to the right in the periodic table) minus the number of bonds shared with electropositive elements (*i.e.*, those to the left in the periodic table). The oxidation number of nitrogen ranges from −3, as in ammonia (NH_3), through zero, as in elemental nitrogen (N_2), to +5, as in nitric acid (HNO_3), and nitrogen occurs in known organic compounds in these and all intermediate oxidation states (−2, −1, +1, +2, +3, +4).

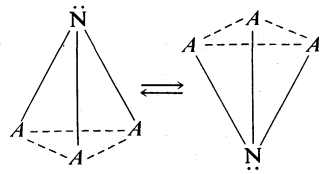
Organic nitrogen compounds are classified on the basis of the inorganic (*i.e.*, not containing carbon) constituent that can be considered the parent compound; *i.e.*, the compound from which the organic compound is produced by removing a hydrogen atom and replacing it with a group of atoms containing carbon. Nitrogen occurs in nature bound to hydrogen, oxygen, carbon, and other nitrogen atoms; it can also form bonds with other elements, including sulfur, phosphorus, fluorine, chlorine, bromine, and iodine.

Nitrogen-compound structures. In almost all organic nitrogen compounds, the three additional electrons required to fill the octet are obtained by sharing an electron pair with another atom, which contributes one electron to the shared pair, thus forming a covalent bond. Three such bonds, adding three electrons to the normal five, result in a filled octet; the fourth pair of electrons of the nitrogen atom is not shared. An example is the reaction shown below (in which *A* represents any atom with one electron available to share, and each electron is represented by a dot); one nitrogen atom combines with three *A* atoms to form a compound in which three pairs of electrons are shared in covalent bonds, and the outer electron shell of the nitrogen atom is filled.

$$\cdot\ddot{N}\cdot + 3A\cdot \longrightarrow A{:}\ddot{\underset{\ddot{A}}{N}}{:}A$$

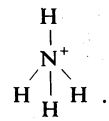
Shapes of nitrogen compounds

It has been shown by experiment and is found consistent with theoretical views that the orbitals (*i.e.*, the regions around the nucleus occupied by the four electron pairs) point to the four corners of a tetrahedron (triangular pyramid). Simple compounds in which three atoms share covalent bonds with nitrogen thus have the shape of a flattened pyramid, with the nitrogen atom at the apex and the unshared electron pair above it. In fact, however, this shape is very mobile and inverts (turns inside out like an umbrella) very rapidly. The interaction between the two forms is represented by an equation in which each bonded pair of electrons is a single line, and each single electron is a dot.



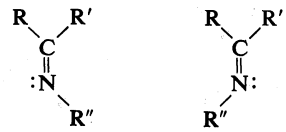
The unshared electron pair is available to form a fourth bond by sharing with another atom having an incomplete outer shell. Such atoms include oxygen (O), hydrogen ion (H^+; an ion is any electrically charged atom or molecule), and carbonium ion (*e.g.*, R_3C^+), a positively charged organic ion in which the outer shell of a carbon atom contains only six electrons shared in covalent bonds to three atoms or groups of atoms, the behaviour of which is explained by assuming that the resulting positive charge is localized on the carbon atom. In the formula, *R* can represent any of a number of groups of atoms bonded to the carbon atom; the symbol *Ar* (for example, ArH_2C^+) often is used to represent an aromatic group; *e.g.*, one having the benzene ring (a six-sided cyclic compound, formula C_6H_6).

The fourth bonding results in the formation of a tetrahedral molecule (or ion) with all four corners occupied by other atoms; an example is ammonium ion, NH_4^+, which can be represented:



Such species of bonded atoms maintain their form more or less rigidly when none of the four surrounding atoms is a hydrogen, and, if all four bonded atoms are different, the same kind of dissymmetry (left- and right-handedness, in reference to the position of the four relative to each other) as is encountered in carbon compounds may arise.

When a nitrogen atom shares two electron pairs with the same atom, a double bond results. The atoms joined by the double bond and the atoms attached directly to them then lie all in the same plane, together with the unshared electron pair on the nitrogen. Such compounds may have two geometrically isomeric forms (*i.e.*, forms differing only in the arrangement in space of the groups about a double bond); for example, many imines (compounds in which nitrogen is doubly bound to carbon) have isomers that may be represented as



The formation of a triple bond by sharing three electron pairs is most commonly encountered between a nitrogen and a carbon atom. Such compounds, called nitriles, have a linear structure (*i.e.*, the atoms lie in a straight line), as in the case of a single radical (*R*) bonded to a carbon atom that also shares a triple bond with a nitrogen atom, expressed as $R—C{\equiv}N$. (Molecular nitrogen and certain less common classes of compounds, the azides and diazonium salts, are considered to have triple bonds between two nitrogen atoms.)

CLASSIFICATION OF ORGANIC NITROGEN COMPOUNDS

Parent compounds

The parent inorganic compounds from which organic nitrogen compounds are obtained include ammonia (NH_3), hydrazine ($H_2N—NH_2$), hydroxylamine ($H_2N—OH$), hydrogen azide (HN_3), nitrous acid (HNO_2), nitric acid (HNO_3), and a few very unstable substances: diazene, or diimide, (HN=NH), nitroxyl (HNO), amino radical ($H_2N\cdot$), aminoxyl ($H_2N—O\cdot$), triazene ($H_2N—N{=}NH$).

Amines. Compounds derived from ammonia by replacement of one or more hydrogen atoms by hydrocarbon groups, which have varying numbers (indicated by *m* and *n*) of carbon and hydrogen atoms, C_mH_n, are known as amines; alkyl (*R*), or aryl (*Ar*). When only one hydrogen atom is replaced, the resulting compounds are called primary amines and are represented as RNH_2; when two hydrogens are replaced, they are secondary (R_2NH), and, when all three hydrogens are replaced, they are tertiary (R_3N). All such compounds resemble ammonia in that they are basic (*i.e.*, they form salts with acids by virtue of the unshared electron pair). As the hydrocarbon portion (the part consisting of groups containing only carbon and hydrogen atoms) of the molecule becomes larger, the amines become less volatile (*i.e.*, have less tendency to vaporize at ordinary temperature), less soluble in water, and generally more like hydrocarbons. Amines have lower melting points than most other classes of nitrogen compounds, and are more strongly basic. Most naturally occurring compounds of nitrogen are amines or derivatives of them.

Amines are named by prefixing the name of the attached groups to the stem amine, as in dimethylamine $(CH_3)_2NH$. If an aryl group such as the phenyl radical, $—C_6H_5$ (from benzene, C_6H_6), is attached, the compound is usually considered to be a derivative of aniline and is named as substituted aniline, as in 4-chloro-*N*-methylaniline ($ClC_6H_4NHCH_3$).

Derivatives of the ammonium (the "-ium" ending indicates, as in carbonium carbon, the presence of a localized positive charge) ion bearing four hydrocarbon groups (R_4N^+) are quaternary ammonium salts. The quaternary ammonium ions themselves are not basic (there being no longer an unshared electron pair), but their hydroxides (compounds in which the positive charge on the nitrogen atom is neutralized by the negative charge of a hydroxide ion, OH^-), $R_4N^+OH^-$, are strong bases that resemble sodium and potassium hydroxides. They are named by prefixing the names of the substituents to the stem ammonium, as in the tetramethylammonium ion ($[CH_3]_4N^+$).

Double bonds

Imines. Compounds in which nitrogen is bound to carbon by a double bond are known as imines (sometimes called Schiff bases). Imines are distinctly weaker bases than the amines. They are most commonly named by following the name of the related carbonyl compound by the word imine, as in diethyl ketone *N*-methylimine, $(C_2H_5)_2C{=}N—CH_3$, related to diethyl ketone, $(C_2H_5)_2C{=}O$.

Carbinolamines. Closely related to the imines, although containing no double bond, are the carbinolamines. They are produced by addition of the elements of water to the double bond and may usually be dehydrated to form imines (providing there is a hydrogen on the carbinolamine nitrogen atom). They are given names on the basis of being amines bearing a 1-hydroxy substituent (OH), as in 1-hydroxyethyldimethylamine,

$$CH_3CH(OH)N(CH_3)_2.$$

Triple bonds

Nitriles. The nitriles have a triple nitrogen-to-carbon bond (which is written $—C{\equiv}N$), generally have low melting points, and are for ordinary considerations totally nonbasic (do not neutralize acids to form salts). They have a close affinity with carboxylic acids—*i.e.*, organic acids containing the carboxylic group, —COOH, such as acetic acid, CH_3COOH, and benzoic acid, C_6H_5COOH; the carbons are in the same oxidation state, having in each case three bonds to an atom having higher affinity for electrons (oxygen or nitrogen). They are named by using the root name of the corresponding acid, with the suffix nitrile, as in acetonitrile, $CH_3C{\equiv}N$ (from acetic acid, CH_3COOH).

Amides. The replacement of an acidic hydroxyl group by an amino group produces the class of nitrogen compounds known as amides. The carboxamides, $RCONH_2$, are the most important group; other substituent groups may be attached to the nitrogen. Amides known as sulfonamides, RSO_2NH_2, are also produced from the sulfonic acids, RSO_3H. Most amides are solids, apt to have high melting points, and are essentially nonbasic and of low volatility. Most are named from the corresponding acid, using the suffix "-amide," as in acetamide, CH_3CONH_2, and *N*-ethylbenzenesulfonamide,

$$C_6H_5SO_2NHC_2H_5.$$

Imides. Compounds of the imide class have two carbonyl (*i.e.*, containing the organic radical $>C=O—$ *e.g.*, acetyl, $CH_3CO—$) groups attached to the same nitrogen, as in acetimide, $(CH_3CO)_2NH$. They are weakly acidic (*i.e.*, neutralize alkalies to form salts) if the nitrogen is not otherwise substituted. If both oxygens of a carboxyl group (—COOH, the radical characteristic of most organic acids) have been replaced by nitrogen, the amidine structure results. Amidines are moderately strong bases. The general formula is $RC(NH_2) = NH$.

Cyanic acid

Carbonic acid derivatives. The compounds derived from carbonic acid are usually considered as a class by themselves. The nitrile of carbonic acid is cyanic acid, $HO—C{\equiv}N$, which in turn can provide two groups of derivatives: the cyanates, $R—O—C{\equiv}N$, and the isocyanates, $R—N{=}C{=}O$. They are named by combining the class name with the name of the substituent, as in ethyl isocyanate, $C_2H_5—N{=}C{=}O$.

The amide of cyanic acid, cyanamide, $H_2N—C{\equiv}N$, and its tautomer (tautomers are compounds of the same chemical formula but different molecular structure that interconvert reversibly) carbodiimide, $HN{=}C{=}NH$, also yield a group of derivatives.

Carbonic acid, HOCOOH, which has two hydroxyl groups, reacts with ammonia to produce amides in two stages, carbamic acid, NH_2COOH, and urea,

$$NH_2CONH_2.$$

Carbamic acids are unstable in the free state but their salts and esters are well known. A salt is the compound other than water formed by reaction of an acid with a base, and an ester is a derivative of an acid in which a hydrocarbon group is attached to an oxygen atom in place of the acidic hydrogen atom. The esters are often called urethans, after the primitive name of the ethyl ester, $C_2H_5O—CO—NH_2$. Urea and its derivatives are almost invariably solids of high melting point and low volatility and are commonly named as substituted ureas, as *N,N'*-dimethylurea, $CH_3—NH—CO—NH—CH_3$. Thiourea is the compound in which a sulfur atom replaces the oxygen atom of urea.

The amidine of carbamic acid goes by the special name guanidine. The guanidines, $H_2NCNHNH_2$, are in general very strong bases, approaching hydroxide ion (OH^-) in strength.

Hydroxylamine compounds. Nearly all the various derivatives of ammonia (NH_3) have their parallel with

hydroxylamine ($H_2N—OH$). Hydroxylamine may carry one, two, or three substituents, which may be bonded to either the nitrogen or the oxygen atom; *e.g.*, *N*,*N*-dimethylhydroxylamine, $(CH_3)_2N—OH$, and ethoxylamine, $C_2H_5—O—NH_2$. Such compounds are similar to amines but are weaker bases. Quaternary compounds corresponding to those of ammonia are of two types: tetrasubstituted salts and amine oxides.

The hydroxylamine compounds analogous to imines are the oximes, $RR'C{=}NOH$, and nitrones,

$$RR'C=\overset{+}{N}RO^-.$$

Nearly all are solids, but the oximes and derivatives have considerably lower melting points than the nitrones.

The most important hydroxylamine derivatives of the carboxyl group are the *N*-acyl compounds, called hydroxamic acids, $RC(=O)NHOH$.

Hydrazine compounds. Hydrazine is also the parent of a family of structures analogous to the organic derivatives of ammonia. There are compounds of from one to four substituents (mono-, di-, tri-, and tetrasubstituted hydrazines); all are weakly basic. Quaternary hydrazinium compounds

$$R_2N-\overset{+}{N}R_3\ X^-$$

(in which X^- represents any negatively charged atom or group of atoms) are also well known. The analogues of imines, $R_2C{=}N—NR_2$, are called hydrazones and may have any of a range of organic groups bonded to the nitrogen atom. Azines, $R_2C{=}N—N{=}CR_2$, have two imine groups joined by an N—N bond. The only important hydrazine derivatives of the carboxyl group are the hydrazides, which are weakly basic.

Diazene compounds. The derivatives of diazene that have two substituted groups (*i.e.*, are disubstituted), $R—N{=}N—R$, are called azo compounds, such as azobenzene, $C_6H_5—N{=}N—C_6H_5$. They are neutral substances, generally coloured and solid. Monosubstituted diazenes are known but are very unstable.

The diazonium compounds,

$$Ar-\overset{+}{N}\equiv N\ X^- \rightleftharpoons Ar-N=N-X,$$

are closely related. The salts of strong acids are truly ionic (*i.e.*, electrically charged) and show the properties to be expected of a triple bond. They are generally rather unstable substances but of great importance as intermediates (chemicals produced as necessary steps between starting material and finished product) in synthesis. The diazo compounds, such as diazomethane,

$$CH_2=\overset{+}{N}=\overset{-}{N} \rightleftharpoons \overset{-}{C}H_2-\overset{+}{N}\equiv N,$$

also show evidence of a triple bond.

Compounds with three or more nitrogens. Triazenes and azides are among the types of compounds having chains of three or more nitrogens. The word triazene denotes three nitrogen atoms and a double bond, $H_2N—N{=}NH$, and, in analogous fashion, names for longer chains and different degrees of unsaturation (that is, capable of combining directly with additional groups of atoms) may be developed.

Azides of various types, such as methyl azide, $CH_3—N_3$, phenyl azide, $C_6H_5—N_3$, and acetyl azide, $CH_3CO—N_3$, all contain the linear (*i.e.*, all the atoms lie in a straight line) azido group

$$-N=\overset{+}{N}=\overset{-}{N} \rightleftharpoons -\overset{-}{N}-\overset{+}{N}\equiv N.$$

The nitroso and nitro compounds, $R—N{=}O$ and $R—NO_2$, represent higher oxidation states. Both are neutral, and the former are all a beautiful blue or green, whereas the latter are colourless to yellow. Most nitroso compounds, such as nitrosobenzene, $C_6H_5—N{=}O$, are in equilibrium with colourless dimers (a dimer is a molecule formed by the union of two simpler molecules), with the nitrogen molecules bonded together. Both classes of compound are invariably named by means of the "nitroso-" or "nitro-" prefix.

Nitrous and nitric acid compounds. The highest oxidation states of nitrogen are represented by the derivatives of nitrous acid (HNO_2) and nitric acid (HNO_3). Among them are the esters, $R—O—N{=}O$ and $R—O—NO_2$, which contain no carbon-to-nitrogen bond. The nitrites that are exceptionally volatile, and the nitrates, such as nitroglycerin (or glyceryl trinitrate), are potentially explosive in proportion as the ratio of oxygen to carbon and hydrogen in the compound approaches the optimum for complete internal combustion to water, oxides of carbon, and molecular nitrogen. Nitrous and nitric acids can give rise to amides, hydroxamic acids, and hydrazides, just as do other acids. All of them are known in the form of organic derivatives.

The volatility of nitrites and nitrates

THE AMINES AND THEIR DERIVATIVES

Physical and chemical properties. The smallest amines are gases, but most amines are liquids, and all of them except the very large have fishy or musty odours. The volatility (*i.e.*, the tendency to vaporize at ordinary temperatures) decreases with increasing molecular weight and increasing number of hydrogens attached to the nitrogen. Primary amines thus have the highest boiling points and tertiary the lowest for comparable molecular weights. The smaller amines are soluble or even completely miscible in water (that is, can be mixed completely without separating into two layers) but, when the size of the molecule exceeds five carbon atoms, the solubility drops rapidly. Nearly all amines are soluble in the common organic solvents. Aliphatic amines (*i.e.*, those in which the nitrogen atom is not attached to a benzene ring) are slightly stronger bases than ammonia and will turn litmus indicator paper blue, but aromatic amines (those in which the nitrogen atom is attached to a carbon atom present in a benzene ring) are considerably weaker, and most of them will not affect litmus.

Aliphatic amines

Salts. Nearly all salts of amines are nonvolatile solids, very soluble in water, to which they give a weak acidic reaction if they are salts of strong acids, such as hydrochloric, nitric, or sulfuric; generally they do not dissolve in nonpolar solvents such as ether and the hydrocarbons. Quaternary ammonium compounds are completely ionic in nature—the hydroxides are alkalies as strong as sodium hydroxide (lye).

Sources and preparation. Some simple aliphatic amines occur naturally, mostly as the products of decay, but there are no practical natural sources. Industrially, aliphatic amines are mostly made from fatty acids (organic acids derived from natural fats and oils), olefins (unsaturated aliphatic compounds whose molecules contain only carbon and hydrogen atoms; *i.e.*, they are hydrocarbons), or alcohols (alcohols are hydroxyl compounds, containing the oxygen–hydrogen radical, OH, bonded to a saturated carbon). The fatty acids (or the fats and oils in which they occur bound) may be reduced to alcohols (one important step in production of detergents), which can then be converted to amines catalytically by reaction with ammonia or by indirect, laboratory methods involving initial conversion to a suitably reactive derivative. Alternatively, the fatty acids may be converted to nitriles (by way of their amides) by reaction with ammonia, followed by hydrogenation (combination of hydrogen with another substance, usually an unsaturated organic compound). These reactions also give rise to secondary and tertiary amines.

A few amines, notably aniline and some of its derivatives, are made by reduction (reduction is an increase in the number of electrons) of nitro compounds, although in recent years it has become more common to produce aniline from chlorobenzene in a continuous process under high pressure.

Reductive alkylation (a process of adding alkyl radicals and hydrogen), whereby an aldehyde or ketone is treated with ammonia and hydrogen in the presence of a catalyst, has both industrial and laboratory utility.

There are three types of reaction by which amines may be prepared on a laboratory scale: displacements on halides or derivatives; reduction of nitriles, amides, oximes, nitro compounds, etc.; and rearrangements (Hofmann,

Curtius, Lossen, Beckmann, Schmidt). Many variations of them are described in specialized works.

Chemical behaviour of amines

The chemical behaviour of amines is dominated by their basicity. Nearly all of them form salts even with moderately weak acids, and the aliphatic amines even extract carbon dioxide from moist air to form salts. Since salts in general are soluble in water, most amines will thus dissolve in dilute solutions of acids. Salts also generally crystallize well, and, by reaction with acids, liquid or gaseous amines can be converted to solid derivatives useful for identification.

Oxidation characteristics. Aliphatic amines are not easily oxidized (oxidation is loss of electrons), but aniline and its derivatives commonly become coloured slowly on prolonged contact with air, owing to oxidation of the activated benzene ring to form complex dyes. Since amines are the lowest oxidation state of nitrogen, they are not ordinarily reducible (except for hydrogenolysis of benzylamines and some quaternary ammonium salts, in which a C—N bond is broken with addition of hydrogen to each fragment).

Chlorine and bromine replace the *N*-hydrogens of amines to form *N*-haloamines. Tertiary amines undergo more deep-seated changes.

The reaction of amines with nitrous acid is an old and important reaction. It appears that all primary amines give a diazonium salt, $R\text{–}N_2^+X^-$, but only with aniline and its derivatives can such products be isolated (or even detected). The observed result with aliphatic primary amines is the formation of mixtures, consisting of olefins, cyclopropanes, and isomeric alcohols, with evolution of nitrogen. The formation of nitrogen gas (effervescence) has been used as a qualitative test for primary amines. From secondary amines, nitrosamines precipitate as nonbasic, yellowish oils, and tertiary amines are degraded to mixtures of nitrosamines with aldehydes or ketones derived from one of the alkyl groups.

Aromatic primary amines are a special case. Cold (about 0° C), they react with dilute nitrous acid to form solutions of diazonium salts, which easily lose nitrogen on warming and form phenols ($ArOH$).

Alkyl halides, sulfates, and related alkylating agents react more or less readily with free amines so as to add an alkyl group to the nitrogen atom. The process may be repetitive, especially if an inorganic base is present, so that most amines are eventually converted to quaternary salts, if bulk does not interfere.

Reactions with other organic compounds. The reactions of amines with other classes of organic compounds form a vast subject. In the most general terms, they do not react with alcohols or olefins except at high temperatures in the presence of surface catalysts, when they may become alkylated on the nitrogen. Aldehydes and ketones will usually react with primary amines to form imines by addition at the carbonyl group (C=O) followed by elimination of water, as shown in the equation

$$R-NH_2+R'_2C=O \longrightarrow R'_2C(OH)-NHR \longrightarrow R'_2C=NR+H_2O.$$

Carboxylic acids form salts, which may lose water when heated very strongly, forming amides, as in the reaction

$$R-NH_2+R'CO_2H \longrightarrow R-NH_3^+RCO_2^- \xrightarrow{\Delta} R'CONH-R+H_2O.$$

Primary and secondary amines react to form amides, vigorously with acid chlorides, slightly less vigorously with acid anhydrides, and extremely sluggishly with esters.

$$RCOX+2R'NH_2 \longrightarrow RCONHR'+R'NH_3^+X^-$$

Amines do not react with other bases, weak or strong, except insofar as amines are liberated from their salts. Quaternary ammonium salts, however, undergo the Hofmann reaction when they are converted to the hydroxides (usually by treating them with moist silver oxide) and then heated strongly. One group is cleaved from the nitrogen as an olefin leaving a tertiary amine.

$$R_3\overset{+}{N}-CH_2CH_2ROH^- \xrightarrow{\Delta} R_3N+CH_2=CHR+H_2O$$

Compounds with amide linkages. *Physical properties.* The carboxamides and imides are mostly solids of low volatility and thus little odour, although they are likely to have flavour. Unsubstituted amides, $RCONH_2$, have the highest melting and boiling points; substitution of hydrogen by alkyl groups (paraffin hydrocarbon radicals) reduces the intermolecular attraction by removing the capability for hydrogen bonding between N—H of one molecule and C=O of another. The small amides and imides (less than six carbon atoms) are soluble in water, and formamide is completely miscible.

Common characteristics of amides

Similar remarks can be made about most of the other types of amides: thionamides, sulfonamides, ureas, urethans (carbamates), amidines, and imidates (imino esters). The last two types, however, which have no carbonyl group, have lower melting points, and the simple imidates are liquid at room temperature.

Amides in general are neutral in water solution, but they show a feeble basicity toward very strong acids, especially in the absence of water. Imidate esters, however, are almost as basic as alkylamines, and amidines are the strongest organic bases short of the quaternary ammonium hydroxides. Sulfonamides show no basic properties, but those having at least one hydrogen on the nitrogen have acidities similar to those of phenols.

Sources and preparation. There are no practical natural sources of simple amides, although polyamides occur in great abundance as the protein of living systems. Simple amides are made industrially from ammonium salts of carboxylic acids by strong heating; some *N*-substituted amides are made analogously. Esters react with ammonia to form amides, but the general laboratory method is to treat an amine with an acid chloride or anhydride. It is usually desirable to add an equivalent of alkali (base, usually having a hydroxyl group that ionizes to hydroxide ion, OH^-, which then combines with the hydrogen ion, H^+, of an acid to form water, a process called neutralization) to neutralize the acid that would otherwise bind a second equivalent of amine (Schotten–Baumann procedure).

Imides do not occur naturally except for some complex heterocyclic compounds. For preparative purposes, imides can generally be prepared from amides by treating them with acid chlorides or anhydrides or with metallic sodium or potassium, but cyclic imides, derived from dicarboxylic acids, can in many instances be prepared by strongly heating the ammonium salt of the acid or the half-amide. Imides can also be prepared by heating nitriles with carboxylic acids.

The derivatives of carbonic acid are a special case. Urea is prepared commercially by heating ammonia with carbon dioxide, but substituted ureas are best prepared by the reaction of an isocyanate with an amine. Urethans are also made from isocyanates, by reaction with alcohols.

The most characteristic reaction of amides is hydrolysis (a chemical reaction with water), by which they are converted to an acid and an amine. Amides are generally inert to pure water and require a mineral acid, alkali, or enzyme to catalyze hydrolysis. Wool and nylon, both of which are polyamides, can thus be dissolved by heating them with strong acid or base such as potassium hydroxide (KOH). The equation is

$$RCONR_2+KOH \longrightarrow RCO_2K+R_2NH.$$

Amides can also be dehydrated, most commonly by heating with phosphorus pentoxide or phosphoryl chloride. Unsubstituted amides give nitriles, and monosubstituted amides give ketenimines, but disubstituted amides are inert. Monosubstituted formamides can be dehydrated to isocyanides.

Amides are not readily oxidized and are inert to most reducing agents. Catalytic hydrogenation will convert amides of aliphatic acids to amines at somewhat high pressures and temperatures, and electrolytic reduction is generally successful with aromatic amides, but the most generally effective reagent is lithium aluminum hydride.

Conditions may generally be chosen so as to favour either amines, alcohols, or aldehydes as the products.

Aldehydes but not ketones generally react with amides that have at least one hydrogen on the nitrogen. The first step is addition to the aldehyde carbonyl group, but this may be followed by reaction with a second molecule of amide and elimination of water. The technique of preserving flesh by immersion in formaldehyde solution makes use of these reactions, and the condensation of formaldehyde with urea (carbamide) produces the industrially important urea–formaldehyde resins.

$$H_2NCONH_2 + H_2C{=}O \longrightarrow H_2NCONHCH_2OH + (-NHCONHCH_2-)_n$$

Amides may react with acid chlorides or anhydrides, especially if a base is present, to form imides.

The halogen elements and the hypohalous acids, HO*X*, replace a hydrogen on the nitrogen in the presence of base. The resulting *N*-haloamides are not very stable and easily undergo the Hofmann rearrangement when warmed with excess base. The overall effect is to convert the amide to a primary amine having one fewer carbon atom.

Amides suffer replacement of the amino group by an aryl or alkyl group, to form ketones (formamides yield aldehydes) in two ways: reaction with Grignard reagents or treatment with an aromatic hydrocarbon, such as benzene, in the presence of phosphoryl chloride (Vilsmeier–Haack reaction).

Compounds with double and triple C—N bonds. *Imines, nitriles, isocyanates, etc.* The simplest members of each class in this section are liquids, although little more can be said about the small imines, for they polymerize so readily. Hydrogen cyanide is actually the smallest nitrile. Acetonitrile has an unusually high dielectric constant and is a polar solvent capable of dissolving many salts as well as most other organic compounds. The isomer methyl isocyanide is a liquid of exceptionally revolting stench; methyl isocyanate has a pungent, biting odour; and methyl isothiocyanate smells somewhat savory as well as pungent. Cyanamide forms transparent crystals. The isomeric (forming isomers) carbodiimides polymerize too rapidly for the simpler ones to have been obtained pure, but in general they are liquids similar to the cyanamides.

The small imines, nitriles, cyanamides, and carbodiimides are miscible or soluble in water, but isocyanides, isocyanates, isothiocyanates, cyanates, and thiocyanates are nearly insoluble.

Imines are most generally prepared either by reaction of a ketone or aldehyde with a primary amine or by the addition of Grignard reagents to nitriles. Oxidation (or catalytic dehydrogenation) of amines can also be made to give imines.

Preparation of isocyanates

Isocyanates are nearly always prepared by the reaction of primary amines with phosgene or by a process called, after its discoverer, the Curtius rearrangement of acyl azides, although other methods exist, such as the addition of cyanic acid to olefins, the reaction of potassium cyanate with alkylating agents, such as methyl sulfate, and the cleavage of certain urethans by heat or salts of silver or mercury. Isothiocyanates, R—N=C=S, can be obtained naturally in some cases; allyl isothiocyanate, for example, comes from mustard seeds, and an old term for isothiocyanates as a class is mustard oils. Laboratory preparation starts with primary amines, which react with carbon disulfide to form dithiocarbamate salts. These can be decomposed to isothiocyanates, sometimes by heating but more generally by treating with salts of silver or of lead. An important source of thiocyanates is the reaction of alkyl thiosulfates with alkali metal cyanides.

Sources of nitriles

The commonest source of nitriles is the reaction of alkyl halides with sodium (or potassium) cyanide, which goes well with primary alkyl halides but fails when they are tertiary. When silver cyanide is used, a complex is formed, which liberates an isocyanide when broken up with excess potassium cyanide. Aryl halides, which are generally inert toward alkali metal cyanides, can be converted to cyanobenzenes by heating with cuprous cyanide. Both nitriles and isocyanides can be made by dehydrating amides.

An industrially important source of nitriles is the catalytic reaction of hydrocarbons bearing methyl groups with ammonia and oxygen at high temperatures (ammoxidation).

Carbodiimides are prepared from thioureas by treating with reagents that extract hydrogen sulfide (*e.g.*, mercuric oxide), as well as by other methods.

Principal reactions. The double and the triple C—N bonds are susceptible to hydrolysis (reaction with water), generally with catalysis by either acid or base (a catalyst alters the speed of reaction enormously, without being itself used up, generally at a much lower temperature than needed to initiate the reaction without a catalyst), with varying degrees of ease. The ultimate result is replacement of all C—N bonds by C—O bonds, such that imines yield aldehydes or ketones, nitriles yield carboxylic acids, isocyanides yield primary amines and formic acid, and cyanates, isocyanates, and their sulfur analogues yield carbon dioxide (or carbon oxysulfide). When conditions are mild enough or the amount of water is limited, hydrolysis may be stopped at the intermediate stage of amides (from nitriles), formamides (from isocyanides,) or ureas (from isocyanates and carbodiimides). Hydrolysis is paralleled by reactions with alcohols, of which the two most important examples are the conversion of nitriles to imidate esters and to ortho esters and of isocyanates to urethans. Ammonia and amines react analogously, particularly with isocyanates or isothiocyanates, and are thereby converted to ureas or thioureas, respectively.

Reduction (a category of reactions, the reverse of oxidation and here usually meaning reaction with hydrogen in the presence of a catalyst) in general saturates the double and triple C—N bonds with hydrogen, converting the compounds to amines of one kind or another.

A class of compounds called Grignard reagents reacts readily with nitriles and isocyanates, converting them to imines and amides, respectively. The same conversion can be accomplished by reaction with aromatic hydrocarbons and a class of catalysts called Friedel–Crafts catalysts; *e.g.*, aluminum chloride.

COMPOUNDS WITH N—N BONDS

Physical properties. The smaller hydrazines are hygroscopic liquids with unpleasant odours and are nearly as strongly basic as amines when they bear only one or two alkyl substituents. Successive substitution lowers the basicity, and the tetraalkylhydrazines, R_2N—NR_2, are only feebly basic. Substitution of methyl groups for hydrogen in hydrazine lowers the boiling point, by reducing hydrogen bonding, but larger substituents add so much weight that their net effect is to raise the boiling point.

Substitution of an acyl group for hydrogen, giving a hydrazide, raises both the melting and boiling point, and such compounds are nearly all solids, basic enough to dissolve in dilute mineral acid but weaker than hydrazine. Hydrazones are intermediate in properties and generally are liquids or low-melting solids (unless there are large substituents on the nitrogen atom) and of feeble basicity. Azines, $R_2C{=}N{-}N{=}CR_2$, may be liquids or solids and are generally yellowish and nonbasic.

The derivatives of azobenzene, a typical azo compound, in general are solids, always coloured, insoluble in water, and nonbasic; the larger aliphatic azo compounds are liquids. Monosubstituted analogues, usually named as diazenes, are of low stability, and much less is known about them.

Azoxy compounds resemble the azo compounds but are of paler colour; they are distinct from the oxadiaziridines, with which they were at one time confused.

Diazonium salts are generally colourless, crystalline, water-soluble substances that decompose slowly at room temperature and are capable of exploding. The salts are neutral in reaction, which implies that the diazonium hydroxides are strong bases. In basic solution, however, diazonium salts become converted to diazotate ions, Ar—N=N—O$^-$, which can exist in stereoisomeric forms

related to those of the imines illustrated previously. Closely related are the diazo compounds, of which diazomethane, $CH_2{=}N_2$, is the best known; they are coloured yellow to purple, generally low-melting, and seldom very stable; they are inert to bases but are destroyed by acids.

The azides are colourless compounds, commonly liquid and insoluble in water. They are neutral but may be destroyed by concentrated acids, and though they are more stable than diazo compounds they may be dangerously explosive, especially if they are small.

The most important of the remaining compounds with N—N bonds are the triazenes, $R_2N{-}N{=}N{-}R$, but all are rarely encountered.

Sources and preparation. Alkyl hydrazines are produced when hydrazine is treated with alkylating agents, and, although mixtures with dialkyl and even trialkyl hydrazines may result, this is usually a satisfactory preparative method. Some alkyl hydrazines are more easily prepared from hydrazides by reduction with lithium aluminum hydride. Reduction of nitrosamines is a good route to *N,N*-disubstituted hydrazines, and reduction of hydrazones has been used to prepare trisubstituted hydrazines. Aryl hydrazines can be obtained in wide variety by reducing diazonium salts; although this is the preferred method, a good alternative is the reaction of azodicarboxylic ester with aromatic hydrocarbons. This reaction produces hydrazine analogues of urethans, which can be hydrolyzed cleanly to aryl hydrazines.

Hydrazones are usually made by the reaction of ketones or aldehydes with hydrazines, although they can also be made by tautomerization of azo compounds. Hydrazides are made by reaction of esters with hydrazine; if acid chlorides are used, diacyl hydrazines are produced.

Aromatic azo compounds can be prepared in great variety by the reaction of diazonium salts with compounds containing reactive benzene rings; this is the method used to prepare azo dyes. Alternatively, an aromatic primary amine and a nitroso compound can be converted to an azo compound by the elimination of water. When the desired azo compound is symmetrical, such as azobenzene, the same effect can be obtained by reducing the corresponding nitrobenzene under suitable conditions. Aliphatic azo compounds are nearly always prepared by oxidizing a hydrazine.

Azoxy compounds are formed by the oxidation of azo compounds with peroxides or by reaction of nitroso compounds with hydroxylamines, with elimination of water. Reduction of a nitrobenzene may be controlled so as to produce the latter two reagents, thereby producing an azoxybenzene.

Diazonium compounds are always prepared by diazotization, the reaction of a primary aromatic amine with cold nitrous acid (although there are some unimportant reactions that also give rise to them).

The most general method among many for preparing diazoalkanes is the treatment of an *N*-nitroso-*N*-alkylamide with a strong base. Oxidation of simple hydrazones is a useful preparative method for larger diazoalkanes.

Aliphatic azides are usually prepared by a simple displacement reaction between an alkyl halide and sodium azide, but aryl azides, which cannot in general be prepared in this way, are easily prepared from diazonium salts and sodium azide. Both aryl and acyl azides can be prepared by the reaction of nitrous acid with the corresponding hydrazine derivative.

Reactions. The most characteristic reactions of alkyl and aryl hydrazines are salt formation and alkylation. Alkyl halides, sulfates, etc., will replace the hydrogens stepwise, usually continuing at the same nitrogen until a quaternary compound is formed. Hydrazones and hydrazides undergo the same type of reaction, although not as readily.

Most organic hydrazine derivatives are reducing agents in some degree, and those that bear no more than one substituent on each nitrogen are especially easily oxidized, forming azo compounds. With two hydrogens on the same nitrogen, oxidation is a little more difficult but can be accomplished with strong oxidizing agents so as to convert disubstituted hydrazines to diazenium salts and hydrazones to diazoalkanes. The N—N bond in hydrazines can be cleaved by reduction to give two molecules of amine; the preferred reagent is hydrogen in the presence of a nickel catalyst.

Nitrous acid attacks most hydrazine derivatives; the initial product in all cases is probably a nitroso (*i.e.*, the molecule contains the —NO group) derivative, but only when the hydrazine is trisubstituted can it usually be isolated. Monoarylhydrazines and simple hydrazides are converted all the way to azides, in an important preparative reaction.

Azo compounds are relatively unreactive. They are inert to alkylating agents, to aldehydes and ketones, and to acylating agents, and are not easily attacked by strong acids or bases. Reduction, their most characteristic reaction, takes place in two stages, first to a hydrazine and then to a pair of amines. Azo compounds having a hydrogen on an adjacent carbon isomerize (*i.e.*, rearrange their molecular structure) rather easily to hydrazones, especially in the presence of a base.

Reaction with acids

The most important general reaction of diazoalkanes is that with acids, which are converted to their esters, usually quantitatively and rapidly. Another important general reaction is the loss of nitrogen to form transient carbenes (electron-deficient species with divalent carbon) when heated or exposed to light. Diazomethane and to a lesser extent other diazoalkanes react with acid chlorides to form diazoalkyl ketones. These undergo the Wolff rearrangement when heated (usually in the presence of a silver catalyst) and form ketenes (or, in the presence of alcohols, water, or ammonia, the corresponding esters, acids, or amides). The overall process thus converts a carboxylic acid through its acid chloride to an acid with one more carbon atom (Arndt–Eistert synthesis).

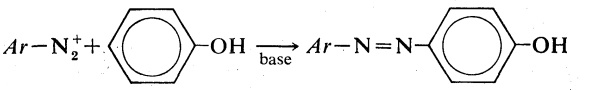
The most important reactions of azides are decomposition by heat, light, or concentrated acids, reduction, and addition of certain very strong bases. Heat or light causes loss of two-thirds of the nitrogen as N_2. In some cases this is accompanied by rearrangement, as in the Curtius rearrangement of acyl azides to isocyanates. In others, a reactive, electron-deficient intermediate, called a nitrene, is produced, which cannot be isolated but reacts very rapidly with itself or the solvent or both.

Many reducing agents, such as hydrogen (with a platinum catalyst) or sodium arsenite, convert azides cleanly to primary amines, for which this process is sometimes a useful preparative method.

The reactions of diazonium salts have been studied intensively since the mid-19th century, and the possibilities are so many and complex that only a small selection can be mentioned here. One important family of reactions involves loss of nitrogen and replacement of the diazonium grouping by another atom or group of atoms. Warming in water solution brings about replacement by OH, giving a phenol. In the presence of inorganic iodides or azides, aryl iodides or azides form rapidly. Replacement by chloride, bromide, or cyanide is brought about with the help of a copper(I) salt as catalyst (Sandmeyer reaction). Replacement by hydrogen may be brought about by alcohol or hypophosphorous acid.

Formation of azo compounds

Another important family of reactions consists of the attachment of another group, usually through a carbon atom, at the end of the diazonium system, to give various types of azo compounds. The most important examples involve reaction with phenols or anilines, forming substituted azobenzenes. When these substances also bear functions that help in binding them to textile fibres, they are azo dyes.



COMPOUNDS WITH N—O BONDS

Classes, sources, and preparation. Hydroxylamines are liquids or low-melting solids, volatile if small, and weakly basic. Oximes are still weaker bases and feeble acids and hydroxamic acids are weak acids similar to phenols. Nearly all hydroxylamine derivatives are soluble

in water if they do not contain more than six carbon atoms.

Compounds with a nitroso group attached to carbon are bright blue or green, low-melting, neutral substances, but few of them are well known, owing to their capacity for rapid dimerization (usually reversible). The dimers are colourless solids having structures analogous to azoxy compounds. With many nitrosobenzenes, the colourless dimer dissociates to the green monomer simply on melting or dissolving in an inert solvent.

Most simple nitro compounds are liquids and are only slightly soluble in water, if at all. Many nitrobenzene derivatives are yellow. Nitro compounds are in general neutral, but those that bear a hydrogen on the same carbon as the nitro group are pseudoacids; that is, although only feebly acidic themselves, they react slowly with bases to form salts derived from tautomeric forms (nitronic acids, or *aci*-nitro compounds), which have acid strengths similar to those of carboxylic acids. Trinitromethane, however, exists entirely in the nitronic acid form and is a strong acid. Nitro compounds are polar, and the smaller nitroalkanes dissolve many salts.

The esters of nitrous and nitric acids are liquids, volatile if small, and insoluble in water. The nitrites are not very stable to storage and are also easily hydrolyzed; the nitrates are considerably more stable, if they are pure, and are difficult to hydrolyze. Nitrate esters are sensitive to heat and shock, especially when impure, and can be violently explosive.

Stable organic derivatives, known as nitrosamines, of the unstable amide of nitrous acid are well known. There is no practical natural source of compounds containing N—O bonds, although some of them occur in nature.

The preparation of organic derivatives of hydroxylamine relies heavily on hydroxylamine itself, which reacts with aldehydes and ketones to form oximes and with esters, anhydrides, or acid chlorides to form hydroxamic acids.

Nitroso compounds are most generally made by oxidation of *N*-substituted hydroxylamines, although many nitrosobenzenes can be made conveniently by oxidizing the corresponding aniline with peroxysulfuric acid, and a few can be made by direct nitrosation (addition of nitroso group, NO) of the benzene ring by nitrous acid (*e.g.*, *p*-nitrosodimethylaniline).

Nitro compounds are prepared in great variety by treating aromatic hydrocarbons with nitric and sulfuric acids; with some severe limitations, nitration of aliphatic hydrocarbons (generally at high temperature in the vapour phase) is also of preparative value, but most aliphatic nitro compounds are made by further reactions of the few commercially available nitroalkanes, such as nitromethane.

Esters of nitrous acid are easily made by reaction between dilute nitrous acid and alcohols, but a mixture of nitric and sulfuric acids is generally required to convert alcohols to alkyl nitrates. Nitrosamines are formed easily when secondary amines are brought into contact with nitrous acid, but nitramines require special methods, such as treating primary amines with α-cyanoisopropyl nitrate or converting an amine to a carbamyl chloride and then treating it with silver nitrate.

Reactions. Although most hydroxylamines are indefinitely stable as their salts, the free bases do not generally store well; *N*-substituted hydroxylamines slowly transform into amines and azoxy compounds, but *O*-alkyl hydroxylamines are more stable. *N*-aryl hydroxylamines differ from the alkyl analogues in that acid easily brings about a rearrangement to form *p*-aminophenols and other ring-substituted anilines.

Both alkylation with alkyl halides and acylation with esters, anhydrides, and acid chlorides take place easily. Aldehydes and ketones react with *O*-substituted hydroxylamines to form oxime derivatives, but *N*-substituted hydroxylamines generally react only with aldehydes, with which they form nitrones.

Hydroxylamines and oximes can be reduced by various reagents, most generally by catalytic hydrogenation, and then form amines.

Hydroxamic acids are hydrolyzed to the parent carboxylic acids by hot aqueous mineral acid. In the form of their alkali metal salts, they can be alkylated or acylated.

The chemistry of nitroso compounds is dominated by their oxidation and reduction. They are active oxidizing reagents, and a wide variety of reducing agents will convert them first to hydroxylamines and then to amines. Because nitroso compounds react with both of these functional groups, the actual products obtained from a reduction may be further transformation products—azo or azoxy compounds, hydrazines, benzidines, etc. Oxidation of nitroso to nitro compounds requires a strong oxidizing agent, such as permanganate or a peroxy acid.

Aliphatic nitroso compounds are generally so tightly bound up as the dimers that reactions specifically of the monomer are not easily observed. If there is a hydrogen on the same carbon as the nitroso group, the monomer easily and irreversibly isomerizes to an oxime. The nitroso group is an effective trap for free radicals, which add to the nitrogen to form a nitroxide, $RR'N{-}O\cdot$. Nitroso compounds react with many other types of compounds under suitable conditions: olefins, aldehydes, Grignard reagents, compounds having active methylene groups, etc.

Reduction of nitro compounds

By far the most important reaction of nitro compounds is their reduction, which, although not so easily accomplished as with nitroso compounds, can still be brought about by a large variety of reducing agents; hydrogen (with catalysts), dissolving metals (iron, tin, zinc, etc.), sulfides, sodium alkoxides, etc. The products may be any of the oxidation states of nitrogen but generally not nitroso compounds, for these are invariably reduced further. If the reducing medium is acidic, reduction almost always goes all the way to an amine. In neutral medium, it is in many cases possible to stop at the hydroxylamine stage. In basic solution, N—N bonds are usually formed, and the products are hydrazo, azo, or azoxy compounds.

The more important part of the chemistry of aliphatic nitro compounds has to do with the reactivity of the hydrogens on the same carbon as the nitro group. Concentrated sulfuric acid isomerizes primary nitroalkanes to hydroxamic acids, which are usually hydrolyzed to carboxylic acids and hydroxylammonium sulfate. Their salts, the nitronates, are not very stable and undergo some unusual reactions. Strong mineral acid cleaves them to aldehydes or ketones and nitrous oxide (Nef reaction). Strong base causes more complex reactions; primary nitroalkanes condense to heterocyclic compounds (isoxazoles) and nitromethane itself is converted in stages to a salt of nitroacetic acid. Heavy-metal salts have distinctive reactions, such as that of the mercuric salt of nitromethane, which loses water to form mercuric fulminate.

Most nitroalkanes decompose above 300° C and form olefins, water, and oxides of nitrogen. With aldehydes (and some ketones), primary nitroalkanes react in a way analogous to the base-catalyzed aldol condensation, adding to the carbonyl group to form a nitro alcohol or its dehydration product, a nitro olefin (Henry reaction).

Alkyl nitrites decompose rather easily under the influence of heat or light to produce mixtures of carbonyl and *C*-nitroso compounds or their further transformation products. The first step is apparently loss of NO to form an alkoxy free radical (*i.e.*, a group of atoms with one unpaired electron), $R{-}O\cdot$. Alkyl nitrites also react as nitrosating agents, similar to nitrous acid.

Alkyl nitrates begin to decompose near 150° C. If the heating is carefully controlled, alkyl nitrites may be formed, along with smaller fragments and oxides of carbon, but otherwise explosion ensues, and the entire molecule may be converted to oxides of carbon, water, and nitrogen. Nitrates are most conveniently converted to the parent alcohol by reduction (zinc and acetic acid, ferrous chloride, etc.), which at the same time converts the nitrogen to nitric oxide or ammonia. Alkyl nitrates can function as alkylating agents or nitrating agents.

BIBLIOGRAPHY. N.V. SIDGWICK, *The Organic Chemistry of Nitrogen*, 3rd ed. rev. by I.T. MILLER and H.D. SPRINGALL (1966), a comprehensive account of the subject for readers who have at least the equivalent of college-level organic chemistry; P.A.S. SMITH, *The Chemistry of Open-Chain Organic*

Nitrogen Compounds, 2 vol. (1965–66), more detailed but narrower coverage of the subject and with more references; C.A. STREULI and P.R. AVERELL, *The Analytical Chemistry of Nitrogen and Compounds*, 2 vol. (1970), comprehensive treatment of this subject; *Kirk-Othmer Encyclopedia of Chemical Technology*, 2nd ed., 22 vol. (1963–70), presents industrial and technological aspects of nitrogen compounds under many individual headings throughout the various volumes; S. PATAI, *The Chemistry of the Amino Group* (1968); J.Z. ZABICKY (ed.), *The Chemistry of Amides* (1970); Z. RAPPOPORT, *The Chemistry of the Cyano Group* (1971); and H. FEUER (ed.), *The Chemistry of the Nitro and Nitroso Groups*, 2 vol. (1969–70), highly detailed presentations at an advanced level.

(P.A.S.S.)

Organic Phosphorus Compounds

Organic phosphorus compounds are carbon-containing compounds that also contain one or more atoms of the element phosphorus. Of the several million known organic compounds, several hundred thousand contain phosphorus.

The organic compounds of phosphorus are typically colourless liquids or solids, not physically distinct from other organic compounds of similar molecular size and character. Several types are very toxic, including the phosphines, but they present little danger in normal use because their exceedingly offensive odours generally prevent overexposure. Many toxic phosphorus compounds of other classes, however, are particularly hazardous because they are colourless, almost odourless liquids. Some of them are used as insecticides. Others were developed during World War II as a particularly unpleasant class of chemical warfare agent, the nerve gases.

Organic phosphates

Insecticides and nerve gases belong to the general family of organic phosphates, which are probably the most important of the organic phosphorus compounds. Also included in this category are many biochemical compounds essential to life processes, such as the nucleic acids (control factors in heredity) and nucleotide coenzymes (compounds that permit enzymes to function properly). Other organic phosphates find industrial uses—as solvents and flame-retardation agents, for example. Other classes of organic phosphorus compounds are less useful, but some play important roles in the synthesis of complex chemical compounds of various types, such as pharmaceuticals.

There are many problems connected with the naming of organic phosphorus compounds, largely because the first organic phosphorus compounds were made and named in the early part of the 19th century, before complete knowledge of their structures was obtained. Furthermore, different authors in different countries have frequently used different names for the same compound. This situation was much improved, however, after an international agreement on nomenclature was reached in 1952, but the problem still exists. In this article standard international nomenclature is used.

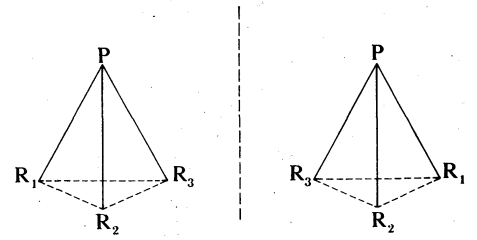
GENERAL PROPERTIES

The chemistry of the organic phosphorus compounds is best understood in terms of the influence of the phosphorus atom on the organic molecule. This influence depends in large measure on the electronic character of that atom.

Electronic configuration. Atoms of phosphorus, like those of other elements, contain a central, positively charged nucleus and, surrounding it, successive concentric rings, or shells, of negatively charged electrons. The electrons of the outermost shell are capable of joining with the electrons of other atoms to form chemical bonds, and they are, therefore, called bonding, or valence, electrons—the term valence referring to the number of other atoms with which a single atom of a given element can combine. In its outermost shell, a phosphorus atom has five electrons in two subshells called—for no important reason—*s* and *p* subshells. Within these subshells the electrons move in particular paths, or orbitals. The five outermost, or valence, electrons of a phosphorus atom occupy one *s* orbital (containing two electrons) and three *p* orbitals. The three *p* electrons are unpaired (each orbital being capable of holding two electrons) and therefore are available to form three single bonds, giving rise to trivalent compounds of phosphorus. Alternatively, however, the two *s* electrons can also become involved in bonding, and this situation leads ultimately to pentavalent phosphorus compounds.

Structure and bonding. Trivalent phosphorus compounds, with the general formula PX_3 (in which P represents a phosphorus atom and X any other atom or group), have pyramidal structures (one atom at each vertex of a three cornered pyramid), that are closely analogous to those of the corresponding and better known nitrogen compounds. One of the main differences between the phosphorus and nitrogen compounds is that the angles between the different P–X bonds are considerably smaller than the corresponding angles between the N–X bonds (in which N is the symbol for nitrogen)—close, in fact, to the 90° expected if bonding involves only the *p* orbitals of phosphorus. Another difference between phosphorus and nitrogen compounds is that the pyramidal configuration in the phosphorus series is much more stable to inversion—that is, to turning inside out, like an umbrella. As a result, trivalent organic phosphorus compounds (phosphines) with three different organic substituents, unlike the corresponding nitrogen compounds (amines), can exist in two separate forms (isomers), which are stable under mild conditions. The two forms are mirror images of one another, as shown by the diagrams below:

Pyramidal structures



In these diagrams, R_1, R_2, and R_3 represent the different organic groups joined to a phosphorus atom. The solid lines represent bonds between the phosphorus atom and the organic groups (and also the edges of the pyramidal structures of the molecules). The dotted lines indicate the bases of the pyramids, and the dashed line is the position of the mirror plane (indicating that each isomer is a mirror image of the other.)

The great majority of phosphorus compounds, however, are not pyramidal but rather have four substituents arranged tetrahedrally about the phosphorus atom. These tetrahedral compounds include salts of the formally trivalent compounds—formed by utilizing the two *s* electrons and based on the phosphonium ion, PH_4^+ (in which H is hydrogen)—as well as the much larger group of truly pentavalent compounds. Pentavalent compounds can be formed because the phosphorus atom has available unused orbitals—called *d* orbitals—which, in the tetrahedral PX_4^+ state, readily accept electrons from other (donor) atoms to form a fifth bond to phosphorus. The donor atom may be in a separate molecule, in which case a pentacovalent species PX_5 results; several scores of compounds with this structure are known. But the utilization of *d* orbitals is most efficient when the donor atom is already bonded to phosphorus. The phosphorus–oxygen bond, for example, is stabilized in compounds such as the phosphine oxides, $R_3P{=}O$ (in which O is an oxygen atom, R represents an organic group, and the double line indicates a double bond), the *p* electrons of oxygen interacting with the vacant *d* orbitals of phosphorus to form a so-called pi bond. This type of multiple bonding stabilizes an adjacent negatively charged centre very effectively in the phosphinimines, $R_3P{=}NR$, and phosphinemethylenes, $R_3P{=}CR_2$ (C representing a carbon atom), as well as being responsible for the great stability of the phosphorus–oxygen double bond (P=O).

Tetrahedral structures

It is also instructive to compare these higher valence compounds of phosphorus to the comparable compounds of nitrogen. Nitrogen atoms, like phosphorus atoms, carry *s* electrons that they are able to use in forming salts—derivatives of the ammonium ion, NH_4^+, just as the related phosphorus compounds are derived from the phos-

phonium ion, PH_4^+. The nitrogen atom, however, does not have vacant *d* orbitals available for sharing donor electrons, so that pi-bonded compounds comparable to those formed by phosphorus do not occur. Thus, the amine oxides, compounds formulated as $R_3N^+–O^-$, cannot form a carbon–nitrogen double bond and are markedly less stable than the comparable phosphorus–oxygen compounds, which do form a phosphorus–oxygen double bond.

This exceptional strength of the phosphorus–oxygen double bond in the phosphoryl compounds has an important influence on much of phosphorus chemistry.

Chemical reactivity. Trivalent phosphorus compounds are readily converted to their oxides and undergo many other reactions that generate products containing the phosphorus–oxygen double bond. Furthermore, the stabilizing effect of pi bonding can be exerted on more than one phosphorus–oxygen bond simultaneously. As a result, anions (negatively charged ions) such as that formulated as $R_2P(=O)(O^-)$ are unusually stable (in having two P–O bonds, both of which can show pi bonding). Indeed, compounds formulated as $R_2P(O)OH$ are strong acids; *i.e.*, they readily give up a hydrogen ion to achieve the stable anionic condition. Even such anions as $RP(=O)(O^-)(OH)$ have acidic properties and in neutral solution are largely dissociated to dianions, RPO_3^{2-}. In these anions, the various oxygen atoms are equivalent and share in the negative charge, while the various phosphorus–oxygen bonds are also equivalent and have equal amounts of partial double bond character.

Importance of the phosphorus–oxygen double bond

Chemically, the phosphorus atom of an organophosphorus compound is a centre of high reactivity. Trivalent phosphorus has a lone pair of nonbonded *s* electrons, which gives it basic (the chemical opposite of acidic) character, as well as nucleophilic character (that is, a tendency to be attracted to positively charged groups). Nucleophilic substitution reactions proceed readily at both trivalent and pentavalent phosphorus centres. The only important exceptions to this generalization are the tertiary phosphine oxides, $R_3P=O$. These have no readily displaced substituent and, therefore, do not undergo substitution reactions of any kind. Furthermore, the doubly bonded phosphorus–oxygen group, P=O, does not undergo the ready addition reactions characteristic of the carbon–oxygen doubly bonded group (C=O, or carbonyl group), the behaviour of which in aldehydes and ketones is an important aspect of general organic chemistry. Thus, the tertiary phosphine oxides, $R_3P=O$, show little or none of the chemical reactivity associated with the ketones, $R_2C=O$.

Analysis of phosphorus in compounds

Phosphorus in organic compounds is usually assayed (measured) as inorganic phosphate. Some compounds, especially those with phosphorus–carbon bonds, are degraded to inorganic phosphate only by powerful oxidizing agents; *e.g.*, concentrated perchloric acid at the boiling point. The inorganic phosphate formed is converted to a substance called phosphomolybdate, which can be determined either gravimetrically (by weight) or colorimetrically (by measurement of colour intensity).

MAJOR CLASSES

Organic derivatives of phosphorus acids. The organic derivatives of the acids based on phosphorus are derived from the six parent acids shown in Table 1. Conversion to organic derivatives can be considered formally to consist of replacement of any or all of the hydrogen atoms by organic groups (R). In addition, one or more of the oxygen atoms may be replaced by sulfur atoms, giving rise to the large family of thioacids of phosphorus.

The free acids themselves exist overwhelmingly in the pentavalent (P=O) forms (shown in the second column of the table), but esters and other derivatives of the trivalent are stable. If this generalization is not appreciated, the nomenclature can be very confusing. Trimethyl phosphite $(CH_3O)_3P$, for example, is a stable liquid that can readily be hydrolyzed (converted by water) to the dimethyl ester. This ester is commonly known, not unreasonably, as dimethyl phosphite, but it exists almost exclusively in the tetrahedral form, $(CH_3O)_2P(=O)H$, and is, therefore, correctly termed dimethyl phosphonate.

Table 1: Trivalent and Pentavalent Forms of the Phosphorus Acids

trivalent form		pentavalent form	
name	formula	name	formula
Phosphinous acid	H_2POH	phosphine oxide	$H_3P=O$
Phosphonous acid	$HP(OH)_2$	phosphinic acid	$H_2P(=O)(OH)$
Phosphorous acid	$(HO)_3P$	phosphonic acid	$(HO)_2P(=O)H$
		phosphoric acid	$(HO)_3P=O$

Most organic compounds of phosphorus formally are derivatives of one of the phosphorus acids—the exceptions being those compounds with three or more phosphorus–carbon or phosphorus–hydrogen bonds. The neutral organic compounds of phosphorus are described below, in sections devoted to the trivalent and pentavalent derivatives, respectively. This section deals specifically with the acid forms, including those acid esters in which the parent acid has more than one acidic group. (Most important by far of these are the esters of phosphoric acid.)

Organic phosphorus acids

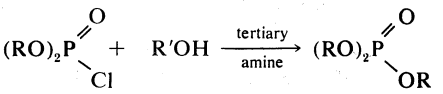
Organic acids of phosphorus are moderately strong acids, with dissociation constants (the standard way of comparing acids) in the region of 10^{-1}–10^{-2}. Acids with two hydroxyl (–OH) groups have a second dissociation constant, usually between 10^{-7} and 10^{-8}. The lower molecular weight compounds are liquids that are soluble in water. The higher molecular weight acids and those containing aromatic groups (a particularly stable organic grouping) become progressively less soluble in cold water. Even these, however, generally dissolve in neutral or alkaline aqueous solution to form anions. As usual, solubility in organic solvents follows almost the reverse trend, being high for nonionized acids and for esters with high molecular weights. In Table 2 is collected a representative selection of acids and their derivatives based on simple organic groups; it illustrates the system of nomenclature now in use, as well as typical physical properties.

Phosphate esters—the esters of phosphoric acid—like other tetrahedral compounds of pentavalent phosphorus, are usually prepared by nucleophilic substitution of

Table 2: Organic Phosphorus Acids and Derivatives

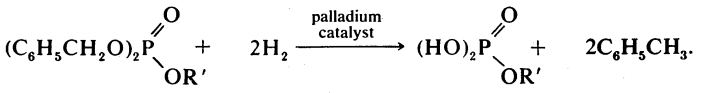
structure	name	melting point (°C)	boiling point (°C)
$(C_2H_5)_2P(=O)OH$	diethylphosphinic acid	19	310
$C_6H_5P(=O)(OH)OC_6H_5$	phenyl hydrogen phenylphosphonate	70–72	...
$(C_2H_5O)_3P=O$	triethyl phosphate	...	215
$C_2H_5OPOCl_2$	ethyl phosphorodichloridate	...	167
$C_6H_5OPO(OH)_2$	phenyl dihydrogen phosphate	97–98	...
$C_2H_5OPO_3^{2-}\ Mg^{2+}$	magnesium ethyl phosphate	...	...
$H_3\overset{+}{N}PO_3^{2-}\ Na^+$	sodium phosphoramidate	...	...
$(C_2H_5O)_2P(O)OPO(OC_2H_5)_2$	tetraethyl pyrophosphate	...	137/1mm
$C_2H_5OP(=O)(O^-)–O–P(=O)(O^-)–O–PO_3^{2-}\ 4\ Na^+$	tetrasodium ethyl tripolyphosphate	...	...

groups located on the phosphorus atom (that is, by attack of a negative group at the positive centre). Syntheses of simple derivatives are based on the readily available substance phosphorus oxychloride, $POCl_3$, which reacts consecutively to lose one, two, or three chloride ions. Triesters are formed when a dialkyl phosphorochloridate (prepared from two molecules of the alcohol to one of phosphorus oxychloride) reacts with an alcohol or phenol in the presence of a tertiary amine, as in the following chemical equation:



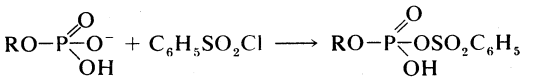
dialkyl phosphorochloridate — alcohol — trialkyl phosphate

This method provides an unambiguous route to unsymmetrical triesters and is readily adapted to the synthesis of monoesters if the first two alkyl groups (R in the above example) can be removed under conditions that leave the third group. The benzyl group ($C_6H_5CH_2-$) is a frequently used blocking group of this kind in the preparation of phosphate esters. It can readily be removed with hydrogen, as in the following example:

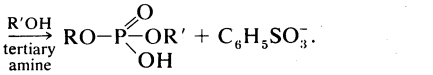


alkyl dibenzyl phosphate — hydrogen — alkyl phosphate — toluene

Important variations of this type of reaction can be carried out with condensing agents, such as acid chlorides or anhydrides, on phosphate anions. These condensing agents convert the negatively charged oxygen atom to a group that is easily displaced. An example of such an easily displaceable group is the benzenesulfonate group (introduced by reaction with benzenesulfonyl chloride). The reactions involved are the following:

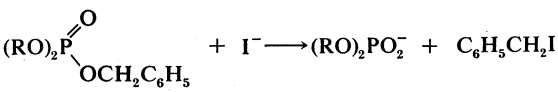


alkyl phosphate anion — benzenesulfonyl chloride



dialkyl phosphate — benzenesulfonate anion

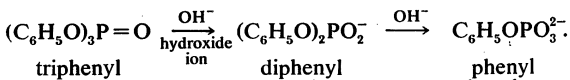
Alternatively, a carbon–oxygen bond may be broken when a powerful but weakly basic nucleophile attacks an ester with a reactive alkyl group, as shown below:



dialkyl benzyl phosphate — iodide ion — dialkyl phosphate anion — benzyl iodide

Beginning in the 1950s the requirement for phosphate esters of known structure in the nucleotide field (see below) greatly stimulated work on methods of phosphorylation.

In their reactions, phosphate esters, especially the triesters, resemble the better known carboxylic esters. Strongly basic nucleophiles, such as hydroxide ions, readily effect nucleophilic displacement of groups on phosphorus atoms in diesters and triesters, as in this example:



triphenyl phosphate — hydroxide ion — diphenyl phosphate anion — phenyl phosphate dianion

Apart from their biochemical importance, organic phosphates play a definite role in the synthesis of organic chemicals containing phosphorus. Organic phosphates and related compounds also find important uses as corrosion and oxidation inhibitors; as plasticizers that confer fire resistance to textiles; as agents for the selective extraction of metals; and as insecticides. Some phosphorylating agents react rapidly and specifically to inhibit certain enzymes, particularly cholinesterase, which is essential for the transmission of nerve impulses. As a result, such compounds are extremely toxic to warm-blooded animals; known as nerve gases, they are particularly hazardous because they can be absorbed through the skin as well as by swallowing or breathing.

Biologically important phosphate esters

Phosphate esters play a vital part in the chemistry of the life process. The genetic code is carried by the deoxyribonucleic acids (DNA), which are high-molecular-weight phosphate esters, built up from four different subunits, called nucleotides. Nucleotides are the phosphate esters of nucleosides (comprised of sugars and organic bases), and in the nucleic-acid structure each nucleotide subunit is joined to the phosphate group of the next by a second ester linkage. The nucleotides themselves also take part in a large number of essential biochemical processes, not only in their monophosphate form but also as diphosphates and triphosphates. Nucleoside triphosphates, especially adenosine triphosphate (ATP), play an important role in the process of phosphate transfer (phosphorylation). This is a key biochemical reaction, because it provides the chemical means by which an organism harnesses the chemical energy available from the degradation of food to do useful work. Phosphorylation is also an important biosynthetic reaction, because a large proportion of the compounds involved in metabolism are present in the organism as their phosphate esters; *e.g.*, the simple-sugar phosphates and phospholipids (compounds structurally similar to fats but containing a phosphate diester linkage).

Several other types of phosphorus compounds are found in nature to a minor extent, including several with phosphorus–nitrogen bonds and a few with phosphorus–carbon bonds (phosphonates).

The acid esters have their own characteristic properties. The monesters, $ROPO(OH)_2$, for example, are stable to alkaline hydrolysis because they are converted to the unreactive dianions, $ROPO_3^{2-}$. At the same time, however, these monoesters are rapidly hydrolyzed under slightly acidic conditions in a unique reaction of the monoanion, which is the predominant ionic form under these conditions. Phosphate diesters are relatively unreactive. They show no special reactivity at any acidity and are hydrolyzed only slowly in alkali—under which conditions, as anions, their reaction with hydroxide ion is inhibited but not completely prevented by electrostatic repulsion; that is, repulsion between electrical charges of the same sign.

Certain structural characteristics confer high reactivity on phosphate esters. Diesters and triesters with their phosphorus atoms contained in a five-membered ring structure, for example, are hydrolyzed some 10^7 times as rapidly as the corresponding open-chain compounds. Also activated toward hydrolysis are many esters with neighbouring ionizable groups close enough to interact with the phosphorus centre. The best known example of this situation occurs in the ribonucleic acids, one of the two principal classes of nucleic acids. Ribonucleic acids are rapidly degraded by alkali under conditions in which the deoxyribonucleic acids, the other principal class, are quite stable. The only structural difference between the two classes of compounds is the presence of an extra hydroxyl group adjacent to the phosphate–ester linkage in the ribonucleic acids.

Trivalent organic phosphorus compounds. Compounds with three substituents bonded to a phosphorus atom have structures resembling those of the corresponding and more familiar nitrogen compounds, but the former group includes a much larger number of stable compounds with bonds to electronegative elements, such as oxygen and the halogens (fluorine, chlorine, bromine

and iodine). Nomenclature of the phosphorus compounds is more complicated than that of the nitrogen derivatives, which are named generally as amines. Naming of phosphorus compounds depends on whether phosphorus is joined solely to carbon or hydrogen atoms or both or whether it is bonded to an electronegative element. Compounds with bonds from phosphorus only to carbon or hydrogen are named as derivatives of phosphine, PH_3, by a system similar to that used for the amines. Thus, methylphosphine, CH_3PH_2, dimethylphosphine, $(CH_3)_2PH$, and trimethylphosphine, $(CH_3)_3P$, are primary, secondary, and tertiary phosphines, respectively. The metal salts are called phosphides, and the protonated forms (compounds to which a hydrogen ion is added) and alkylated forms (compounds to which an organic group is added) are called phosphonium compounds. Compounds with one or more bonds from phosphorus to any element other than carbon or hydrogen are named as derivatives of phosphinous, phosphonous, and phosphorous acids (see Table 1). Compounds with groups formulated as P—OR and P—SR are regarded as esters of the appropriate acid and thio (sulfur-containing) acid, respectively. Amino and halogen compounds are considered to be amides and acyl halides of the respective acids. A representative selection of trivalent phosphorus compounds based on simple organic groups is listed in Table 3. In all, a wide range of compounds is possible, with phosphorus making three bonds to any combination of the following groups: H, R (any organic group), NH_2, NHR, NR_2, OR, SR, SeR (Se being selenium), F, Cl, Br, and I. Compounds belonging to nearly all these categories can be made by nucleophilic substitution reactions at the phosphorus atom, using the readily available phosphorus trihalides, especially phosphorus trichloride, PCl_3. Amines —and alcohols and phenols in the presence of added base—readily displace one or more chloride ions in a phosphorus chloride. Alkyl groups can be introduced by using organic magnesium compounds. A reaction sequence with examples of both reactions is:

Nomenclature of trivalent phosphorus compounds

Table 3: Trivalent Phosphorus Compounds

structure	name	boiling point (°C)
$(C_2H_5)_4P^+\,Br^-$	tetraethylphosphonium bromide	*
$(C_2H_5)_3P^+H\,Cl^-$	triethylphosphonium chloride	
$(C_2H_5)_3P$	triethylphosphine	127
$(C_2H_5)_2P—P(C_2H_5)_2$	tetraethyldiphosphine	
$(C_2H_5)_2PH$	diethylphosphine	85
$C_2H_5PH_2$	ethylphosphine	25
$(C_2H_5)_2P^-\,K^+$	potassium diethylphosphide	
$(C_2H_5)_2PCl$	diethylphosphinous chloride	60–70/15mm
$(C_2H_5)_2PN(C_2H_5)_2$	tetraethylphosphinous amide	181
$C_2H_5P(NH_2)_2$	ethylphosphonous diamide	—
$(C_2H_5)_2POC_2H_5$	ethyl diethylphosphinite	80–85/15mm
$C_2H_5P(OC_2H_5)_2$	diethyl ethylphosphonite	137–9
$(C_2H_5O)_3P$	triethyl phosphite	156–7
$(C_2H_5S)_3P$	triethyl phosphorotrithioite	140–3/18mm

*Decomposes at 300° C.

$$C_2H_5MgCl + PCl_3 \longrightarrow C_2H_5PCl_2 + MgCl_2$$

ethylmagnesium chloride; phosphorus trichloride; ethylphosphonous dichloride; magnesium chloride

$$C_2H_5PCl_2 + 2C_2H_5OH \xrightarrow{\text{base}} C_2H_5P(OC_2H_5)_2 + 2HCl.$$

ethylphosphonous dichloride; ethyl alcohol; diethyl ethylphosphonite; hydrogen chloride

The lower phosphines, with phosphorus–hydrogen bonds, can be converted to metal salts, and these are readily alkylated by alkyl halides, as in the following example:

$$K^+PH_2^- + C_2H_5Br \longrightarrow C_2H_5PH_2 + KBr.$$

potassium phosphide; ethyl bromide; ethylphosphine; potassium bromide

Although the metal salts are easily accessible and useful as synthetic intermediates, the lower phosphines are only very weakly acidic.

The most important class of reactions of trivalent phosphorus compounds is one in which the lone pair of *s* electrons on phosphorus is attacked by electrophilic reagents. The most reactive phosphorus compounds are the phosphines themselves; electronegative substituents on phosphorus and increasing molecular weight both reduce reactivity. Many aliphatic and some aromatic phosphines, for example, are so readily oxidized by atmospheric oxygen that they are spontaneously flammable in air, and these same compounds react explosively with reactive alkylating agents, such as methyl iodide; but triphenyl phosphite, $(C_6H_5O)_3P$, and triphenylphosphine, $(C_6H_5)_3P$, are oxidized and alkylated only relatively slowly and can be handled without special precautions. Similarly, triphenylphosphine is only weakly basic, although aliphatic tertiary phosphines (but not primary or secondary) are bases as strong as the corresponding amines.

Three reactions typical of the phosphines are illustrated by the following equations:

$$(C_2H_5)_3P + C_2H_5I \xrightarrow{\text{cold ether solvent}} (C_2H_5)_4P^+I^-,$$

triethylphosphine; ethyl iodide; tetraethylphosphonium iodide

$$(C_6H_{13})_3P + \tfrac{1}{2}O_2 \xrightarrow{\text{stand in air}} (C_6H_{13})_3P{=}O,$$

tri-*n*-hexylphosphine; oxygen; tri-*n*-hexylphosphine oxide

$$(C_2H_5)_3P + Br_2 \xrightarrow[\text{solvent}]{\text{cold benzene}} (C_2H_5)_3P{-}Br^+\,Br^-.$$

triethylphosphine; bromine; bromotriethylphosphonium bromide

Many trivalent phosphorus compounds remove atoms of oxygen from various molecules, especially when they are attached to nitrogen or sulfur or to another phosphorus centre; for example,

$$(C_2H_5O)_3P + R_3\overset{+}{N}{-}O^- \longrightarrow (C_2H_5O)_3P{=}O + R_3N.$$

triethyl phosphite; tertiary amine oxide; triethyl phosphate; tertiary amine

Many classes of trivalent compounds also form series of complex coordination compounds with metal ions.

As stated above, hydroxyl compounds of trivalent phosphorus exist almost exclusively in the tetrahedral P=O form. Extremely small amounts of the trivalent hydroxyl forms are present in equilibrium, however (often much less than one part per million), and these are thought to be responsible for the observed reactions of the compounds with reactive electrophilic reagents. An example is given below:

diethyl phosphite (tetrahedral form) $(C_2H_5O)_2P(=O)H$

equilibrium

$$(C_2H_5O)_2P{-}OH + Br_2 \longrightarrow (C_2H_5O)_2P^+(OH)(Br)\;Br^-$$

diethyl phosphite (trivalent form); bromine; intermediate

$$\longrightarrow (C_2H_5O)_2P(=O)Br + HBr.$$

diethyl phosphorobromidate; hydrogen bromide

The availability for bonding of the *d* orbitals of phosphorus has a decisive effect on the chemistry of phos-

Phosphonium salts

phonium compounds. Generally, the reactions of phosphonium compounds are more important than those of ammonium compounds, and their chemical behaviour is quite different. Some of these compounds form a fifth bond to phosphorus, as in the following example:

$$(C_6H_5)_4P^+\ Br^- + C_6H_5Li \longrightarrow (C_6H_5)_5P + LiBr.$$

tetraphenylphosphonium bromide — phenyllithium — pentaphenylphosphorane — lithium bromide

Usually, the fifth bond is a pi bond, as in the following:

$$(C_6H_5)_3\overset{+}{P}CH_2C_6H_5\,Br^- + C_6H_5Li$$

benzyltriphenylphosphonium bromide — phenyllithium

$$\longrightarrow (C_6H_5)_3P{=}CHC_6H_5 + LiBr + C_6H_6.$$

triphenylbenzylidenephosphorane — lithium bromide — benzene

Most commonly the multiple bond formed is the very stable P=O bond. Alkoxyphosphonium compounds, for example, are rapidly dealkylated by most nucleophiles to give phosphine oxides:

$$(C_2H_5)_3\overset{+}{P}OC_2H_5\ Br^- \longrightarrow (C_2H_5)_3P{=}O + C_2H_5Br.$$

triethylethoxyphosphonium bromide — triethylphosphine oxide — ethyl bromide

As a result of this reaction, phosphoryl compounds are the final products when any aliphatic esters of trivalent phosphorus acids react with normal alkylating agents, as shown in the following example:

$$(C_2H_5)_2POC_2H_5 + C_2H_5Br \xrightarrow{2\ \text{steps}} (C_2H_5)_3P{=}O + C_2H_5Br.$$

ethyl diethylphosphinite — ethyl bromide — triethylphosphine oxide — ethyl bromide

When the same alkyl group is involved in both ester and alkylating agent, as in the example shown, the alkylating agent is regenerated at the end of the reaction, and only catalytic amounts of it are required. In many instances, the same overall reaction can be brought about simply by strong heating, without the addition of any alkylating agent at all.

Phosphonium compounds are hydrolyzed by strong alkali in an unusual reaction in which an organic group, *R*, is displaced from phosphorus, apparently as the anion R^-. This remarkable result is feasible only because of the great stability of the P=O group of the product and because a mechanism is possible in which two molecules of hydroxide cooperate to break the carbon–phosphorus bond. The steps involved are shown in the equations below:

$$(C_6H_5CH_2)_4P^+ + OH^- \rightleftharpoons (C_6H_5CH_2)_4P{-}OH,$$

tetrabenzylphosphonium ion — hydroxide ion — tetrabenzylhydroxyphosphorane

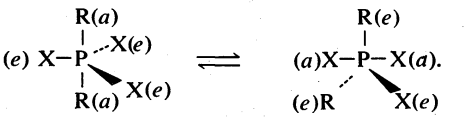
$$(C_6H_5CH_2)_4P{-}OH + OH^- \rightleftharpoons (C_6H_5CH_2)_4P{-}O^- + H_2O,$$

tetrabenzylhydroxyphosphorane — hydroxide ion — anion — water

$$(C_6H_5CH_2)_4P{-}O^- + H_2O \longrightarrow (C_6H_5CH_2)_3P{=}O + OH^- + C_6H_5CH_3.$$

anion — water — tribenzylphosphine oxide — hydroxide ion — toluene

Pentavalent organic phosphorus compounds. The organic compounds of pentavalent phosphorus are of two structural types. A small number have five single bonds to phosphorus and trigonal bipyramidal geometry (that is, a structure formed by two pyramids joined base to base). Examples are pentaphenylphosphorane, $(C_6H_5)_5P$, and pentaethoxyphosphorane, $(C_2H_5O)_5P$. In compounds of this structure, two substituents are in axial (*a*) positions, and three are in equatorial positions (*e*), as shown in the structures below. Compounds with substituents in one position can in some cases be distinguished from those with substituents in the other. Complications arise, however, because axial and equatorial substituents can exchange positions by a process of molecular reorganization known as pseudorotation. An example of pseudoration is shown below:



In this example, the two axial substituents (**R**) become equatorial substituents by changing of the bond positions. In these diagrams, ordinary lines represent bonds in the plane of the paper, dotted lines are bonds extending to the rear, and the wedges indicate bonds extending forward. Compounds with the PX_5 structure are generally very reactive. They readily revert to tetrahedral compounds, usually by way of an initial reversible ionization of the type $PX_5 \rightleftharpoons PX_4^+ + X^-$. An equilibrium of this sort is possible for any phosphonium system, and whether a particular compound exists predominantly as PX_5 or as $PX_4^+X^-$ depends in borderline cases on the environmental conditions (*e.g.*, solvent, temperature, and so on), as well as on the structure of the molecule. Compounds of the PX_5 structure have electrophilic (Lewis-acid) character also, and they may accept electrons from suitable donors to form the rare six-coordinate compounds PX_6^-. One or two organic examples of this class are known.

Compounds with pi bonds

The great majority of pentavalent organophosphorus compounds and, thus, also of all organic phosphorus compounds have four single bonds to phosphorus, one or more of which is reinforced by an additional bond, a pi bond. For purposes of nomenclature these compounds are classified in the same way as the corresponding trivalent derivatives: those with three bonds to carbon or hydrogen have names based on the parent phosphines, and the remainder are named as derivatives of phosphorus acids—in this case phosphinic or phosphonic acids. The range of structures known is essentially that given above for trivalent compounds, with the addition of the fourth bond, which is commonly a phosphorus to oxygen bond, P=O, or phosphorus–sulfur bond, P=S, but may also be a phosphorus–nitrogen or phosphorus–carbon bond, P=NR or $P{=}CR_2$. The selection of compounds listed in Table 4 illustrates the physical properties of these compounds, as well as the ways of naming them.

Tetrahedral compounds of pentavalent phosphorus are generally prepared by reactions involving nucleophilic substitution at phosphorus or oxidation of the appropriate trivalent derivatives. Many reactions of trivalent compounds that give phosphonium compounds as the initial product also lead eventually to phosphoryl derivatives. Several such reactions have been described above. One that is of considerable importance in organic synthesis is the Wittig reaction (named for its discoverer, the German chemist Georg Wittig), in which the oxygen atom of a carbonyl group (carbon and oxygen atoms forming a double bond, C=O) of a ketone or aldehyde is replaced by a doubly bonded carbon atom (methylene group) from a methylenephosphorane, producing an olefin, usually in good yield. The Wittig reagent (the methylenephosphorane) is generated from a phosphonium compound by removing a proton with a strongly basic reagent:

The Wittig reaction

$$(C_6H_5)_3\overset{+}{P}CH_2C_6H_5Cl^- + CH_3(CH_2)_3Li \longrightarrow CH_3(CH_2)_2CH_3 + LiCl$$

triphenylbenzylphosphonium chloride — *n*-butyllithium — butane — lithium chloride

$$+ (C_6H_5)_3P{=}CHC_6H_5\ \{\text{or } (C_6H_5)_3\overset{+}{P}{-}\bar{C}HC_6H_5\}.$$

triphenylbenzylidenephosphorane

Table 4: Pentavalent Phosphorus Compounds

structure	name	melting point (°C)	boiling point (°C)
$(C_6H_5)_5P$	pentaphenylphosphorane	124	—
$(C_2H_5)_3P=O$	triethylphosphine oxide	50	238–40
$(C_2H_5)_3P=S$	triethylphosphine sulfide	94	—
$(C_2H_5)_3P=NC_2H_5$	tetraethylphosphine imine (or imide)	—	—
$(C_2H_5)_3P=C(C_2H_5)_2$	pentaethylmethylenephosphorane	—	—
$(CH_3CH_2CH_2CH_2)_2P(=O)H$	dibutylphosphinous acid (or dibutylphosphine oxide)	58–60	—
$C_8H_{17}P(=O)H_2$	octylphosphine oxide	46–48	—
$(C_2H_5)_2P(=O)Cl$	diethylphosphinic chloride	—	109/16mm
$C_2H_5P(O)Cl_2$	ethylphosphonic dichloride	—	34/3mm
$(C_2H_5)_2P(O)N(C_2H_5)_2$	tetraethylphosphinic amide	—	134/16mm
$(C_2H_5)_2P(O)OC_2H_5$	ethyl diethylphosphinate	—	92/14mm
$(C_2H_5O)_2P(=O)H$	diethyl phosphonate	—	187–88
$C_2H_5PO(OH)_2$	ethylphosphonic acid	61–62	—
$C_2H_5PO(OC_2H_5)_2$	diethyl ethylphosphonate	—	198
$(C_2H_5O)_3P=O$	triethyl phosphate	—	215–16
$(C_2H_5O)_3P=S$	*O,O′,O″*-triethyl phosphorothioate	—	105/20mm
$(C_2H_5S)_3P=S$	triethyl phosphorotetrathioate	—	110/0.2mm

Although the negative charge in the so-called ylide form is largely neutralized by the adjacent positive phosphorus centre (P^+), the unsaturated carbon atom involved remains strongly nucleophilic. In particular, it adds readily to the carbonyl group of aldehydes and ketones, as shown in the example below:

$$(C_6H_5)_3P{=}CHC_6H_5 + (CH_3)_2C{=}O$$

triphenylbenzylidenephosphorane — acetone

$$\longrightarrow (C_6H_5)_3\overset{+}{P}-\underset{\displaystyle (CH_3)_2\overset{|}{C}-O^-}{\underset{|}{CHC_6H_5}}.$$

adduct

The initial adduct rapidly decomposes to the final products, which are the desired olefin and a phosphine oxide, perhaps by way of a cyclic intermediate, as shown below:

$$\underset{\displaystyle (CH_3)_2\overset{|}{C}-O^-}{\overset{H}{\overset{|}{C_6H_5C}}-\overset{+}{P}(C_6H_5)_3} \rightleftharpoons \underset{\displaystyle (CH_3)_2\overset{|}{C}-O}{\overset{H}{\overset{|}{C_6H_5C}}-P(C_6H_5)_3}$$

adduct — intermediate

$$\longrightarrow \overset{H}{C_6H_5C}{=}C(CH_3)_2 + (C_6H_5)_3P{=}O.$$

dimethylphenylethylene (olefin product) — triphenylphosphine oxide

As in earlier examples, the formation of the stable P=O group is an important driving force for this reaction.

The important reactions undergone by the tetrahedral pentavalent organophosphorus compounds have already been described for the phosphate esters and their derivatives. The phosphorus–carbon bond is a strong one, and, with the exception of the Wittig reaction, electronegative substituents on phosphorus are the ones normally involved in chemical reactions.

BIBLIOGRAPHY. R.F. HUDSON, *Structure and Mechanism in Organophosphorus Chemistry* (1965); A.J. KIRBY and S.G. WARREN, *The Organic Chemistry of Phosphorus* (1967), modern mechanistic treatments of the subject, far from comprehensive in their compound coverage; G.M. KOSOLAPOFF, *Organophosphorus Compounds* (1950), deals almost exclusively with compound synthesis, now out of date but the only work of its kind; K. SASSE, *Organische Phosphorverbindungen*, vol. 12/1 and 12/2 of E. MUELLER (gen. ed.), *Methoden der organischen Chemie* (1963–64), a comprehensive and indispensable source book for the preparation and properties of organic phosphorus compounds.

(A.J.K.)

Organic Sulfur Compounds

Organic sulfur compounds constitute a diverse and important subdivision of the class of organic substances. They are widely distributed in nature, often betraying their presence by the strong odours they impart (*e.g.*, to crude petroleum, to certain plants, and to animal secretions). Sulfur-containing amino acids—cysteine, cystine, methionine, and taurine—are important components of biologically important substances, including hormones, enzymes, and coenzymes. Synthetic organic sulfur compounds include numerous insecticides, pharmaceuticals, dyes, solvents, and agents used in flotation processes for refining ores, in improving the performance of lubricating oils, in preparing rubbers, and in making rayon.

Organic compounds make up an enormous class of chemical substances composed of molecules in which atoms of the element carbon are linked to each other and to atoms of other elements, most often hydrogen. Compounds containing only these two elements, the hydrocarbons, undergo characteristic chemical reactions and display distinctive physical properties, but the presence of even one atom of another element in an organic molecule usually has such profound effects upon its chemical and physical properties that organic compounds frequently are classified according to the presence of these additional elements. After carbon and hydrogen, the elements most commonly occurring in organic compounds are oxygen and nitrogen; sulfur is neither extremely common nor extremely rare, but its effects are sufficiently distinctive that organic sulfur compounds comprise a recognized area of specialized study.

The properties of organic sulfur compounds are best understood in the context of knowledge of the composition and transformations of matter, different aspects of which are treated in the articles CHEMICAL COMPOUNDS, ORGANIC; MOLECULAR STRUCTURE; CHEMICAL BONDING; and CHEMICAL REACTIONS.

THE NATURE OF CHEMICAL COMPOUNDS

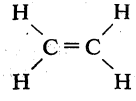
General physical characteristics. *Atoms, nuclei, and electrons.* All matter is made up of one or more of the 100-odd chemical elements, the individual particles of which are called atoms. Each atom is composed of a small nucleus and a surrounding cloud of electrons. In the nucleus, there are particles called protons, each bearing a positive electrical charge, and approximately the same number of uncharged particles, called neutrons. The chemical identity of an atom is determined by the number—the atomic number—of protons in its nucleus; and the chemical behaviour, by the number and arrangement of electrons surrounding it. The electrons are concentrated in well-defined patterns called orbitals, which Orbitals are grouped into sets, or shells, of increasing average distance from the nucleus. Each shell beyond the first is subdivided into subshells, sets of orbitals that differ slightly in energy. The larger the number of the shell, the more numerous are these subshells: the first shell has a single orbital, denoted 1*s*; the second has four, one designated 2*s* and three designated 2*p*; the third has nine, one 3*s*, three 3*p*, and five 3*d*. The approach of other atoms ordinarily affects only the outermost electrons, which therefore determine the kinds of compounds and chemical bonds formed by the atom: that is, its valence.

Particular stability is associated with atoms in which the outermost shell contains eight electrons; an atom that differs from that electron configuration tends to gain, lose, or share electrons with other atoms to attain it. These processes leave the atom with a net electrical charge, and such charged atoms are called ions. Oppositely charged ions attract each other, becoming arranged

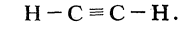
in regular geometric patterns in compounds such as calcium sulfide.

Covalent bonds and molecules. In many molecules, particularly those of organic compounds, atoms are held together by the sharing of electrons between pairs of atoms. Each atom shares one, two, or three of its valence electrons to form a single, double, or triple bond with another atom that contributes an equal number. Such bonds are called covalent. A carbon atom shares four electrons in this way, and a hydrogen atom shares one; simple organic compounds composed of carbon and hydrogen are methane, ethylene, and acetylene. The formulas of covalent compounds, like those of ionic substances, indicate the number of atoms present in the compound: the formula of methane is CH_4, denoting the presence of one atom of carbon (C) and four atoms of hydrogen (H); the formulas of ethylene and acetylene are C_2H_4 and C_2H_2, respectively.

Covalent bonds have specific orientations in space and distances over which they are effective. Often, these directions and distances are indicated in structural formulas, in which each atom is represented by its chemical symbol and each pair of shared electrons by a line:

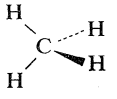


ethylene



acetylene

The structural formula of methane is often drawn as if the molecule were planar, but a more accurate idea of its true shape is conveyed by a figure in which full lines stand for bonds lying in the plane of the paper, a dotted line stands for a bond projecting away from the viewer, and a wedge for a bond projecting toward the viewer:



methane

Ethylene and acetylene, which have multiple covalent bonds, are classed as unsaturated compounds because they contain carbon atoms bonded to fewer than the maximum possible number, four, of other atoms. Methane, on the other hand, is a saturated compound.

Nomenclature

Structural and functional groups. In this article, many organic compounds are mentioned by name, and it may be helpful to the reader to review briefly some of the salient features of their terminology. In most chemical reactions, only one or two of the covalent bonds in a molecule are affected, while the rest of the structure is unchanged. An unchanged group of atoms, regarded as transferred intact from one compound to another, usually is named after the compound that would be formed by attaching a hydrogen atom to it. For example, methyl bromide and hydrosulfide ion react to produce methanethiol and bromide ion, as represented by the equation:

$$\underset{\text{methyl bromide}}{CH_3Br} + \underset{\text{hydrosulfide ion}}{SH^-} \longrightarrow \underset{\text{methanethiol}}{CH_3SH} + \underset{\text{bromide ion}}{Br^-}.$$

Of the four covalent bonds in each molecule of the starting material, only the one linking the carbon and bromine atoms is broken, while the other three are unaffected. The group CH_3, unchanged in the reaction, is conveniently regarded as a single entity and named methyl to show its structural relationship to methane, CH_4. Such groups derived from saturated hydrocarbons, as methyl from methane, are collectively known as alkyl groups. Analogous groups derived from aromatic hydrocarbons, as phenyl from benzene, are called aryl groups.

A group of atoms that undergoes the same general reactions regardless of the structure of the molecule of which it forms a part is called a functional group: typical examples are hydroxyl (—OH), carbonyl ($>C=O$), carboxyl ($-C\overset{O}{\underset{OH}{<}}$), and amino ($-NH_2$).

The sulfur atom. An uncombined sulfur atom has 16 electrons, 10 of which completely fill the 1*s*, 2*s*, and 2*p* orbitals; of the remaining six (*i.e.*, the valence electrons), two occupy the 3*s* orbital and four are present in 3*p* orbitals: The distribution of electrons in a sulfur atom is abbreviated to $1s^22s^22p^63s^23p^4$. An element closely resembling sulfur in electron configuration is oxygen, which has eight electrons in the configuration $1s^22s^22p^4$. The identical distribution, s^2p^4, of the valence electrons in the two elements leads to many similarities in their chemical behaviour. Important differences, however, arise from the fact that the valence electrons of sulfur interact less strongly with the nucleus than do those of oxygen and that the third electron shell of sulfur includes five *d* orbitals. Although these *d* orbitals are not occupied in the uncombined sulfur atom, their availability makes possible the formation of types of compounds not formed by oxygen. In these compounds, the 3*s* and 3*p* orbitals can, in effect, blend together into a new set of four "hybrid" orbitals that can interact with three or four other atoms, or the 3*s*, the 3*p*, and two of the 3*d* orbitals can hydridize to form a set of orbitals that participate in the formation of hexacovalent compounds.

In Table 1, members of several important groups of organic sulfur compounds are compared with the oxygen

Table 1: Organic Sulfur Compounds and Corresponding Organic Oxygen Compounds

sulfur compounds			oxygen compounds		
group of compounds	characteristic structural unit	example	group of compounds	characteristic structural unit	example
Thiols (aliphatic)	—SH	methanethiol	alcohols	—OH	methanol
Thiols (aromatic)	—SH	benzenethiol (thiophenol)	phenols	—OH	phenol
Sulfides	—S—	dimethyl sulfide	ethers	—O—	dimethyl ether
Disulfides	—S—S—	dimethyl disulfide	peroxides	—O—O—	dimethyl peroxide
Thioaldehydes	S ‖ —C—H	ethanethial (thioacetaldehyde)	aldehydes	O ‖ —C—H	ethanal (acetaldehyde)
Thioketones	S ‖ —C—	dimethyl thione (thioacetone)	ketones	O ‖ —C—	2-propanone (acetone)
Thiolocarboxylic acids	O ‖ —C—SH	thioloacetic acid			
Thionocarboxylic acids	S ‖ —C—OH	thionoacetic acid	carboxylic acids	O ‖ —C—OH	acetic acid
Dithiocarboxylic acids	S ‖ —C—SH	dithioacetic acid			

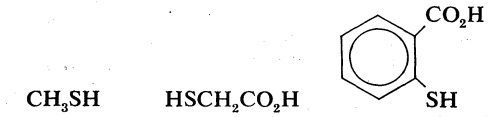
Table 2: Organic Sulfur Compounds with No Oxygen Analogues

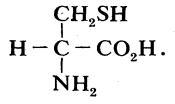
group of compounds	characteristic structural unit	example
Trisulfides	—S—S—S—	dimethyl trisulfide
Polysulfides	(—S—)$_n$; n = 4,5,6, . . .	dimethyl tetrasulfide, dimethyl pentasulfide, etc.
Sulfoxides		dimethyl sulfoxide
Sulfones		dimethyl sulfone
Sulfenic acids		methanesulfenic acid
Sulfinic acids		methanesulfinic acid
Sulfonic acids		methanesulfonic acid

compounds possessing similar structures. Table 2 lists some of the most important types of sulfur compounds that have no counterparts among oxygen compounds.

ORGANIC COMPOUNDS OF BIVALENT SULFUR

Thiols. The organic sulfur compounds most closely resembling their oxygen analogues are the thiols, also called mercaptans. The functional group of the thiols is the mercapto group (—SH), comparable to the hydroxyl group (—OH) present in alcohols and phenols. In preferred names for thiols, the suffix -thiol is appended to the name of the appropriate hydrocarbon. If another group is designated by a suffix, the prefix mercapto- is used. A few thiols are named by using the prefix thio- to denote the replacement of an oxygen atom of a related compound by a sulfur atom, and a few have names that do not convey structural information. Examples are as follows:



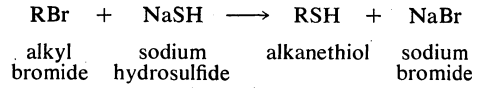


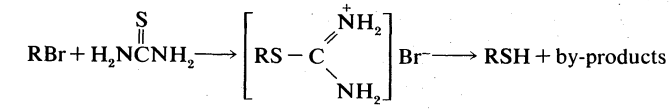
Natural thiols

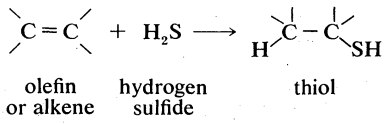
A few thiols are found in nature: crude petroleum contains methanethiol, ethanethiol, and other members of the group. Although they constitute a small fraction of most petroleums, thiols are a serious nuisance. They impart objectionable odours, they are corrosive to the equipment and interfere with the action of catalysts used in refining processes, and their combustion creates the noxious gas sulfur dioxide. Small amounts of thiols are converted into chemical products, but no uses require the amounts that could be recovered from crude oil.

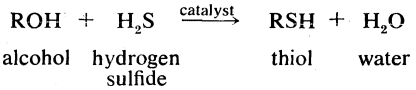
Methanethiol arises from the bacterial decomposition of proteins, such as albumin or gelatin, and butanethiol is present in the defensive secretion of the skunk. A thiol group plays an important role in the natural functions of several proteins and of coenzyme A, which participates in many metabolic reactions.

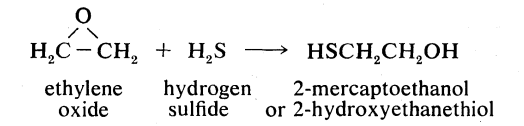
Preparation. Several methods are known for the preparation of thiols. Thiols in which the mercapto group is not attached to an aromatic ring may be made according to the reactions represented by the following equations:

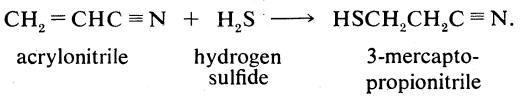




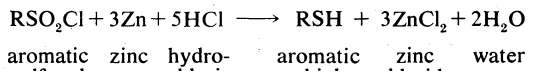


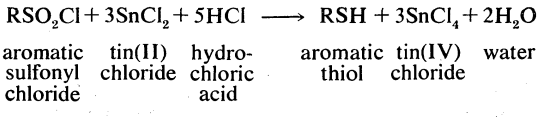


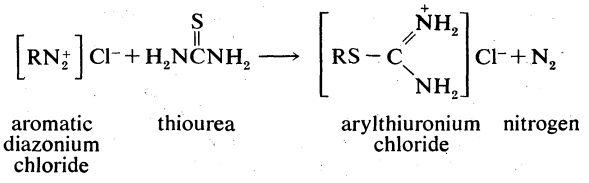


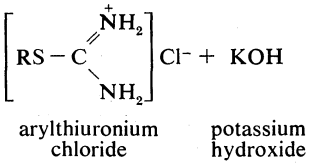


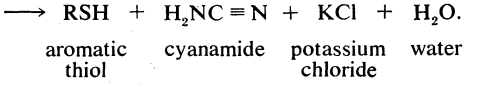
Other methods must be used to make aromatic thiols because of the different nature of reactions typical of aromatic compounds in general (see CHEMICAL REACTIONS). Applicable procedures include those formulated as follows:











Reactions. In most of their reactions, thiols resemble alcohols or phenols, the differences being those of degree rather than kind. Thiols react more completely with alkalies, forming salts, than do the corresponding hydroxyl compounds; that is, the thiols are stronger acids, although very weak in comparison to, say, acetic acid. Aromatic thiols are stronger acids than the aliphatic thiols, just as phenols are stronger acids than alcohols. In the presence of salts of heavy metals (as mercury, lead, zinc, or copper), thiols form mercaptides, soluble in organic solvents (as ether, isopropyl alcohol, benzene, chloroform) but insoluble in water. The characteristic formation of these compounds from mercury salts led to the name mercaptan (Latin *mercurium captans*, "seizing mercury"). The

mercury mercaptide, sodium ethylmercurithiosalicylate (thimerosal, Merthiolate), is a well-known germicide.

Thiols form sulfides (thioethers) and thioesters in reactions similar to those of hydroxyl compounds. They react with aldehydes and ketones to yield thioacetals and thioketals, respectively; sulfur-containing compounds of this type are more readily formed and more stable than the oxygen compounds and are useful in suppressing the reactivity of carbonyl groups while chemical reactions are performed on another part of a molecule. Dithiols are useful for this purpose. Thiols combine with unsaturated compounds, particularly those that have a carbonyl group adjacent to the double bond, forming sulfides.

Thiols differ from hydroxyl compounds in their reactions with oxidizing agents (as oxygen, iodine, hydrogen peroxide). Alcohols usually are converted to aldehydes or ketones, but thiols are transformed into disulfides; further oxidation results in formation of sulfonic acids:

$$2RSH + H_2O_2 \longrightarrow RSSR + 2H_2O$$

thiol; hydrogen peroxide; disulfide; water

$$RSSR + 5H_2O_2 \longrightarrow 2RSO_3H + 4H_2O.$$

disulfide; hydrogen peroxide; sulfonic acid; water

Sulfides and sulfonium salts. Compounds in which two organic groups are bonded to a sulfur atom are called sulfides; the structurally related oxygen compounds are ethers. The organic groups may be both alkyl, both aryl, or one of each. If no other functional group is present in the molecule, sulfides are named as such: dimethyl sulfide is $CH_3—S—CH_3$, methyl phenyl sulfide is $CH_3—S—C_6H_5$, and diphenyl sulfide is $C_6H_5—S—C_6H_5$. If another functional group forms part of the molecule, a sulfide group is designated by the particle -thio-:

thiodiacetic acid; 4-(methylthio)benzoic acid

Properties of sulfides

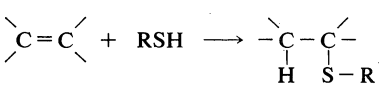
The sulfur atom in sulfides does not cause their physical properties to differ much from those of hydrocarbons of similar molecular size and shape. The sulfides have low solubilities in water, but they are miscible with many organic solvents; they are colourless liquids or solids, many of them possessing odours that are unpleasant, though not as intense as those of the thiols.

Certain sulfides occur in nature: garlic contains diallyl sulfide, and many proteins contain the amino acid methionine, which is a sulfide.

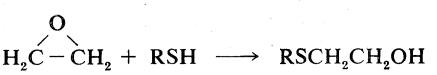
Preparation. Sulfides most often are prepared from thiols; just as preparing a thiol usually involves replacing one of the two hydrogen atoms of hydrogen sulfide, preparation of a sulfide is the replacement of the second, often by similar procedures. Examples are:

$$RSNa + R'Br \longrightarrow RSR' + NaBr$$

thiol sodium salt; alkyl bromide; sulfide; sodium bromide



olefin; thiol; sulfide



ethylene oxide; thiol; alkyl hydroxyethyl sulfide

$$CH_2=CHC\equiv N + RSH \longrightarrow RSCH_2CH_2C\equiv N.$$

acrylonitrile; thiol; 3-alkylthiopropionitrile

The reaction of certain olefins with disulfur dichloride produces chlorine-containing sulfides; the best known example of this process is the manufacture of mustard gas, a blister-forming chemical warfare agent, from ethylene. This reaction is not general, however; the products obtained from other olefins are disulfides or mixtures of monosulfides and disulfides.

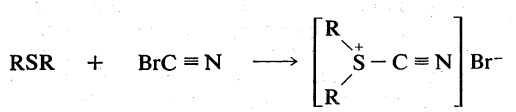
$$2H_2C=CH_2 + S_2Cl_2 \longrightarrow ClCH_2CH_2SCH_2CH_2Cl + S$$

ethylene; disulfur dichloride; mustard gas (bis[2-chloroethyl] sulfide); sulfur

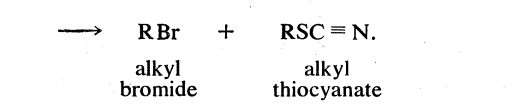
$$2CH_3CH=CH_2 + S_2Cl_2 \longrightarrow ClC_3H_6SSC_3H_6Cl$$

propylene; disulfur dichloride; bis(chloropropyl) disulfide

Reactions. Sulfides combine with chlorine, bromine, iodine, or salts of heavy metals to form crystalline compounds. Cyanogen bromide forms an unstable compound that decomposes by breaking a carbon–sulfur bond:

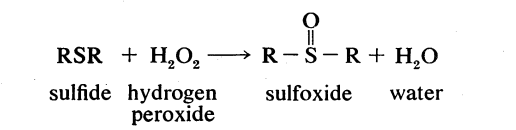


sulfide; cyanogen bromide; unstable sulfonium compound

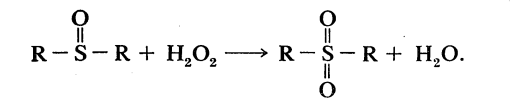


alkyl bromide; alkyl thiocyanate

Oxidizing agents convert sulfides into sulfoxides; if a sufficient amount of the oxidizing agent is present, the sulfoxides undergo oxidation to sulfones:

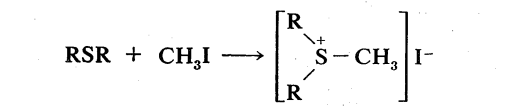


sulfide; hydrogen peroxide; sulfoxide; water



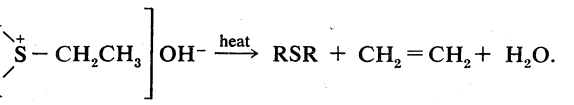
sulfoxide; hydrogen peroxide; sulfone; water

Compounds such as methyl iodide react with sulfides to produce sulfonium salts, in which three organic groups are attached to the positively charged sulfur atom. This process has many parallels in organic chemistry, especially in the reactions of amines and phosphines; it occurs less readily with ethers.



sulfide; methyl iodide; sulfonium iodide

The formation of sulfonium salts from sulfides and alkyl halides can be reversed by heat. The heat-induced decomposition of quaternary ammonium hydroxides (*i.e.*, compounds in which four organic groups are bonded to a positively charged nitrogen atom) is exactly duplicated in the case of sulfonium hydroxides:



sulfonium hydroxide; sulfide; ethylene; water

Disulfides and polysulfides. Disulfides have the structure R—S—S—R′, in which R and R′ represent organic

groups of any kind. They are named by designating the groups attached to the sulfur atoms or by use of the particle -dithio-:

CH_3SSCH_3 — dimethyl disulfide

$CH_3SSCH_2CO_2H$ — methyldithioacetic acid

$HO_2CCH_2SSCH_2CO_2H$ — dithiodiacetic acid

Disulfides. The formation of disulfides by oxidation of thiols or by the reaction of disulfur dichloride and olefins has already been mentioned. A few disulfides have been isolated from natural sources, such as onion, garlic, and asafetida. The amino acid cystine, a disulfide, is an important component of many proteins: the sulfur–sulfur bond is partly responsible for maintaining the molecules in shapes essential for their biological activity. The coenzyme lipoic acid, a disulfide necessary for growth of certain micro-organisms, may also be involved in photosynthesis and other vital processes of plants and animals.

Most important reactions

In the most important reactions of disulfides, the bond between the two sulfur atoms is broken, as by reducing agents, including hydrogen; by strong alkalies; or by alkyl halides:

$$RSSR + H_2 \longrightarrow 2RSH$$

disulfide, hydrogen, a thiol

$$RSSR + KOH \longrightarrow RSH + (RSO)K$$

disulfide, potassium hydroxide, a thiol, alkanesulfenate (salt)

$$RSSR + 4RI \xrightarrow{HgI_2} 2(R_3S^+)I^- + I_2 .$$

disulfide, alkyl iodide, sulfonium iodide, iodine

Chlorine or bromine splits the —S—S— bond of aromatic disulfides but attacks the adjacent carbon atom of an aliphatic disulfide:

$$C_6H_5SSC_6H_5 + Br_2 \longrightarrow 2C_6H_5SBr$$

diphenyl disulfide, bromine, benzenesulfenyl bromide

$$CH_3SSCH_3 + Cl_2 \longrightarrow CH_3SSCH_2Cl + HCl .$$

dimethyl disulfide, chlorine, chloromethyl methyl disulfide, hydrogen chloride

Polysulfides. The tendency to form chains in which atoms of the same element are joined together by covalent bonds, especially pronounced with carbon, is shown to some extent by sulfur. Organic compounds containing such chains of three, four, or more sulfur atoms are called polysulfides. They are less stable than the corresponding carbon compounds, and their properties are not as fully understood. A group of organic polysulfides has received considerable attention as rubbers and as solid propellants for rockets; these are the Thiokols, which are polymeric compounds, having molecules composed of long chains of repeating units built up from simpler compounds. An example is provided by the product made from ethylene dichloride and sodium tetrasulfide:

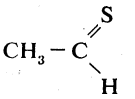
$$nClCH_2CH_2Cl + nNa_2S_4$$

ethylene dichloride, sodium tetrasulfide

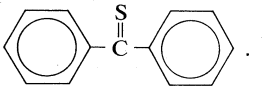
$$\longrightarrow (CH_2CH_2S_4)_n + 2nNaCl .$$

Thiokol A, a polymer; sodium chloride

Thiocarbonyl compounds. The thiocarbonyl functional group ($>C{=}S$), analogous to the carbonyl group, forms part of the structure of thioaldehydes and thioketones. These compounds are named by analogy with the corresponding oxygen compounds:



thioacetaldehyde or ethanethial



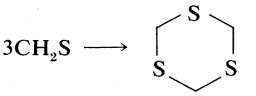
thiobenzophenone or diphenyl thione

The thiocarbonyl compounds have exceptionally repulsive odours. The general method of preparation, from carbonyl compounds and hydrogen sulfide in the presence of acidic catalysts, is not difficult:

$$R-C(=O)-R + H_2S \longrightarrow R-C(=S)-R + H_2O .$$

carbonyl compound, hydrogen sulfide, thiocarbonyl compound, water

The reactions of thiocarbonyl compounds are similar to those of their oxygenated relatives, although they show a greater tendency to undergo cyclotrimerization, in which three molecules unite to form a ring structure:



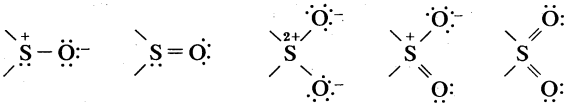
thioformaldehyde, trithiane

Although the monomers are blue, the cyclic trimers (trithianes) are colourless; the latter somewhat resemble the dialkyl sulfides in their reactions.

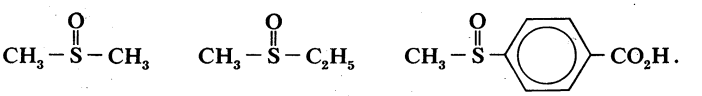
COMPOUNDS CONTAINING POLYVALENT SULFUR

Sulfoxides and sulfones. Two groups of organic sulfur compounds that have no counterparts among organic oxygen compounds are the sulfoxides and the sulfones. In some respects, these compounds resemble carbonyl compounds, containing the group $>C{=}O$, in which the carbon and the oxygen atom each interact with eight electrons, four of which are shared in the double bond. Doubly bonded structures for the sulfinyl ($>SO$) and sulfonyl ($>SO_2$) groups present in sulfoxides and sulfones, respectively, require that the sulfur atom interact with 10 electrons in sulfoxides or with 12 in sulfones. Because these numbers are larger than eight, the number of electrons usually associated with each atom (except hydrogen) in stable covalent compounds, has been the basis of opposition to these formulations. Covalent bond formation by a sulfur atom, however, involves electrons in the third shell, which is larger than the second shell (the one engaged in carbon and oxygen bonding) and has additional orbitals. It is considered that these orbitals (the *d* orbitals) participate in the bonding in sulfoxides and sulfones and that the bonds differ from those of the carbonyl group, showing considerable polarity (separation of electrical charges). Their structures are regarded as intermediate between those with single bonds of high polarity and those with double bonds of low polarity but high electron density on the sulfur atom.

Sulfur–oxygen double bonds

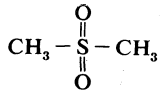


Sulfoxides are named by simply designating the two organic groups attached to the $>SO$ group, with the word sulfoxide, or by forming a prefix from the name of the simpler of the two groups and the particle -sulfinyl-, as shown in the examples:

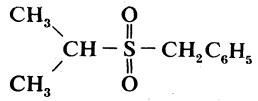


dimethyl sulfoxide; ethyl methyl sulfoxide; 4-(methylsulfinyl) benzoic acid

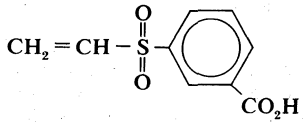
The nomenclature of sulfones is like that of sulfoxides; the particle -sulfonyl- is used in complicated structures:



dimethyl sulfone



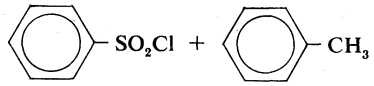
benzyl isopropyl sulfone



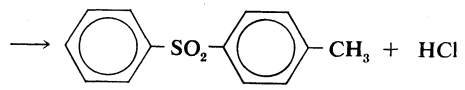
3-(vinylsulfonyl) benzoic acid

Most sulfoxides are colourless liquids or solids with low melting points; they are more soluble in water and possess much higher boiling points than hydrocarbons or carbonyl compounds of similar molecular size. Sulfoxides of low molecular weight have faint odours and tastes described as metallic or garlic-like. In general, sulfones are colourless, crystalline solids at room temperature.

Occurrence and preparation. Several sulfoxides are found in the turnip, garlic, and several species of mustard; another has been isolated from the cockroach. Oxidation of sulfides to sulfoxides and sulfones by hydrogen peroxide has been mentioned; certain aromatic sulfones can be made by the reaction of sulfonyl chlorides (see below *Sulfonic acids*) with aromatic hydrocarbons.

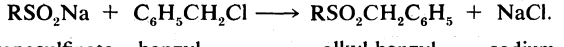


benzenesulfonyl chloride — toluene



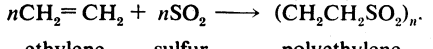
phenyl *p*-tolyl sulfone — hydrogen chloride

The reaction of a metal salt of a sulfinic with a halogen compound is sometimes used to make sulfones:



alkanesulfinate (salt) — benzyl chloride — alkyl benzyl sulfone — sodium chloride

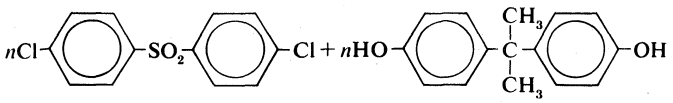
A class of polymeric sulfones results from the reaction of sulfur dioxide with olefins:



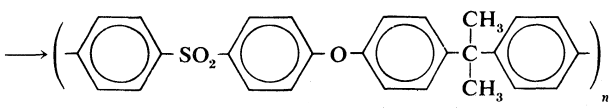
ethylene — sulfur dioxide — polyethylene polysulfone

Polymeric sulfones

These products have not been found useful, although members of a different class of polysulfones have properties that make them valuable as wire coatings. The latter group is made from chlorine-containing aromatic sulfones and a compound (bisphenol A) of two molecules of phenol linked through a hydrocarbon group:



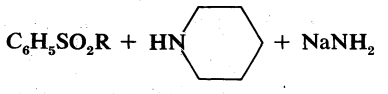
bis(*p*-chlorophenyl) sulfone — bisphenol A



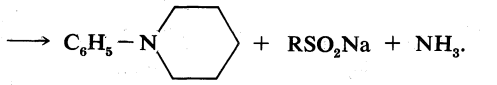
polymer

Reactions. Sulfoxides can be converted to sulfides by powerful reducing agents, such as lithium aluminum hydride, hydriodic acid, or zinc in the presence of sulfuric acid. Sulfoxides are very weak bases, forming salts with strong acids (*e.g.*, hydrochloric acid); they are also very weak acids, giving up a proton only to very strong bases, such as sodium hydride or sodium amide.

Sulfones are generally unreactive compounds and are not attacked by most reducing agents. The bond joining the sulfur atom to an aromatic ring can be cleaved by piperidine in the presence of sodium amide:

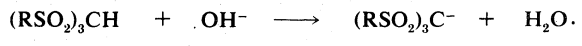


alkyl aryl sulfone — piperidine — sodium amide



N-phenyl-piperidine — sulfinate (salt) — ammonia

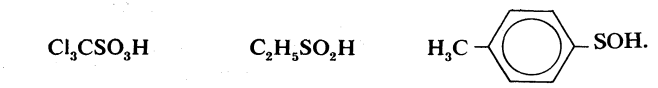
If two or three sulfonyl groups are bonded to the same carbon atom, a hydrogen atom bonded to that atom shows somewhat acidic properties; that is, it can be removed by a base:



trisulfonylmethane — hydroxide ion — trisulfonylmethide ion — water

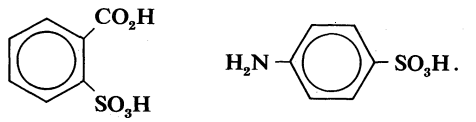
Dimethyl sulfoxide has been investigated as a topical analgesic and as a carrier for drugs (it rapidly penetrates the skin); the disulfones sulfonal, trional, and tetronal have been used in medicine as hypnotics but have been replaced by less toxic compounds.

Sulfonic, sulfinic, and sulfenic acids. Three sulfur-containing functional groups confer acidity upon compounds in which they are present; these compounds are the sulfonic acids, in which the group SO_3H is present; the sulfinic acids, with the group SO_2H; and the sulfenic acids, with the group SOH. All three types of compounds are named by attaching the name of the functional group to the name of the compound in which that group replaces a hydrogen atom:



trichloromethanesulfonic acid — ethanesulfinic acid — *p*-toluenesulfenic acid

Occasionally, sulfonic acids are named using the prefix sulfo- or by arbitrary names:



o-sulfobenzoic acid — sulfanilic acid

The sulfonic acids are very strong—comparable to the mineral acids, hydrochloric, nitric, or sulfuric—and are the most common of the sulfur-containing acids. Most of them are colourless, odourless, crystalline compounds; the characteristic water solubility of these acids and most of their salts has made them useful as detergents that perform well in hard water and as dyes that can be applied to textiles from aqueous solutions.

Sulfinic acids are weaker, less soluble in water, and less stable than sulfonic acids; they are most often prepared in the form of their metal salts, which are more stable.

Sulfenic acids and their salts are unstable compounds, rarely isolated; many substances named as derivatives of sulfenic acids are not actually obtainable from them.

Sulfonic acids. Aromatic sulfonic acids usually are made by the reaction (sulfonation) of an aromatic hydrocarbon with sulfuric acid. Aliphatic hydrocarbons seldom react similarly with sulfuric acid, but aliphatic sulfonic acids may be obtained by oxidation of thiols or other sulfur-containing starting materials or by treatment of certain halogen compounds with sodium sulfite:

$$C_6H_6 + H_2SO_4 \longrightarrow C_6H_5SO_3H + H_2O$$

benzene; sulfuric acid; benzenesulfonic acid; water

$$CH_3SH + 6HNO_3 \longrightarrow CH_3SO_3H + 6NO_2 + 3H_2O$$

methanethiol; nitric acid; methanesulfonic acid; nitrogen dioxide; water

$$ClCH_2CO_2Na + Na_2SO_3 \longrightarrow NaO_3SCH_2CO_2Na + NaCl.$$

sodium chloroacetate; sodium sulfite; disodium sulfoacetate; sodium chloride

The formation of sulfonic acids from sulfuric acid and aromatic compounds can be reversed by reaction with water, although strenuous conditions are sometimes necessary (*e.g.*, use of superheated steam). When aromatic sulfonic acids are heated with caustic alkalies, the sulfo ($-SO_3H$) group is replaced by a hydroxyl ($-OH$) group; this reaction is useful for preparing phenols:

$$C_6H_5SO_3H + NaOH \longrightarrow C_6H_5OH + NaHSO_3.$$

benzenesulfonic acid; sodium hydroxide; phenol; sodium bisulfite

In certain cases, the sulfo group can be replaced by various others, including cyano ($-CN$), nitro ($-NO_2$), formyl ($-CHO$), amino ($-NH_2$), or alkylthio ($-SR$).

Sulfonyl chlorides may be prepared from sulfonic acids by reaction with phosphorus pentachloride, chlorosulfuric acid, or certain other reagents:

$$C_6H_5SO_3H + PCl_5 \longrightarrow C_6H_5SO_2Cl + POCl_3 + HCl.$$

benzenesulfonic acid; phosphorus pentachloride; benzenesulfonyl chloride; phosphoryl chloride; hydrogen chloride

Sulfonyl chlorides are widely employed in preparing sulfonamides, sulfonates (esters of sulfonic acids), and sulfinic acids; their use in making sulfones has already been mentioned. Other sulfonyl halides (fluorides, bromides, or iodides) are less frequently made.

Sulfonamides result from the reaction of sulfonyl chlorides with ammonia or primary or secondary amines, compounds that all have the structure R_2NH, in which R represents a hydrogen atom or an organic group:

$$R'SO_2Cl + 2R_2NH \longrightarrow R'SO_2NR_2 + R_2NH_2^+Cl^-.$$

sulfonyl chloride; primary or secondary amine or ammonia; sulfonamide; hydrochloride

Sulfonates are obtained by treating alcohols or phenols with sulfonyl chlorides in the presence of a tertiary amine:

$$RSO_2Cl + R'-OH + R''_3N \longrightarrow RSO_2OR' + R''_3NH^+Cl^-.$$

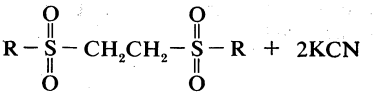
sulfonyl chloride; alcohol; tertiary amine; sulfonate ester; amine hydrochloride

Sulfinic acids. Sulfinic acids usually are made by treating a sulfonyl chloride with finely divided zinc or with sodium sulfite:

$$2RSO_2Cl + 2Zn \longrightarrow (RSO_2)_2Zn + ZnCl_2.$$

sulfonyl chloride; zinc; sulfinic acid (zinc salt); zinc chloride

Aliphatic sulfinic acids can be made by this reaction, but the necessary sulfonyl chlorides are not always available; in such cases, the reaction of ethylene disulfones with potassium cyanide can be employed:



ethylene disulfone; potassium cyanide

$$\longrightarrow 2RSO_2K + N\equiv CCH_2CH_2C\equiv N.$$

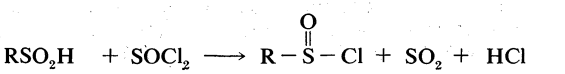
potassium sulfinate; succinonitrile

Several other reactions have been utilized for making sulfinic acids, but they do not have general applicability. One of these reactions is the basis of the manufacture of the sodium salt of hydroxymethanesulfinic acid, sometimes called sodium formaldehydesulfoxylate, used in stripping dyes from textiles and in the discharge printing process for producing dyed designs on fabrics:

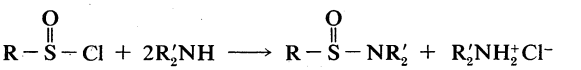
$$Na_2S_2O_4 + CH_2O + H_2O \longrightarrow HOCH_2SO_2Na + NaHSO_3.$$

sodium dithionite; formaldehyde; water; sodium hydroxymethanesulfinate; sodium bisulfite

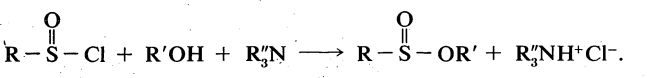
Sulfinic acids are oxidized to sulfonic acids by hydrogen peroxide or by nitric acid but to sulfonyl halides by chlorine, bromine, or iodine. The reaction of a sulfinic acid with thionyl chloride can be used to prepare sulfinyl chlorides, which are useful in the synthesis of sulfinamides and sulfinic esters (sulfinates):



sulfinic acid; thionyl chloride; sulfinyl chloride; sulfur dioxide; hydrogen chloride



sulfinyl chloride; amine; sulfinamide; amine hydrochloride



sulfinyl chloride; alcohol; tertiary amine; sulfinate (ester); tertiary amine hydrochloride

Sulfenic acids. Sulfenic acids or their salts are formed as the initial products in reactions in which the sulfur–sulfur bond of disulfides is broken by the attack of alkalies; the molecules of sulfenic acids react rapidly with one another, producing sulfinic acids and thiols:

$$2RSOK \longrightarrow RSO_2K + RSK.$$

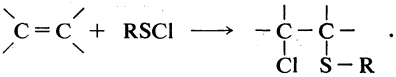
sulfenate (salt); sulfinate (salt); thiol (salt)

The sulfenyl chlorides are more stable than the sulfenic acids; they are formed in the reaction of chlorine with disulfides (see above *Disulfides and polysulfides*). The chlorides can be converted into other derivatives of sulfenic acids, such as amides, esters, or anhydrides:

$$2RSCl + H_2O \longrightarrow RSOSR + 2HCl.$$

sulfenyl chloride; water; sulfenic anhydride; hydrogen chloride

They also react with olefins to produce chlorine-containing sulfides:



olefin; sulfenyl chloride; chloroalkyl sulfide

LESS COMMON ORGANIC SULFUR COMPOUNDS

Organic derivatives of inorganic sulfur-containing acids. Several inorganic acids contain one or more atoms of sul-

fur in their molecules. The most important of these are sulfuric acid, H_2SO_4, and sulfurous acid, H_2SO_3. Most other inorganic acids contain one or more oxygen atoms that can be replaced by sulfur atoms. Nearly all of these acids have organic derivatives, such as esters or amides.

Sulfuric acid. Several esters of sulfuric acid are important industrial chemicals; dimethyl sulfate and diethyl sulfate, made from the alcohols and oleum (a solution of sulfur trioxide in sulfuric acid), are used to introduce methyl and ethyl groups into organic molecules. Certain monoesters of sulfuric acid occur as water-soluble forms in which substances are eliminated from the body.

Thiosulfuric acid. Monoesters of thiosulfuric acid, called Bunte salts, are prepared as intermediates in the synthesis of thiols:

$$RBr + Na_2S_2O_3 \longrightarrow RSSO_3Na + NaBr$$

alkyl bromide; sodium thiosulfate; Bunte salt; sodium bromide

$$RSSO_3Na + NaOH \longrightarrow R-SH + Na_2SO_3.$$

Bunte salt; sodium hydroxide; thiol; sodium sulfite

Sulfurous acid. Esters of sulfurous acid can be made from alcohols and thionyl chloride:

$$2ROH + SOCl_2 \longrightarrow RO-\overset{\overset{O}{\|}}{S}-OR + 2HCl.$$

alcohol; thionyl chloride; dialkyl sulfite; hydrogen chloride

Xanthates, rayon, and cellophane

Carbonic acid. The derivatives of carbonic acid, H_2CO_3, have counterparts in which any or all of the oxygen atoms have been replaced by sulfur atoms. One of the important groups of these compounds is that of the xanthates (Greek *xanthos*, "yellow," from the colour of their copper salts), which are made from hydroxyl compounds and carbon disulfide:

$$ROH + CS_2 + KOH \longrightarrow RO-\overset{\overset{S}{\|}}{C}-SK + H_2O.$$

alcohol; carbon disulfide; potassium hydroxide; xanthate (salt); water

This reaction is used to produce a soluble form of cellulose that can be extruded into an acidic solution, which disrupts the xanthate group, regenerating the cellulose in the form of fibres (rayon) or films (cellophane). Xanthates of simpler alcohols are used as collectors in ore flotation (that is, agents that preferentially attach themselves to the surface of certain minerals, making air bubbles cling to them so that they float to the surface.

Thiourea, the diamide of thiocarbonic acid, is manufactured from hydrogen sulfide and cyanamide or by heating ammonium thiocyanate:

$$H_2S + H_2NC\equiv N \longrightarrow H_2N-\overset{\overset{S}{\|}}{C}-NH_2$$

hydrogen sulfide; cyanamide; thiourea

$$NH_4SCN \xrightarrow{\text{heat}} H_2N-\overset{\overset{S}{\|}}{C}-NH_2.$$

ammonium thiocyanate

It is used as a component of photocopying papers and silver polishes and in a synthesis of thiols that prevents the formation of sulfides as by-products.

Several derivatives of dithiocarbamic acid are used as accelerators of the vulcanization of rubber. The following equations show typical methods of preparing these compounds:

$$2CS_2 + 2R_2NH + ZnO \longrightarrow \left(R_2N-C(=S)-S\right)_2Zn + H_2O$$

carbon disulfide; secondary amine; zinc oxide; zinc dialkyldithiocarbamate; water

$$2R_2N-C(=S)-SNa + H_2O_2 \longrightarrow R_2N-C(=S)-S-S-C(=S)-NR_2 + 2NaOH$$

sodium dialkyldithiocarbamate; hydrogen peroxide; tetraalkylthiuram disulfide; sodium hydroxide

$$R_2N-C(=S)-S-S-C(=S)-NR_2 + KCN$$

potassium cyanide

$$\longrightarrow R_2N-C(=S)-S-C(=S)-NR_2 + KSCN.$$

tetraalkylthiuram sulfide; potassium thiocyanate

Phosphoric acid. Sulfur-containing derivatives of phosphoric acid, H_3PO_4, are useful insecticides, lubricant additives, and ore flotation agents. Some are made from phosphorus pentasulfide or thiophosphoryll chloride:

$$P_4S_{10} + 8ROH \longrightarrow 4(RO)_2PS_2H + 2H_2S$$

tetraphosphorus decasulfide; alcohol; *O,O'*-dialkylphosphorodithioic acid; hydrogen sulfide

$$2(RO)_2PS_2H + ZnO \longrightarrow [(RO)_2PS_2]_2Zn + H_2O$$

zinc oxide; zinc *O,O'*-dialkylphosphorodithioate

$$(CH_3O)_2PS_2H + HC(-CO_2C_2H_5)=HC(-CO_2C_2H_5) \longrightarrow (CH_3O)_2P(=S)-S-CH(CO_2C_2H_5)-CH_2CO_2C_2H_5$$

O,O'-dimethylphosphorodithioic acid; diethyl maleate; Malathion

$$PSCl_3 + 2C_2H_5ONa \longrightarrow (C_2H_5O)_2P(=S)-Cl + 2NaCl$$

thiophosphoryl chloride; sodium ethoxide; *O,O'*-diethyl phosphorochloridothioate

$$(C_2H_5O)_2P(=S)-Cl + O_2N-C_6H_4-ONa$$

sodium *p*-nitrophenoxide

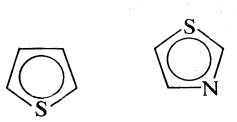
$$\longrightarrow (C_2H_5O)_2P(=S)-O-C_6H_4-NO_2$$

parathion

$$(C_2H_5O)_2P(=S)-Cl + C_2H_5SCH_2CH_2OH \longrightarrow (C_2H_5O)_2P(=S)-OCH_2CH_2SC_2H_5$$

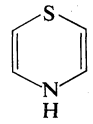
2-ethylthioethanol; Systox

Heterocyclic sulfur compounds. In several groups of organic compounds, sulfur atoms are present in a ring structure (see the article HETEROCYCLIC COMPOUNDS). The most important of these cyclic compounds have the following molecular formulas:

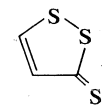


thiophene

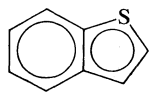
thiazole



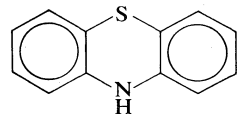
1,4-thiazine



1,2-dithiole-3-thione



thianaphthene or benzothiophene



phenothiazine

BIBLIOGRAPHY. RALPH CONNOR, "Organic Sulfur Compounds," in HENRY GILMAN (ed.), *Organic Chemistry: An Advanced Treatise*, 2nd ed., vol. 1 (1943), is comprehensive but now somewhat dated. W.A. PRYOR, *Mechanisms of Sulfur Reactions* (1962), focusses on reactions of elemental sulfur but contains valuable sections on its organic compounds. NORMAN KHARASCH (ed.), *Organic Sulfur Compounds*, vol. 1 (1961) and NORMAN KHARASCH and C.Y. MEYERS (eds.), *The Chemistry of Organic Sulfur Compounds*, vol. 2 (1966), are collections of chapters on specialized topics. FREDERICK CHALLENGER, *Aspects of the Organic Chemistry of Sulphur* (1959), emphasizes natural products. H. GOLDWHITE treats aliphatic compounds in S. COFFEY (ed.), *Rodd's Chemistry of Carbon Compounds*, 2nd ed., vol. 1, pt. B, ch. 5 (1965); and A.R. FORRESTER and J.L. WARDELL and D.R. HOGG, treat aromatic compounds in the same work, vol. 3, pt. A, ch. 5–6 (1971). Individual groups of organic sulfur compounds are the subject of articles in the 2nd ed. of the *Kirk-Othmer Encyclopedia of Chemical Technology:* M.B. BERENBAUM and R.N. JOHNSON, "Polymers Containing Sulfur," vol. 16, pp. 253–281 (1968); JULIUS FUCHS, "Sulfuric and Sulfurous Esters," vol. 19, pp. 483–498 (1969); W.S. MACGREGOR, "Sulfoxides," vol. 19, pp. 320–337; E.E. GILBERT, " Sulfonic Acids," vol. 19, pp. 311–319; LEONARD DOUB, "Sulfonamides," vol. 19, pp. 255–279; S.D. TURK, "Thiols," vol. 20, pp. 205–218, 1969).

(J.V.K.)

Organometallic Compounds

Organometallic compounds are those substances that contain chemical bonds between carbon atoms and metal atoms, excluding such inorganic salts as the metal carbonates, which show quite different properties. Organometallic compounds constitute an immense group of chemical substances of importance to the history, theory, laboratory practice, and technology of chemistry. In a very general way it may be said that organometallic substances have unique properties that set them apart from both the inorganic and the organic families of compounds. The

The metals in organometallic compounds

metals involved in organometallic compounds may be from any of three categories: (1) the so-called main-group metals of the periodic table of the elements, which include such chemically active metals as lithium and magnesium; (2) the transition series of metals, which include those elements commonly thought of as metallic in character, such as iron, titanium, and platinum; and (3) the metalloids, or partly metal elements, such as silicon and boron. The carbon-containing groups in organometallic compounds also may be of several different varieties, including simple hydrocarbons (groups comprised solely of carbon and hydrogen atoms), unsaturated hydrocarbons (groupings with multiple bonds between carbon atoms), aromatic hydrocarbons (groupings with especially stable multiple bonds), and groups containing atoms other than carbon and hydrogen (oxygen or nitrogen, for example). Metal carbonyls, substances containing metal atoms and units of carbon monoxide, and metal isocyanides, substances containing metals and multiply bonded nitrogen–carbon groups, are classed as organometallic compounds and also as coordination compounds (compounds formed with metal atoms using bonds over and above those normally employed in bonding), revealing the close relationship between organometallic and coordination chemistry.

The carbon–metal bonds in organometallic compounds are of a number of different types: (1) ordinary covalent bonds, or σ-bonds, characterized by pairs of electrons shared between atoms, as, for example, in tetraethyllead, $(C_2H_5)_4Pb$, a gasoline additive; (2) ionic bonds, characterized by complete association of a bonding electron pair with one atom only, resulting in a separation of electrical charge, as in ethylpotassium, $(K^+C_2H_5^-)$; (3) unusual, multicentre covalent bonds, in which bonding extends over a group of atoms (in contrast with the usual two atoms), as in certain polymeric derivatives of light metals, such as ethyllithium, $(C_2H_5Li)_6$; and (4) donor–acceptor bonds, which involve association of a metal atom with an extended system of unsaturation, or overlapping π-orbitals (special electron pathways), as are found in ferrocene, bis(π-cyclopentadienyl)iron, $(\pi\text{-}C_5H_5)_2Fe$.

Physical and chemical properties

A group of compounds as varied as those in the organometallic series naturally includes substances exhibiting a wide range of physical and chemical properties. With regard to stability to heat and to air oxidation they range from alkylpotassium and alkylaluminum compounds, many of which are spontaneously inflammable in air, to ferrocene and the tetraalkyl derivatives of tin, which are very stable. Ferrocene, for example, survives heating to over 470° C in the absence of air. The covalent organometallics vary from crystalline solids, often volatile and usually soluble in organic solvents, to liquids—or even, occasionally, to gases at room temperature; for example, trimethylboron, $(CH_3)_3B$, which boils at −22° C. Many organometallic compounds are highly toxic, particularly those that are volatile.

The first references to organometallic compounds appeared in the early 19th century. In 1827, for example, the Danish chemist W.C. Zeise reported that when platinum chloride was boiled with ethanol, and potassium chloride was added to the solution, a salt was obtained having the composition $KCl\cdot PtCl_2\cdot C_2H_4\cdot H_2O$. More than 100 years later it was realized that Zeise's salt should be formulated as a coordination compound of platinum, viz., $K[C_2H_4PtCl_3]\cdot H_2O$, with carbon–metal bonding. A reasonable theory to account for the bonding of the ethylene molecule (the C_2H_4 unit) to the platinum ion in this compound was not suggested until 1953.

Early historical landmarks in organometallic chemistry were the synthesis of the arsenic compound cacodyl, $(CH_3)_2AsAs(CH_3)_2$, in 1842 by the German chemist Robert Bunsen; the preparation of alkylzinc compounds in 1849 by the English chemist Sir Edward Frankland; the synthesis of nickel carbonyl in 1890 in England by Ludwig Mond, a German-born chemist; and the discovery of organomagnesium compounds in 1900 by the French chemist Victor Grignard. Although organometallic compounds had been studied for over a century, it was not until 1951, when the highly stable substance called ferrocene was discovered and its unusual π-bonded structure became known, that organometallic compounds in themselves became the subject of truly intensive research. Up to that time organometallic compounds had been of importance chiefly as substances that were useful in the synthesis of organic compounds of various types. This situation was particularly true of the organomagnesium compounds which are called, after their discoverer, Grignard reagents.

BONDING IN ORGANOMETALLIC COMPOUNDS

Metals are electropositive elements—that is, their atoms tend to lose electrons to the atoms of other elements with which they combine. The atoms of nonmetals, on the other hand, generally attract electrons in chemical compounds. The concept of electronegativity (the power of an atom in a compound to attract electrons to itself) is useful in understanding the nature of chemical bonds. Thus, all metals have electronegativities less than that of carbon, with the result that in every carbon–metal bond the bonding electrons are attracted somewhat more to the carbon atom than to the metal atom, and the bond is polar (shows a separation of electrical charges) to some degree.

Polar and nonpolar bonding

Differences in electronegativity may be sufficient to produce compounds containing discrete positive and negative ions (atoms that have gained or lost electrons) held together by electrostatic forces. The electronegativity of a carbon atom is enhanced when it is part of an unsaturated system or when halogen atoms, particularly those of fluorine, are joined to it. These charge-transfer, or electrostatic, effects brought about by unsaturation or halogen substitution are responsible for the compound sodium cyclopentadienide having a structure with discrete Na^+ and $C_5H_5^-$ ions, and for the compound bis(trifluoromethyl)mercury, $(CF_3)_2Hg$, resembling more the ionic substance mercuric chloride, $HgCl_2$, than the covalent dimethylmercury, $(CH_3)_2Hg$. The latter substance, unlike its fluorocarbon analogue, is a typically organic, volatile liquid (boiling point 96° C), immiscible with water. More organometallic compounds are covalent in nature than are truly ionic, because electronegativity differences between the organic groups and the metal atoms are not sufficiently pronounced to do other than impart a degree of polarity to a covalent bond. Indeed, with the least electropositive metals, such as tin or lead, their organic derivatives provide excellent examples of covalent compounds, with classical two-electron σ-bonds (bonds formed from ordinary σ-orbitals, as opposed to those from π-orbitals). Consequently, like hydrocarbons (compounds of carbon and hydrogen), these compounds are unaffected by oxygen and water at room temperature.

Organometallic compounds formed from the lighter metals (for example, lithium, beryllium, or aluminum) show somewhat different behaviour. In these compounds the carbon–metal bonds are readily decomposed by oxygen or water, affording metal–oxygen bonds and releasing the organic groups as hydrocarbons. The high reactivity of these covalent carbon–metal bonds is due to their considerable polarity, coupled with the ability of the metals to coordinate (form "secondary" bonds) with reactant molecules as a prelude to decomposition. Indeed, the availability of vacant metal orbitals to accept electron pairs is reflected in the polymeric nature of many of the organic derivatives of lithium, beryllium, and aluminum—*e.g.*, the dimethylberyllium compound, $[(CH_3)_2Be]_x$; the alkyl groups in these compounds link the metal atoms together, and to account for this it is assumed that the electron pairs occupy bonding orbitals embracing more than two atoms. These compounds, in which the number of bonding orbitals exceeds the number of electron pairs, are termed electron-deficient; they represent a bonding akin to that found in the metals, where the electrons are delocalized over many atoms in the structure.

Organometallic compounds of the transition elements

The transition metal–organic compounds include a number of metal alkyls, such as bis(triphenylphosphine)dimethylplatinum $[(C_6H_5)_3P]_2Pt(CH_3)_2$, which contain covalent carbon–metal σ-bonds similar to those found in organometallic compounds of the main-group elements, such as tin. Unless other groups are simultaneously coordinated to the metal atoms, however, this class of compound is usually of low thermal or oxidative stability. Occupation of all coordination sites (typically six) of a metal atom has an important influence on stability. Vacant sites allow coordination of oxygen and water molecules, circumstances that lead to decomposition by paths requiring low activation energies (energies required to initiate reactions). Thus, dimethylmanganese, $Mn(CH_3)_2$, is a bright-yellow powder that readily detonates and reacts instantaneously with air, whereas pentacarbonylmethylmanganese, $CH_3Mn(CO)_5$, is a volatile white solid, mp 95° C, stable to air for long periods in the solid state, and decomposing only slowly in solution.

In addition to the purely σ-complexes, the transition metals also form a large number of coordination compounds in which an unsaturated organic system is associated with the metal. In addition to Zeise's salt and ferrocene (see above), examples include allylpalladium chloride, tricarbonyl(cyclobutadiene)iron, and bis(benzene)-chromium. These compounds are called π-complexes because the π-electrons of the hydrocarbon moieties (combining units) are involved in bonding to the metal atoms. The bonding in these cases is thought to involve reciprocal electron transfers between vacant and occupied π-orbitals of the organic group and filled and unfilled orbitals (electron pathways not involved in ordinary valence bonding) of the metal atom. Generally the metals are in their lower formal valence state—that is, fewer of their *d* electrons are involved in direct bonding—because this condition allows an energetically more favourable interaction of the *d* orbitals with the organic moieties. Many organic derivatives of the transition metals obey the so-called effective-atomic-number rule (a method for determining the degree of electronic saturation), with the result that the final electronic configuration of a fully coordinated metal is that of the next noble gas of higher atomic number. Thus, in the complex π-$C_5H_5Fe(CO)_2CH_3$, the iron atom acquires an additional ten electrons to reach the configuration of the noble gas krypton, which has 18 more electrons than the preceding noble gas, argon; these electrons are supplied by the two CO groups (four electrons), the π-cyclopentadienyl group (five electrons), and the methyl group (one electron).

The effective atomic number rule

Many organic–transition metal complexes show fluxional behaviour—that is, the molecules exist in several chemically equivalent structures and are able to pass from one to another rapidly, a process that may be detected by the technique of nuclear magnetic resonance spectroscopy, which measures certain magnetic phenomena associated with the atomic nuclei.

ORGANOMETALLIC COMPOUNDS OF THE MAIN-GROUP ELEMENTS

Organic compounds of the main-group metals are generally named as a combination of the organic group and the metal; *e.g.*, $Si(C_2H_5)_4$, tetraethylsilicon; C_6H_5HgCl, phenylmercuric chloride; and $Na[B(C_6H_5)_4]$, sodium tetraphenylborate. Occasionally, compounds are named by adapting the name of the metal hydride, in the manner of organic compounds; *e.g.*, $(CH_3)_3SnH$ is called trimethylstannane.

Certain regularities in the trends of properties in the various families of organometallic compounds can be observed. For example, within a family, boiling points increase with increasing atomic weight of the metal atom; thus the boiling points for the tetramethyl derivatives of Group IVa elements (other than carbon) are as follows: silicon, 26.5° C; germanium, 43.4°; tin, 78°; and lead, 110°. Chemical reactivities of organometallic compounds also vary in regular patterns. The small size of the lithium atom results in organolithium compounds having covalent properties, but the other Group Ia organometallics show a steadily increasing reactivity from sodium to cesium.

Methods of preparation. The formation of these compounds may be achieved by three main methods: (1) reaction of a free metal with an organic halide or an unsaturated organic compound; (2) displacement of one metal in an organometallic compound by another, by reaction with a halide of the second metal; and (3) the addition of metal hydrides to unsaturated organic compounds.

1. Reactive main-group metals, such as lithium, sodium, and magnesium, react with organic halides to give either the organometallic compound or an organometallic halide. Examples are:

$$n\text{-}C_4H_9Br + 2Li \longrightarrow n\text{-}C_4H_9Li + LiBr$$

$$(C_6H_5)_3CCl + 2Na \longrightarrow (C_6H_5)_3CNa + NaCl$$

$$CH_3I + Mg \longrightarrow CH_3MgI.$$

Some hydrocarbons—*e.g.*, acetylene and cyclopentadiene—are sufficiently acidic to react directly with a metal, the reaction being accompanied by evolution of hydrogen:

$$C_6H_5C{\equiv}CH + Na \longrightarrow C_6H_5C{\equiv}CNa + \tfrac{1}{2}H_2$$

$$C_5H_6 + K \longrightarrow C_5H_5K + \tfrac{1}{2}H_2.$$

2. Treatment of a metal halide with an organic deriva-

tive of a different metal often leads to the substitution of one metal for the other. This is the most common synthetic method, and it is one of wide scope. A variety of examples are shown:

$$3C_6H_5MgBr + (C_2H_5)_2OBF_3 \longrightarrow (C_6H_5)_3B + 3MgBrF + (C_2H_5)_2O$$

$$(n\text{-}C_3H_7)_3Al + (C_2H_5)_3NBF_3 \longrightarrow (n\text{-}C_3H_7)_3B + (C_2H_5)_3N + AlF_3$$

$$3(C_2H_5)_3Al + GaBr_3 \longrightarrow (C_2H_5)_3Ga + 3(C_2H_5)_2AlBr$$

$$(C_6H_5)_4Sn + 2BCl_3 \longrightarrow 2C_6H_5BCl_2 + (C_6H_5)_2SnCl_2$$

3. The readily available hydrides of metals of Groups IIIa and IVa add across carbon–carbon double and triple bonds, on heating or on exposure to ultraviolet light, often in the presence of a catalyst:

$$(C_6H_5)_3SnH + CH_2{=}CHCN \longrightarrow (C_6H_5)_3SnCH_2CH_2CN$$

$$Cl_3SiH + CH{\equiv}CH \longrightarrow Cl_3SiCH{=}CH_2$$

$$6n\text{-}C_4H_9CH{=}CH_2 + B_2H_6 \longrightarrow 2(C_6H_{13})_3B.$$

In an extensively used industrial process, olefins with terminal methylene, $=CH_2$, groups (1-olefins) react with aluminum and hydrogen under pressure to give the alkyls directly:

$$Al + \tfrac{3}{2}H_2 + 3(\text{l-olefin}) \longrightarrow \text{alkyl}_3Al.$$

Specific synthetic procedures

The syntheses of many specific organometallics are dictated by the particular properties or reactions of a single compound or group of compounds. Thus, cyclopentadienylthallium is insoluble in water, and it may be prepared from cyclopentadiene and thallium hydroxide in aqueous solution because it readily separates from the solution. A large number of organomercury compounds are available by unique reactions: (1) mercuration of aromatic compounds with mercuric acetate, (2) reaction of aryldiazonium salts with mercury, and (3) interaction of arylsulfinic acids with mercuric chloride:

$$C_6H_6 + Hg(O_2CCH_3)_2 \longrightarrow C_6H_5HgO_2CCH_3 + CH_3CO_2H$$

$$ArN_2Cl + Hg \longrightarrow ArHgCl + N_2$$

$$ArSO_2H + HgCl_2 \longrightarrow ArHgCl + SO_2 + HCl,$$

where Ar is an aryl group.

The organosilicon halides, which are technologically important because they are precursors to the silicones, are prepared by a special reaction using copper as catalyst.

$$2CH_3Cl + Si \xrightarrow[300°]{Cu} (CH_3)_2SiCl_2$$

Organogermanium and organotin halides may be prepared similarly.

Many organoboron compounds have compact molecular structures based on cages of boron atoms and carbon atoms ("carboranes"). The most easily prepared carborane is $C_2B_{10}H_{12}$, which has a structure based on an icosahedron (a polyhedron with 20 faces), with ten boron and two adjacent carbon atoms at the apexes and one hydrogen atom attached to each of these. This carborane can be prepared as follows:

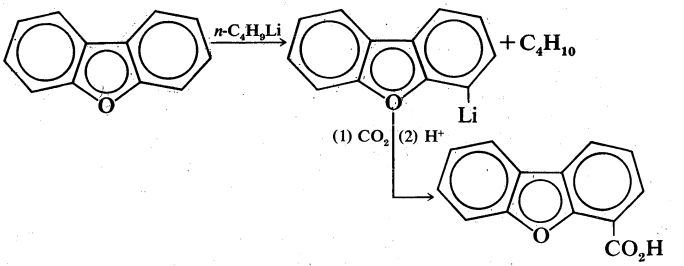
$$B_{10}H_{14} + HC{\equiv}CH \xrightarrow[C_6H_6]{2(C_2H_5)_2S} C_2B_{10}H_{12} + 2H_2.$$

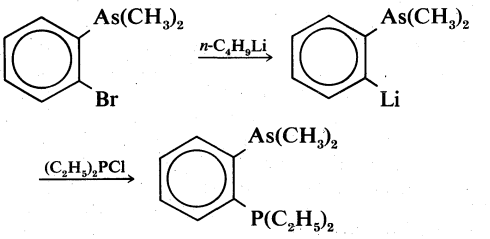
The compound is high melting (294.5–295.5°C) and is very resistant to oxidation and hydrolysis. Substitution at the carbon atoms may be effected by treatment with butyllithium to give $Li_2C_2B_{10}H_{10}$, followed by reaction with reactive halides; *e.g.*,

$$Li_2C_2B_{10}H_{10} + 2(C_6H_5)_2PCl \longrightarrow [(C_6H_5)_2P]_2C_2B_{10}H_{10} + 2LiCl.$$

Principal reactions. The reactions of main-group organometallic compounds are extremely varied, and certain classes of compound—*e.g.*, lithium, boron, and aluminum compounds, and the Grignard reagents—are very useful synthetically.

Organolithium compounds often are not isolated but are used as soon as they are formed, and a frequently employed sequence of reactions is metalation of an organic compound with *n*-butyllithium (a commercially available product) followed by reaction with a selected reagent that reacts with the organometallic intermediate.





A useful reaction occurs with alkyltriphenylphosphonium salts to give phosphinemethylenes. These intermediates react with carbonyl compounds to give olefins, a reaction known as the Wittig reaction.

$$(C_6H_5)_3P \xrightarrow{CH_3Br} (C_6H_5)_3P(CH_3)^+Br^-$$

$$\xrightarrow{C_6H_5Li} (C_6H_5)_3P{=}CH_2 \xrightarrow{R_2CO} (C_6H_5)_3PO + R_2C{=}CH_2$$

Hydroboration

The most important reactions of organoboron compounds, R_3B, are cleavage of the C–B bonds in various ways to give alkanes, alcohols, ketones, and other compounds. The addition of B–H bonds to carbon–carbon double bonds is called hydroboration; the reaction proceeds rapidly in ethers (as solvents) affording diverse species, R_3B, in which the R groups can be complex organic moieties. Many syntheses of purely organic compounds depend on the addition of diborane (B_2H_6), often generated *in situ*, to olefins using a polyether—*e.g.*, diglyme, $(CH_3OCH_2CH_2)_2O$—as solvent.

$$12RCH{=}CH_2 + 3NaBH_4 + 4(C_2H_5)_2OBF_3 \longrightarrow 4(RCH_2CH_2)_3B + 3NaBF_4 + 4(C_2H_5)_2O$$

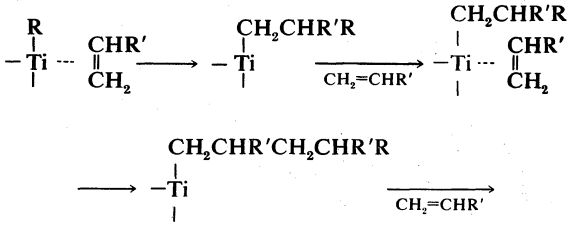
The resulting organoboron compounds are able to undergo a variety of reactions. For example, treatment with alkaline hydrogen peroxide yields an alcohol RCH_2CH_2OH, whereas protonation (addition of a hydrogen ion) with an acid gives RCH_2CH_3; *i.e.*, produces an overall hydrogenation of the starting olefin. Oxidation of $(RCH_2CH_2)_3B$ with chromic acid gives ketones or acids, while treatment with diethylchloramine, $(C_2H_5)_2NCl$, gives RCH_2CH_2Cl (corresponding to an overall addition of hydrogen chloride to the olefin in a reverse manner). Thus, a family of related organic syntheses depends on hydroboration.

The industrial availability of triorganoaluminum compounds has resulted in their becoming reagents of great utility. In the preparation of alkyl derivatives of other metals their use avoids the employment of ethers, which are necessary with Grignard reagents. Aluminum alkyls also cause dimerization, oligomerization, or polymerization of olefins—that is, joining of two olefin molecules, several olefin molecules, or many, respectively. The reactions are of use in the manufacture of detergents, which often are made from long-chain alcohols, as follows:

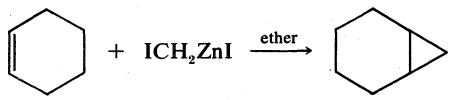
$$(C_2H_5)_3Al \xrightarrow[\sim 100 \text{ atm}]{C_2H_4,\ \sim 100°} [C_2H_5(C_2H_4)_n]_3Al \xrightarrow[\text{oxidation}]{\text{controlled}} H(C_2H_4)_{n+1}OH.$$

Olefin polymerization

If a compound $[CH_3(CH_2)_n]_3Al$ is heated under pressure, the alkyl chain becomes detached in the form of a medium-to-high-molecular-weight linear polyethylene. Such polymers, however, are better obtained by using aluminum–transition metal catalysts (Ziegler–Natta catalysts). For example, treatment of titanium tetrachloride, ($TiCl_4$), with triethylaluminum, $(C_2H_5)_3Al$, in heptane solvent gives a complex that reacts with such olefins as ethylene or propylene. The polymers produced by this reaction are isotactic (sterically regular) solids with desirable properties, as opposed to disordered (atactic) polymers that are oils at room temperature. The mechanism of this industrially important process is a matter of some controversy but probably involves an initial reduction and alkylation of the titanium by the aluminum alkyl, followed by coordination and insertion of olefin molecules in a stereoregular way.



A mixture of zinc and copper reacts with di-iodomethane to give iodomethylzinc iodide (stable in solution only). This compound is useful in reacting with olefins to give cyclopropanes:



norcarane
(bicyclo [4.1.0] heptane)

Organozinc compounds also are formed as transient intermediates in a standard reaction for producing highly substituted esters (the Reformatsky reaction):

$$BrCH_2CO_2C_2H_5 \xrightarrow{Zn} [BrZnCH_2CO_2C_2H_5]$$

$$\xrightarrow[H_2O]{R'_2CO} R'_2C(OH)CH_2CO_2C_2H_5.$$

The most important reaction of organocadmium compounds is the reaction with acid chlorides to give ketones:

$$2RCOCl + R'_2Cd \longrightarrow 2RCOR' + CdCl_2.$$

This method is preferred to the use of Grignard reagents, since the latter react rapidly with the ketone produced.

Analytical chemistry. Organic derivatives of main-group metals may be analyzed by combustion analysis, in which the compound is burned and the carbon and hydrogen present are determined as carbon dioxide and water (the residual metal oxide may be weighed directly). Spectroscopic methods—*e.g.*, infrared and nuclear magnetic resonance spectroscopy—provide a great deal of information, especially about the nature of the organic groups present. Mass spectrometry also is extensively used, often yielding precise molecular-weight measurements.

Industrial applications. Aluminum alkyls are employed industrially in olefin polymerization and in the synthesis of long-chain alcohols (see above). Tetraethyllead is used on a large scale as an additive (antiknock agent) in gasoline, since it is a suitable combustion catalyst, which permits higher compression ratios in the engine. About 1,000,000,000 pounds of tetraethyllead are used each year. The commercial process for manufacture of tetraethyllead is as follows:

Tetraethyllead

$$4NaPb + 4C_2H_5Cl \longrightarrow (C_2H_5)_4Pb + 4NaCl + 3Pb.$$

In gasoline, tetraethyllead is mixed with dibromoethane and dichloroethane; these substances act as "scavengers" within the engine by converting the resulting lead oxide into volatile halides, which are emitted in the exhaust. Volatile lead compounds, however, are highly toxic, and their expulsion into the atmosphere is undesirable. Hence the use of lead alkyls in motor fuels is declining, and other methods are being sought for the production of fuels with superior combustion characteristics (higher "octane number").

Controlled hydrolysis of organosilicon halides produces polymers containing silicon–oxygen and silicon–carbon bonds. The silicones have unusual properties and are widely used for a number of purposes. Organotin compounds are used to stabilize polyvinyl chloride, a polymer used in many common plastic articles. They are also used as fungicides in agriculture and in treating plant diseases that infect potatoes, sugar beets, cacao, groundnuts (peanuts), and other crops. Tricyclohexyltin hydroxide also is used in agriculture, to kill mites. Tributyltin oxide is important as an industrial biocide and surface disinfectant; thus, it is used in preserving wood and in keeping clean the recirculating water of cooling towers.

Certain organomercury compounds have been widely used in pharmacy, for example, as diuretics, but have now been largely replaced by more specific organic compounds. Mercurochrome (2,7-dibromo-4-hydroxymercurifluorescein) is a familiar antiseptic.

ORGANOMETALLIC COMPOUNDS OF THE TRANSITION ELEMENTS

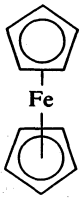
Organic compounds of transition metals are generally named as a combination of the organic group and the metal; *e.g.*, $Mn(CH_3)_2$, dimethylmanganese. When other donor ligands are present, their names precede that of the organic group; *e.g.*, $[(C_6H_5)_3P]_2Pt(C_6H_5)I$ iodobis(triphenylphosphine)phenylplatinum. Similarly, the names of π-bonded organic groups precede those of σ-bonded groups; *e.g.*, $(\pi\text{-}C_5H_5)Fe(CO)_2CH_3$, dicarbonyl-$\pi$-cyclopentadienyl-$\sigma$-methyliron. As these names tend to be cumbersome, many of the commonly encountered compounds have acquired simpler names; *e.g.*, $Fe(\pi\text{-}C_5H_5)_2$, ferrocene or bis(π-cyclopentadienyl)iron; $(\pi\text{-}C_5H_5)Mn(CO)_3$, cymantrene or π-cyclopentadienyltricarbonylmanganese. Organic complexes of the transition elements may be classified according to the number of electrons donated by the organic ligand (that is, the substance complexed with the metal atom; see table). A great many separate compounds are known because many members of each particular class of organic molecule can bond to a given metal. There are, for example, a vast number of complexes of formula diene-$Fe(CO)_3$ because the $Fe(CO)_3$ group shows great affinity for any 4π-electron system based on two conjugated double bonds, of which there are many. Moreover, a particular metal may form complexes with all types of ligand; for example, in the series $CH_3Mn(CO)_5$, $[(\pi\text{-}C_2H_4)Mn(CO)_5]^+$, $(\pi\text{-}C_3H_5)Mn(CO)_4$, $(\pi\text{-}C_6H_8)Mn(CO)_3H$, $(\pi\text{-}C_5H_5)Mn(CO)_3$, and $[(\pi\text{-}C_6H_6)Mn(CO)_3]^+$, manganese is associated with organic groups that donate—in order—from one to six electrons to its valence shell. In each case the effective atomic number rule is obeyed, the manganese atom having 18 electrons in its outermost shell.

Representative Ligands in Organotransition-Metal Complexes

number of electrons	class	examples of complexes
1	yl	$CH_3Re(CO)_5$, white, mp 120°
2	alkene	$[(\pi\text{-}C_2H_4)_2RhCl]_2$, orange red, decomp. > 100°
	carbene	$CH_3(CH_3O)CCr(CO)_5$, white, mp 34°
3	enyl*	$[(\pi\text{-}C_3H_5)PdCl]_2$, yellow, mp 145°
4	diene	$(\pi\text{-}C_4H_6)Fe(CO)_3$,† yellow oil
5	dienyl	$(\pi\text{-}C_5H_5)_2Ni$, green, mp 173°
	ollyl	$[(CH_3)_4N][\{\pi\text{-}1,2\text{-}B_9H_9C_2(CH_3)_2\}_2Fe]$‡, red, mp 247°
6	arene	$(\pi\text{-}C_6H_6)_2W$, green, decomposes 110°
	triene	$(\pi\text{-}C_7H_8)Mo(CO)_3$,§ orange red, mp 101°
7	trienyl	$(\pi\text{-}C_7H_7)V(CO)_3$, dark green, decomposes 134°

*Also called "π-allylic" ligands. †C_4H_6 is buta-1,3-diene. ‡The ligand $\{1,2\text{-}B_9H_9C_2(CH_3)_2\}^{2-}$ is an 11-atom B_9C_2 species (an icosahedron with an apex removed) having a face comprising five atoms (B_3C_2) with π-type molecular orbitals similar to those of π-C_5H_5. In forming the complex the metal atom completes the icosahedron. §C_7H_8 is cycloheptatriene.

Although this rule is very useful, it is not followed invariably; the exceptions may be rationalized by considering the relative energies of the *s*, *p*, and *d* orbitals (designations for electron pathways within the various shells). For metals of the scandium and titanium groups of the periodic table, the 3*d* orbitals have a relatively high energy compared with the other orbitals (4*s* and 4*p*) in the same valence band, and their importance in bonding, therefore, is less. At the other end of the transition series (in the neighbourhood of copper and zinc), the *d* orbitals are of relatively low energy and cannot easily be regarded as valence orbitals. Across the series vanadium, chromium, manganese, iron, cobalt, and nickel, the 3*d* orbitals decrease in energy more rapidly than do the 4*s* and 4*p* orbitals, but all three orbital types are generally sufficiently close in energy to be suitable for metal–ligand bonding. Under these conditions, maximum bonding—with associated stability—occurs when all orbitals are filled with electron pairs. This occurs when the metal atom has an 18-electron configuration; *i.e.*, the five *nd*, one $(n+1)s$, and three $(n+1)p$ orbitals ($n = 3$, 4, or 5) are all occupied by metal and ligand electrons. Nowhere is this better exemplified than in the case of ferrocene, a compound of exceptional stability and one in which the iron atom (eight electrons) is located symmetrically between two π-C_5H_5 groups (ten electrons), forming a classical "sandwich" molecular structure.



Exceptions to the effective-atomic-number rule

It follows from this reasoning that exceptions to the effective-atomic-number rule are found at both ends of the transition series, in which cases the 3*d* orbitals are too high (as with titanium) or too low (copper) in energy. Sixteen-electron environments are prevalent with organotitanium compounds—*e.g.*, $(\pi\text{-}C_5H_5)_2Ti(C_6H_5)_2$—and 14 with the metals of the copper group [*e.g.*, $(C_6H_5)_3PAuCH_3$]. With the latter, not only have the *d* electrons become low-energy (core) electrons but the *p* orbitals have become relatively high in energy compared with the *s* orbitals; consequently only one *p* orbital is used in bonding, in combination with the *s* orbital. A similar effect is responsible for the existence of numerous 16-electron complexes of nickel, palladium, and platinum, in which the relative energies of the *nd*, the $(n+1)s$, and the $(n+1)p$ orbitals are such as to lead to the use of one of the *d*, the *s*, and two of the *p* orbitals in bonding—*e.g.*, $[(C_2H_5)_3P]_2Pd(CH_3)I$.

Methods of preparation. It is convenient to consider separately the methods for preparing the π-complexes and the σ-compounds of the transition metals.

π-Complexes. Treatment of a suitable metal halide with an organic derivative of an alkali metal or a Grignard reagent using an ether (often tetrahydrofuran) as solvent is often a method for preparing π-complexes of transition elements.

$$2CH_2=CHCH_2MgBr + NiCl_2 \longrightarrow (\pi\text{-}C_3H_5)_2Ni + 2MgClBr$$

$$2C_5H_5Na + FeCl_2 \longrightarrow (\pi\text{-}C_5H_5)_2Fe + 2NaCl$$

$$2C_8H_8K_2 + UCl_4 \longrightarrow (\pi\text{-}C_8H_8)_2U + 4KCl$$

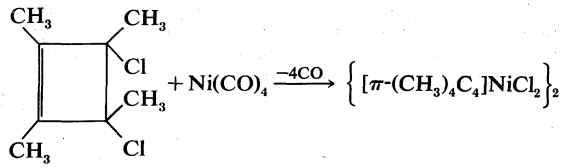
Reactions with olefins

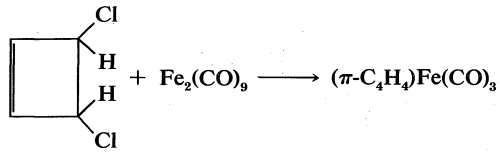
The reaction between a metal complex and an unsaturated hydrocarbon is the most generally useful method of preparing π-complexes; it has several modifications. In one of the simplest forms of this reaction, carbonyls are displaced from metal carbonyl complexes by unsaturated hydrocarbons. The reaction is carried out by heating the reactants together or by exposing them to ultraviolet irradiation in a suitable solvent (often a saturated hydrocarbon).

$$Fe(CO)_5 + CH_2=CHCH=CH_2 \xrightarrow{-2CO} (\pi\text{-}C_4H_6)Fe(CO)_3$$

$$Co_2(CO)_8 + 2C_5H_6 \longrightarrow 2(\pi\text{-}C_5H_5)Co(CO)_2 + 4CO + H_2$$

The more reactive metal carbonyls abstract chlorine atoms from carbon–chlorine bonds with concomitant oxidation of the metal to a metal chloride. This property has been used to isolate organic molecules that are not normally stable in the free state but that can be stabilized by coordination with a metal atom. Cyclobutadiene and its derivatives, sought by chemists for over 80 years, were finally prepared by being stabilized in this way.





Another example of this type of stabilization is found in the preparation of the trimethylenemethane group:

$$CH_2=C(CH_2Cl)_2 + Fe_2(CO)_9 \longrightarrow [\pi\text{-}(CH_2)_3C]Fe(CO)_3.$$

Reactions with acetylenes

Reactions between acetylenes and metal carbonyls (or their derivatives) can be a fruitful source of unusual organometallic compounds. A variety of reactions is possible, including coordination of the triple bond, di- or trimerization of the acetylene, and formation of ligands containing carbonyl groups, with the result that highly complex mixtures of products are obtained. For example, the mixture that results from treatment of iron carbonyls with acetylenes has given complexes of nearly 30 different types, containing such diverse ligands as cyclobutadienes, cyclopentadienones, and quinones.

The preparation of many π-complexes depends upon an initial reduction of a transition-metal compound in the presence of an unsaturated hydrocarbon to serve as the ligand. The synthesis of bis(benzene)chromium depends upon this technique:

$$3CrCl_3 + 2Al + AlCl_3 + 6C_6H_6 \longrightarrow 3[(\pi\text{-}C_6H_6)_2Cr]^+[AlCl_4]^-$$

$$2[(\pi\text{-}C_6H_6)_2Cr]^+ + S_2O_4^{2-} + 4OH^- \longrightarrow 2(\pi\text{-}C_6H_6)_2Cr + 2H_2O + 2SO_3^{2-}.$$

With this procedure, or modifications of it, arene complexes of many of the transition metals have been prepared. Aromatics other than benzene also may be used.

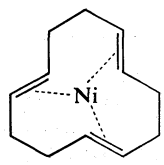
Reduction of nickel acetylacetonate with aluminum alkyls in the presence of triphenylphosphine and ethylene produces a yellow crystalline compound,

$$(\pi\text{-}C_2H_4)\,Ni[P(C_6H_5)_3]_2.$$

With cycloocta-1,5-diene, rather than ethylene and triphenylphosphine, the same method produces

$$(\pi\text{-}C_8H_{12})_2Ni,$$

a yellow solid of melting point 142° C that is unstable in air but is reasonably stable to heat. An unusual complex that is prepared by a similar method is all-*trans*-1,5,9-cyclododecatrienenickel,



in which the metal atom is coordinated only to three double bonds. Many olefin complexes of nickel are used in further syntheses. Thus, $(\pi\text{-}C_{12}H_{18})Ni$ catalyzes the trimerization of butadiene, giving isomers of cyclododecatriene. In some syntheses the hydrocarbon itself functions as the reducing agent. Thus, treatment of rhodium trichloride with butadiene gives bis(butadiene)rhodium(I) chloride, $(\pi\text{-}C_4H_6)_2RhCl$.

Ligand-exchange reactions

Ligand-exchange reactions, those in which an organic group is transferred from one metal to another, also are used to prepare π-complexes. For example, treatment of $(C_6H_5CN)_2PdCl_2$ with diphenylacetylene, followed by reaction with hydrogen halides, gives dimeric tetraphenylcyclobutadienepalladium halides $\{[\pi\text{-}(C_6H_5)_4C_4]PdX_2\}_2$ (X = Cl or Br). The tetraphenylcyclobutadiene group then may be readily transferred to other metals. Thus, reaction of one of the above palladium complexes with $(\pi\text{-}C_5H_5)_2Co$ gives $[\pi\text{-}(C_6H_5)_4C_4]Co(\pi\text{-}C_5H_5)$, and reaction with $Fe(CO)_5$ gives $[\pi\text{-}(C_6H_5)_4C_4]Fe(CO)_3$.

σ-Compounds. Although simple alkyls and aryls of the transition metals are generally very unstable—*e.g.*, $(CH_3)_2Ni$ decomposes above −110° C—derivatives containing such ligands as carbon monoxide, phosphines, and π-bonded hydrocarbons are often reasonably robust chemically. A number of synthetic routes to these compounds are known.

Treatment of transition-metal-complex halides with Grignard reagents or with organolithium compounds frequently leads to replacement of the halogen with alkyl or aryl groups.

$$[(C_6H_5)_3P]_2PdCl_2 + 2C_6H_5MgBr \longrightarrow [(C_6H_5)_3P]_2Pd(C_6H_5)_2 + 2MgBrCl$$

$$CrCl_3(THF)_3 + 3C_6H_5MgBr \longrightarrow (C_6H_5)_3Cr(THF)_3 + 3MgBrCl$$

[THF = tetrahydrofuran, C_4H_8O]

$$(C_6H_5)_3PAuCl + CH_3Li \longrightarrow (C_6H_5)_3PAuCH_3 + LiCl$$

Reactions between complex metal anions and organic halides yield organometallic compounds in which the anion has replaced the halide.

$$Na^+[Mn(CO)_5]^- + CH_3I \longrightarrow CH_3Mn(CO)_5 + NaI$$

$$Na^+[W(CO)_3(\pi\text{-}C_5H_5)]^- + C_2H_5I \longrightarrow C_2H_5W(CO)_3(\pi\text{-}C_5H_5) + NaI$$

The utility of this method is enhanced by the fact that acyl complexes, prepared by reaction of the anions with acid halides, can often be decarbonylated, either thermally or on ultraviolet irradiation, to give the corresponding alkyl compounds:

$$Na^+[Fe(CO)_2(\pi\text{-}C_5H_5)]^- + C_6H_5COCl$$
$$\longrightarrow C_6H_5COFe(CO)_2(\pi\text{-}C_5H_5) + NaCl$$
$$\xrightarrow{h\nu} C_6H_5Fe(CO)_2(\pi\text{-}C_5H_5) + CO.$$

Insertion reactions

Insertion of unsaturated groups into metal–hydrogen or metal–alkyl σ-bonds produces organometallic compounds with more extended alkyl chains.

$$(\pi\text{-}C_5H_5)Fe(CO)_2H + CH_2{=}CHCH{=}CH_2$$
$$\longrightarrow CH_3CH{=}CHCH_2Fe(CO)_2(\pi\text{-}C_5H_5)$$

$$CH_3Mn(CO)_5 + CF_3C{\equiv}CCF_3 \longrightarrow (CH_3)(CF_3)C{=}C(CF_3)Mn(CO)_5$$

A useful synthesis depends on the so-called oxidative-addition or oxidative-elimination reactions. Many transition-metal complexes, in which the metal atom is of low valence and which are not so fully coordinated as they might be, react with electron-acceptor (electrophilic) molecules to form compounds in which the valence state of the metal can be regarded formally as having increased by two units. The reacting molecule, which contains a bond designated A–B, may be cleaved by the metal atom to form in its coordination sphere two separate bonds, M–A and M–B; or, if the bond between A and B is a multiple one, the product may contain an A–B unit coordinated to the metal atom. The following reactions exemplify the utility of these kinds of processes in organometallic syntheses:

$$Fe(CO)_5 + C_3F_7I \longrightarrow C_3F_7Fe(CO)_4I + CO$$

$$[(C_6H_5)_3P]_3RhCl + C_2H_4$$
$$\longrightarrow (\pi\text{-}C_2H_4)Rh[P(C_6H_5)_3]_2Cl + (C_6H_5)_3P$$

$$[(C_6H_5)_3P]_2Ir(CO)Cl + CH_3I$$
$$\longrightarrow CH_3Ir[P(C_6H_5)_3]_2(CO)(Cl)(I)$$

$$[(C_6H_5)_3P]_3Pt + CH_3C{\equiv}CCH_3$$
$$\longrightarrow [\pi\text{-}(CH_3)_2C_2]Pt[P(C_6H_5)_3]_2 + (C_6H_5)_3P.$$

Principal reactions. Each class of organometallic complex (see table) has its own characteristic chemical and physical properties associated with the presence of a particular type of organic group. The degree to which group properties are shown, however, depends on the firmness with which the ligand is attached to the metal atom. Consequently, there are no reactions generally applicable to all those compounds, and even for a particular class large gradations in reactivity are observed.

Many organometallics are endothermic compounds; that is, their formation from the constituent elements involves an increase in free energy. Hence there is an inherent thermal instability with respect to decomposition into the metals and the hydrocarbons that make them up. Nevertheless, many hundreds of these compounds can be prepared, and in some cases they show considerable resistance to thermal decomposition, because it is necessary for a favourable free-energy change to be accompanied by a reaction path of sufficiently low energy if decomposition is to occur at a measurable rate. In many cases, decomposition may be kinetically rather than thermodynamically controlled—that is, it may not be observed normally because the rate is too slow, even though the energy requirements are favourable. The limiting process may be the initial breaking of an M–C bond to form M· and R· radicals (homolytic dissociation) or M^+ and R^- ions (heterolytic dissociation). Such bond dissociations are likely to be facilitated when empty low-lying orbitals are present in the metal atom, since electrons can temporarily reside in these. In both homolytic and heterolytic processes, the carbon fragments are generally reactive and rapidly form stable products; for example, by dimerization. It is the formation of these more stable products that makes the decomposition irreversible; but if there is a high activation energy (energy needed to initiate the process) for the dissociation, decomposition will not occur.

Oxidation

All organometallic compounds are thermodynamically unstable to oxidation because of the large favourable free-energy change in the conversion into metal oxide, carbon dioxide, and water. In this respect organometallic compounds are no different from purely organic compounds except that with the former an additional product (metal oxide) is involved. Again, however, kinetic factors (considerations of reaction rate) are often decisive both in enhancing stability toward oxygen or lowering resistance to oxidative attack. In this case also instability can be associated generally with the presence of empty low-lying orbitals. Similar considerations apply to hydrolytic stability. Hydrolysis often involves an initial bonding with a water molecule, a process facilitated by an empty metal orbital. The M–C bond polarity, however, also is an important factor in hydrolysis because a further step in the process is the elimination of R^- as a hydrocarbon RH.

The existence of a closed-shell (18-electron) configuration has a bearing on the reactivity of a complex (see above). For complexes in which the metal atom already follows the effective-atomic-number rule, reaction may require displacement of one or more ligand molecules in order to provide a vacant coordination site or sites for a reactant molecule. This situation may lead only to the exchange of one ligand for another, but often a molec-

ular rearrangement follows, affording an entirely new structure.

Cleavage and insertion reactions are the hallmark of σ-bonds between transition metals and alkyl or aryl groups. Cleavage reactions are those in which the C–M bond is split, with attendant formation of a new bond to an inorganic atom. Typical reagents that may effect cleavage are acids, hydrogen, halogens, and metal halides.

$$(CH_3)_2Pt[P(C_2H_5)_3]_2 + HCl \longrightarrow CH_3Pt[P(C_2H_5)_3]_2Cl + CH_3Cl$$

$$C_6H_5Pt[P(C_2H_5)_3]_2Cl + H_2 \longrightarrow HPt[P(C_2H_5)_3]_2Cl + C_6H_6$$

$$(C_6H_5)_2Ti(\pi\text{-}C_5H_5)_2 + 2Cl_2 \longrightarrow Cl_2Ti(\pi\text{-}C_5H_5)_2 + 2C_6H_5Cl$$

$$(C_6H_5)_3Cr(THF)_3 + 3HgCl_2 \longrightarrow 3C_6H_5HgCl + CrCl_3 + 3THF$$

The term insertion designates a reaction in which a C–M bond is broken, with the reactant molecule's being placed between the two parts and joined to both. This process is the basis of the Ziegler–Natta olefin polymerizations (see above), but it also is known in many simpler forms.

$$C_2H_5Mo(CO)_3(\pi\text{-}C_5H_5) + CO \longrightarrow C_2H_5COMo(CO)_3(\pi\text{-}C_5H_5)$$

$$C_6H_5Fe(CO)_2(\pi\text{-}C_5H_5) + (C_6H_5)_3P \longrightarrow C_6H_5COFe(CO)[P(C_6H_5)_3](\pi\text{-}C_5H_5)$$

$$CH_3Mn(CO)_5 + SO_2 \longrightarrow CH_3SO_2Mn(CO)_5$$

Conversion of σ- to π-complexes

In many reactions, a σ-bonded organic group is converted into a π-complexed group. A simple example involves the complex $\sigma\text{-}CH_2{=}CHCH_2Mn(CO)_5$, which on heating or exposure to ultraviolet light releases carbon monoxide and forms $(\pi\text{-}C_3H_5)Mn(CO)_4$. Similar decarbonylations (loss of carbonyl groups) have been observed with σ-allyl carbonyl derivatives of cobalt, iron, chromium, molybdenum, and tungsten, in each case π-allyl complexes of these metals being formed.

In 1919 and afterward many polyphenylchromium compounds were prepared by treating chromium chloride ($CrCl_3$) with phenylmagnesium bromide. In 1954 it was shown that several of these compounds were cationic arene–chromium complexes; *e.g.*,

$$[(\pi\text{-}C_6H_6)Cr(\pi\text{-}C_6H_5C_6H_5)]^+,$$

which had been formed by rearrangement of $\sigma\text{-}C_6H_5Cr$ groups either in a hydrolysis step or thermally. Most of these transformations are exceedingly complex, and a multitude of products is formed. The history of the polyphenylchromium complexes is an example of the failure to recognize the true nature of a series of compounds because of the lack of sufficiently sophisticated analytical techniques.

Several σ-alkyl complexes have been converted to cationic olefin π-complexes by hydride abstraction (removal of an H^-) or by protonation (addition of an H^+).

$$\begin{matrix} CH_3CH_2CH_2Fe(CO)_2(\pi\text{-}C_5H_5) \xrightarrow{[(C_6H_5)_3C]^+} \\ CH_2{=}CHCH_2Fe(CO)_2(\pi\text{-}C_5H_5) \xrightarrow{H^+} \end{matrix}$$

$$\left[\begin{matrix} CH_3CH \\ \| \\ CH_2 \end{matrix} \cdots Fe(CO)_2(\pi\text{-}C_5H_5)\right]^+$$

An unusual group of compounds contains a fluorocarbon group (chemical unit of only carbon and fluorine atoms) σ-bonded to a transition metal. These complexes are more stable chemically than their alkyl or aryl counterparts—*e.g.*, $C_2F_5Mn(CO)_5$ decomposes at 150° C, whereas $C_2H_5Mn(CO)_5$ decomposes at 25°; $C_6F_5Mn(CO)_5$ decomposes at 170°, whereas $C_6H_5Mn(CO)_5$ decomposes at 100°. Thermodynamic and kinetic factors both are thought to be responsible for the enhanced stability of the fluorocarbon compounds. The fluorocarbon–metal bonds are more ionic than hydrocarbon–metal bonds; that is, they show more complete association of the shared electron pair with the carbon atom of the bond. This effect is due to the strongly electron-withdrawing character of the fluorocarbon groups, which tend to pull the electron pair farther from the metal atom. Moreover, alkyl–metal bonds frequently decompose by a reaction that is the reverse of an insertion reaction, $M - C_2H_5 \rightarrow M - H + C_2H_4$. The analogous reaction involving elimination of a fluoroolefin and formation of a metal fluoride, although thermodynamically favoured, requires more activation energy in order to proceed at a measurable rate. Because of their stability, bonds between fluorocarbon groups and transition metals show little or no tendency to undergo insertion reactions. More than 500 complexes of this type are known, produced by reactions such as

$$HMn(CO)_5 + CF_2{=}CF_2 \longrightarrow HCF_2CF_2Mn(CO)_5$$

$$NaRe(CO)_5 + C_6F_6 \longrightarrow C_6F_5Re(CO)_5 + NaF$$

$$(\pi\text{-}C_5H_5)_2ZrCl_2 + 2C_6F_5Li \longrightarrow (C_6F_5)_2Zr(\pi\text{-}C_5H_5)_2 + 2LiCl$$

$$(C_4H_9NC)_4Ni + 2CF_2{=}CF_2 \longrightarrow \begin{matrix} CF_2-CF_2 \\ | \qquad \diagdown \\ CF_2-CF_2 \end{matrix} Ni(CNC_4H_9)_2.$$

Reactions of π-complexes

The reactions of π-complexes are extremely extensive; it has been said that the cyclopentadienyl group forms more complexes than any other organic ligand. Whether or not this is strictly true, the range of complexes formed by this five-membered ring is certainly great. Simple cyclopentadienyl derivatives, $(\pi\text{-}C_5H_5)_2M$, of the first row transition elements, from vanadium through nickel, all have the same structure, consisting of a metal atom sandwiched between two five-membered rings. This situation gives rise to a fairly constant melting point (173° C), although the differing electronic configurations result in markedly different reactivities. Thus, in marked contrast with the highly stable ferrocene, chromocene bursts into flame spontaneously in air.

One of the main factors responsible for the many ramifications of coordination chemistry is the ability of a metal atom to bond more than one type of ligand within its coordination sphere at the same time. Moreover, certain ligands stabilize the bonding of others, and this is especially true of the $\pi\text{-}C_5H_5$ group, as illustrated by the existence of the following compounds of metals of the first transition series:

$$(\pi\text{-}C_5H_5)_2Ti(C_6F_5)_2,$$

$$(\pi\text{-}C_5H_5)V(\pi\text{-}C_7H_7),$$

$$(\pi\text{-}C_5H_5)Cr(NO)_2Cl,$$

$$(\pi\text{-}C_5H_5)Mn(\pi\text{-}C_6H_6),$$

$$(\pi\text{-}C_5H_5)Fe(\pi\text{-}C_4H_4N),$$

$$(\pi\text{-}C_5H_5)Co(\pi\text{-}C_5H_6),$$

$$(\pi\text{-}C_5H_5)Ni(CO)CF{=}CFCF{=}CF_2.$$

The cyclopentadienylmetal carbonyls undergo a number of chemical reactions and show several resemblances to the metal carbonyls. For example, the $(\pi\text{-}C_5H_5)M$ group is isoelectronic with (*i.e.*, has the same number of electrons as) the $M'(CO)_3$ group, where M′ is the element preceding M in the periodic table. Many complexes containing the $Fe(CO)_3$ group thus have counterparts containing the $(\pi\text{-}C_5H_5)Co$ group; similarly, the chemistry of $[(\pi\text{-}C_5H_5)Fe(CO)_2]_2$ bears many resemblances to that of $[Mn(CO)_5]_2$, and that of $(\pi\text{-}C_5H_5)Mn(CO)_3$ to

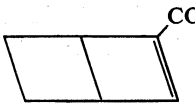
$Cr(CO)_6$. The cyclopentadienylmetal carbonyls also resemble the metal carbonyls in that in both cases there are numerous polynuclear complexes containing metal–metal bonds; *e.g.*, $[(\pi\text{-}C_5H_5)Ni]_3(CO)_2$, $[(\pi\text{-}C_5H_5)RhCO]_3$, or $[(\pi\text{-}C_5H_5)FeCO]_4$.

Aromatic reactions of π-complexes

A striking feature of the chemistry of certain cyclopentadienyl complexes is the ability of the five-membered ring to undergo reactions typical of aromatic systems. Such chemisty was first demonstrated for ferrocene, which can be acylated (substituted with an acyl group) under Friedel-Crafts conditions—that is, with aluminum chloride as catalyst. Furthermore, one of the hydrogen atoms can be replaced by a lithium atom or a mercuriacetate group, using butyllithium or mercuric acetate, respectively. Introduction of these and other functional groups has allowed the synthesis of hundreds of ring-substituted ferrocene compounds. Many other cyclopentadienylmetal compounds, on the other hand, do not undergo substitution reactions for either of several reasons: (1) the metal atom interferes with the reaction—*e.g.*, becomes oxidized; or (2) the complex does not survive the reaction conditions. Aromatic behaviour also has been shown by $(\pi\text{-}C_5H_5)_2M$, M = Ru, Os; $(\pi\text{-}C_5H_5)Mn(CO)_3$, termed "cymantrene"; $(\pi\text{-}C_5H_5)Cr(CO)_2NO$; $(\pi\text{-}C_5H_5)_2ReH$; and certain other compounds.

In tricarbonyl(cyclobutadiene)iron, the metal–ligand bonding is sufficiently strong to allow the cyclobutadiene group to show aromaticity. For example, CH_3CO and HgCl groups may be substituted for H atoms of the ring to give useful intermediates for further syntheses. On the other hand, $(\pi\text{-}C_4H_4)Fe(CO)_3$ can be oxidized, releasing cyclobutadiene, which may be characterized by the reactions it undergoes immediately.

$$(\pi\text{-}C_4H_4)Fe(CO)_3 \xrightarrow[HC{\equiv}CCO_2CH_3]{Ce(IV)}$$



When this reaction is carried out with $C_6H_5C{\equiv}CH$, the organic compound produced is in the unusual situation of having a classical benzene ring bonded to its valence isomer, Dewar benzene (named for the Scottish chemist who proposed this structure for ordinary benzene), the structure shown in the above formula.

Analytical chemistry. Analytical methods for the organic derivatives of the transition metals are generally similar to those for the organic derivatives of the main-group metals. Spectroscopic techniques are an important source of information, and with metal carbonyl derivatives, infrared spectroscopy is extremely valuable in determining the number of carbonyl groups and their arrangement around the metal. Proton magnetic resonance spectra are of help in interpreting the nature of the bonded hydrocarbon groups, although the number of fluxional systems discovered means that care must be exercised in the interpretation of apparently simple spectra. Other spectroscopic methods—*e.g.*, Mössbauer, and electron-spin resonance techniques—have been used where appropriate.

The increasing complexity of the complexes discovered makes X-ray crystallographic methods the final arbiter of structures; about 150 new structures are determined by this method each year. This analytical technique has been greatly facilitated by computer processing of data.

Industrial applications. Relatively few organometallic compounds of the transition metals find direct practical application. On the other hand, their value as reaction intermediates is almost inestimable.

Use in the petroleum industry

As already indicated, organometallics are widely employed in the petroleum-based industries. Certain industrial-scale syntheses depend upon the transient existence of catalysts in the form of metallo-organic complexes, which are constantly regenerated. Hydroformylation, or the conversion of olefins into aldehydes or ketones using carbon monoxide and hydrogen under pressure (100–300° C; 100 atm), is catalyzed by cobalt carbonyl and related complexes, including those of rhodium. The catalytic material in the system is thought to be a hydridocarbonyl—such as $HCo(CO)_3$—derived from $Co_2(CO)_8$, and the various reaction steps include insertion reactions (olefin inserted into metal–hydrogen bonds and CO into metal–alkyl bonds) as well as oxidative-addition and reductive-elimination reactions. The overall process clearly depends upon organometallic intermediates and their cleavage reactions.

$$Co_2(CO)_8 + H_2 \rightleftharpoons 2HCo(CO)_3 + 2CO$$

$$HCo(CO)_3 \xrightarrow[\pi\text{-complex formation}]{RCH{=}CH_2} (\pi\text{-}RCH{=}CH_2)Co(CO)_3H$$

$$(\pi\text{-}RCH{=}CH_2)Co(CO)_3H \xrightarrow{\text{insertion}} RCH_2CH_2Co(CO)_3$$

$$RCH_2CH_2Co(CO)_3 \xrightarrow[\text{insertion}]{CO} RCH_2CH_2COCo(CO)_3 \xrightarrow[\text{oxidative addition}]{H_2} RCH_2CH_2COCo(CO)_3H_2$$

$$RCH_2CH_2COCo(CO)_3H_2 \xrightarrow{\text{reductive elimination}} \underset{\text{product}}{RCH_2CH_2CHO} + \underset{\text{catalyst}}{HCo(CO)_3}$$

The mechanisms of such reactions are related to the mechanisms of catalytic hydrogenation of olefins or acetylenes by certain metal complexes. In both cases, the catalyst carries out three basic reactions: (1) activation of molecular hydrogen by breaking of the H–H bond and generation of two M–H bonds; (2) activation of the olefin by formation of a π-complex; and (3) transfer of hydrogen to the olefin (by insertion to form a σ-metal–alkyl group, followed by cleavage of this group by hydrogen). It is apparent that conversion of a π-complex into a σ-complex represents a key step in the overall process.

Use in hydrolysis reactions

Another useful synthesis depends in part for its success on the fact that coordinated olefins become susceptible to a hydrolysis reaction that leads to oxidation. In this way ethylene is converted to acetaldehyde; propylene, to acetone; and ethylene and acetic acid to vinyl acetate (which serves as a monomer for polymerization). In one variation of this basic reaction, ethylene and oxygen are passed together through an aqueous solution of palladium and cupric salts. The overall reaction ($C_2H_4 + \frac{1}{2}O_2 \rightarrow CH_3CHO$) results from a combination of three steps:

$$(1)\quad C_2H_4 + PdCl_2 + H_2O \longrightarrow CH_3CHO + Pd + 2HCl$$

$$(2)\quad Pd + 2CuCl_2 \longrightarrow PdCl_2 + 2CuCl$$

$$(3)\quad 2CuCl + 2HCl + \tfrac{1}{2}O_2 \longrightarrow 2CuCl_2 + H_2O.$$

Reaction (1), the reaction that forms the product, is itself composed of several steps, including formation of the π-complex anion $[(\pi\text{-}C_2H_4)PdCl_3]^-$ (related to the anion of Zeise's salt) and its rearrangement after hydrolysis to a σ-complex $[CH_2OHCH_2PdCl_3]^{2-}$.

Other industrial reactions catalyzed by transition-metal complexes include the isomerization, dimerization, and polymerization of olefins, the trimerization of acetylenes, the tetramerization of acetylene (to cyclooctatetraene), and the dismutation (conversion to products with both larger and smaller carbon skeletons) of olefins (*e.g.*, the conversion of propene into ethylene and 2-butene). In each of these reactions the making and breaking of carbon–metal bonds are essential steps.

BIBLIOGRAPHY. E.G. ROCHOW, *Organometallic Chemistry* (1964), a highly readable, general introduction to the subject; G.E. COATES, M.L.H. GREEN, and K. WADE, *Organometallic Compounds*, 3rd ed., 2 vol. (1967–68), the most useful general introduction to the subject, with many illustrative examples. These two works treat both main group and transition-metal organometallics. More detailed books on these separate topics are: J.J. EISCH, *The Chemistry of Organometallic Compounds: The Main Group Elements* (1967); and R.B. KING, *Transition-Metal Organometallic Chemistry* (1969). A series of reviews of particular topics may be found in the serial publication *Advances in Organometallic Chemistry*, ed. by F.G.A. STONE and R. WEST (1964–). Volumes appear at approximately yearly intervals.

(M.I.B./F.G.A.S.)

Organs and Organ Systems, Animal

The bodies of multicellular animals are generally organized on the basis of tissues, organs, and organ systems. Constituting a tissue are many cells, usually similar in both structure and function and bound together by intercellular material. An organ is composed of various tissues, not necessarily similar, grouped together into a structural and functional unit. The bodies of some of the simple multicellular animals generally have few clearly distinct organs. Most larger, more advanced animals, however, characteristically have numerous distinct organs, which, in turn, may be organized into groups of organs that cooperate as a functional complex. Such a complex is called an organ system. Ten highly specialized organ systems make up the bodies of advanced animals; lower forms usually are not as complex in structure.

SPECIALIZED ORGAN SYSTEMS

Integumentary system. The integument, or outer covering of the body, is often called the skin in higher forms. The integument and its derivatives make up the integumentary system, which functions as a protective covering that resists harmful substances and organisms and prevents excessive water loss. All animals have an outer covering, the nature of which varies considerably. In protozoans, the simplest animals, it consists of a fine membrane or a firm, elastic film. In certain more advanced invertebrates, inhabiting an aquatic or moist environment, the integument consists of a single layer of cells; in others, the outer cells secrete a noncellular, waxy substance called a cuticle, which has various degrees of resistance. The cuticle may be hard enough to form a shell, as in the case of snails, or a protective exoskeleton, as in the case of insects. The periodic shedding of this outer layer is known as molting, or ecdysis. In vertebrates, the integument consists of an outer layer, the epidermis, and an underlying layer, the dermis. These layers vary in structure in different parts of the body and give rise to numerous integumentary derivatives (see INTEGUMENTARY SYSTEMS).

Types of integuments

Skeletal system. The framework of the vertebrate body is the skeleton. Vertebrates have a jointed internal skeleton (endoskeleton), composed of cartilage, bone, or both, that supports the rest of the body, serves as an attachment for muscles, and protects delicate vital organs. In a few forms, such as turtles and armadillos, skeletal elements may also be present in the dermis of the skin, making up the dermal skeleton. Among certain lower animals (*e.g.*, many protozoans, coelenterates, flatworms, and slugs), no skeletal system exists. As stated above, the cuticle derived from the integument may, in some animals, form an exoskeleton that provides support and protection and serves for muscle attachment. It may be composed of calcareous (containing calcium carbonate), siliceous (containing silica), or organic substances, and it is either rigid, as in corals and mollusks, or jointed, as in certain echinoderms and arthropods. Deposits of lime salts provide increased protection. Exoskeletons of arthropods must be shed periodically in order to permit growth (see SKELETAL SYSTEMS).

Muscular system. In all but the simplest organisms, specialized muscle cells perform the function of converting chemical energy into mechanical energy. The actual contractile machinery of muscle resides in long, threadlike filaments of molecular dimensions called myofilaments. The combination of muscle cells into large, organized groups—*i.e.*, tissues—that act together is called a muscle. Individual muscles may work together to form part of a muscle system. Two types of muscle tissue are found in animals: striated, or striped, muscle and smooth muscle. Nerves control the action of muscles.

Striated muscle fibres are found in a great variety of both vertebrates and invertebrates. Some coelenterates and rotifers and all arthropods have striated muscle, which is rare in worms and mollusks. All muscles of arthropods, in fact, are striated and thus function as both visceral and skeletal muscles. Striated muscle is found in the heart and in the entire skeletal, or voluntary, muscular system of vertebrates. Striated muscle is capable of rapid motions (*e.g.*, the flight muscles of an insect can contract and relax more than 100 times per second) or slow motion (*e.g.*, the heart muscle of some large mammals may contract only one or two times a second); most striated muscles function between the two extremes.

Smooth muscles vary widely in structure among animals; some have spiral or ribbon-shaped units (myofibrils); others have no recognizable structural units. Smooth muscle is found in all animal phyla. In most worms, the body and visceral movements are dependent on smooth muscle. Smooth muscle is found in the feet of many mollusks and is used when the animal crawls on a surface. Smooth muscle forms the greater part of the visceral musculature of vertebrates and forms coats around the alimentary canal and certain associated structures, such as bile ducts, as well as arteries and veins. In general, smooth muscle fibres act more slowly than do striated ones (see MUSCLE SYSTEMS).

Nervous system. Nervous tissue possesses the properties of irritability (the ability to respond to stimuli) and conductivity (the ability to conduct a signal). In Protozoa with minute hairlike structures called cilia, a specialized region (motorium) serves to coordinate and relay impulses from small filaments (fibrils) connecting the bases of the cilia. Many Protozoa also possess organelles; that is, cell parts analogous to organs specialized for stimulus reception, such as eyespots or sensory bristles. Sponges lack definite nerve cells, but all animal groups above the level of sponges have some form of nervous system. Coelenterates have a diffuse nerve net within the cell layer known as the epithelium but no central ganglion, or nerve cluster. Flatworms have two anterior ganglia with nerves to the head region and two separate nerve cords joined by cross connectives. Mollusks, annelid worms, and arthropods have paired anterior ganglia above and below the esophagus, joined by connecting nervous tissue. The echinoderms possess a radially arranged nervous system. The vertebrate nervous system is divided into two main parts: (1) the central nervous system, composed of brain and spinal cord, and (2) the peripheral nervous system, made up of cranial nerves and spinal nerves. Part of the peripheral system, consisting of portions of cranial nerves and spinal nerves, as well as outlying ganglia, controls involuntary functions, such as muscular contraction of the intestinal wall, and is called the autonomic system (see NERVES AND NERVOUS SYSTEMS).

Components of the vertebrate nervous system

Endocrine system. The endocrine system consists of a number of ductless glands that secrete chemical substances called hormones. These substances are carried by the circulating fluid to all parts of the body and exert highly specific effects on various tissues. In invertebrates, there is evidence of hormones in nematodes, mollusks, annelids, and arthropods. In crustaceans, a substance produced in the sinus gland influences the chromatophores —specialized pigment cells—so that the animal's body colour more closely resembles that of his environment. Endocrine glands in vertebrates include the pituitary, thyroid, parathyroid, adrenals, gonads, islets of Langerhans in the pancreas, and parts of the gastric and intestinal mucosa. In some mammals, the placenta, a structure through which the unborn animal is nourished, functions as an endocrine gland (see ENDOCRINE SYSTEMS; HORMONE).

Digestive system. The function of the digestive system is to procure and process, or metabolize, nutrients. The nutritional requirements and the basic processes of digestion are essentially similar in all animals; however, the body plan of animals varies so greatly that the structures involved are often different. *Amoeba*, a protozoan, lacks a permanent structure for digesting or ingesting food. Pseudopodia, temporary extensions of the body mass, are used to surround the food, which then becomes a bubble-like food vacuole in the body. *Paramecium*, however, another protozoan, has a permanent structure, the ciliated oral groove, into which food passes before forming a food vacuole. Sponges capture food by means of collar cells equipped with flagella, long whiplike structures.

Coelenterates have a central saclike digestive pocket, the gastrovascular cavity, which has only one opening and functions as both mouth and anus. In flatworms, the mouth opens into a muscular pharynx that leads to an extensively branched gastrovascular cavity. Animals above the level of coelenterates and flatworms have a complete digestive system—*i.e.*, one with two openings, a mouth and an anus. Food is passed in one direction through a tubular system that may have distinct sections, each specialized for a different function (see DIGESTION AND DIGESTIVE SYSTEMS).

Respiratory system. If nutrient materials are to be completely metabolized to carbon dioxide and water, oxygen is essential. One of the basic tasks of the organism, then, is the procurement of oxygen and the elimination of carbon dioxide. Gas exchange between a living cell and its environment always takes place by diffusion across a moist membrane. In protozoans, sponges, coelenterates, and flatworms, most of which are aquatic, each cell is in direct contact with the surrounding medium or is only a few cells removed from it. These animals, therefore, have not evolved special respiratory devices.

Mechanisms for gas exchange

With few exceptions, the respiratory systems of higher multicellular aquatic animals involve evaginated (outpocketed) exchange surfaces, usually known as gills. Gills are found in such diverse groups as annelids, mollusks, arthropods, and vertebrates. Most terrestrial animals have evolved invaginated (infolded) respiratory systems. These systems are of two principal types: lungs, found in land snails and the higher vertebrates; and tracheae, found in most terrestrial arthropods (see RESPIRATION AND RESPIRATORY SYSTEMS).

Circulatory system. Since every cell of an animal must obtain nutrients and oxygen and rid itself of carbon dioxide and nitrogen-containing wastes, some mechanism is needed for transporting these substances. In protozoans, such transport occurs by the flow, or streaming, of cytoplasm within the cell. In multicellular types such as sponges, coelenterates, flatworms, and roundworms, these substances are transported by simple diffusion between the external environment and internal organs. In higher metazoans (many-celled animals except sponges), most of the body cells are far removed from the external environment; a circulatory system for internal transport, therefore, has developed. Mollusks and annelids have an open circulatory system with dorsal (back-side) three-chambered hearts; the arthropods have a closed system with dorsal and ventral (belly-side) vessels and cross-connective vessels. In vertebrates, the highly specialized system is composed of integrated vascular networks for the separate transport of blood and lymph, which are carried directly to tissues (see CIRCULATION AND CIRCULATORY SYSTEMS).

Excretory system. Excretion is the process by which waste products of metabolism are removed from an organism. Excretory mechanisms also help to regulate water and salt balance. Special excretory structures are absent in many unicellular and simple multicellular animals, such as sponges and coelenterates. Some protozoans, however, have a special excretory organelle, or organ-like structure, the contractile vacuole. Flatworms possess a primitive, so-called flame-cell excretory system. This consists of two or more long, branching tubules, one end of which opens on the body surface through numerous excretory pores, the other ends of which are ciliated hollow bulbs—the flame cells. The arthropods possess Malpighian tubules, tiny pockets in the walls of the digestive tract. They collect wastes from the body fluids and pass them into the lower end of the gastrovascular cavity (hindgut) for excretion. In the earthworm, an annelid, each body segment has a pair of excretory organs called nephridia that consist of an open ciliated funnel, a coiled tubule, an enlarged bladder, and a pore (nephridiopore) leading to the outside. The excretory organs of vertebrates consist of paired kidneys, each with a separate duct (see EXCRETION AND EXCRETORY SYSTEMS).

Specialized excretory structures

Reproductive system. Reproduction is essential if a species is to survive. Asexual and sexual reproduction occur among animals. Asexual, or agamic, reproduction does not involve transfer of sex cells between individuals and occurs in members of the lower animal phyla. Some protozoans reproduce by dividing into "twin" cells, after which each develops into the original form. Budding, which occurs in coelenterates, is a process by which a new individual arises from an outgrowth, or bud, of an older animal. Some flatworms reproduce by fragmentation of their own bodies; each fragment then develops into a complete animal.

Sexual reproduction is the method of propagation common to all but the lower animal phyla. Gametes—*i.e.*, egg cells, or ova, and sperm cells, or spermatozoa—are formed in the reproductive organs, or gonads—ovaries in the female, testes in the male. If both male and female systems are in one individual, as in flatworms and earthworms, the animal is termed monoecious. In nematodes, arthropods, various other invertebrates, and practically all vertebrates, each individual is either male or female; the sexes thus are separate, and such animals are termed dioecious (see REPRODUCTIVE SYSTEMS, ANIMAL).

INTERRELATIONSHIPS BETWEEN ORGAN SYSTEMS

Functional interdependence. An understanding of the complexity of life functions and of the intricacy with which these functions are interwoven has gradually emerged from investigations of the various aspects of the biology of organisms. This complexity is not restricted to multicellular organisms, with their many cooperating cells, tissues, organs, and organ systems; it also characterizes unicellular organisms, which are known to be far from simple. Each cell of a living organism requires an environment that is dependent upon the interrelationships of cellular structure, metabolism, nutrient procurement, gas exchange, internal transport, regulation of levels of salts and other substances, and excretion. Homeostasis, or the state of equilibrium between the internal and external environments, is brought about by the functional interdependence of organelles, in unicellular animals, and of organ systems, in multicellular animals. This is demonstrated by the inability of one organ system to function without the assistance of another.

That living organisms function in an orderly fashion despite their immense complexity shows clearly that control mechanisms are at work. Coordination of the regulatory functions and all the myriad other functions of an organism depends upon two principal types of control mechanisms: chemical control mechanisms, which are found in all organisms, and nervous control mechanisms, which, in the strict sense, are found only in multicellular animals.

Coordination of regulatory functions

Since the multicellular organism is characterized by a division of labour among its parts, it is not unexpected that evolution has led to specialization of certain cells or tissues as producers of chemical controls, usually hormones. These substances often have important control functions in parts of the body far removed from the sites of synthesis and are transported in higher animals from sites of production to sites of action through the blood circulatory system. Chemical controls are produced by specialized tissues or organs, exert highly specific effects on other tissues of the body, and are effective in very low concentrations. Hormonal control, however, is a relatively slow process, there being an appreciable delay between the release of the hormone and its arrival at the target organ. Slow chemical control is compatible with normal activity when instantaneous response is not needed, as in control of digestion, salt and water balance, metabolism, and growth. When rapid response is required, however, nervous control is essential. A nerve impulse can travel several hundred feet per second, thus reducing the interval between stimulus and response to milliseconds.

In vertebrates, the autonomic nervous system plays a large role in the maintenance of homeostasis and provides a fine control of the visceral, or internal, functions of the body. The autonomic system is generally involuntary and acts on the internal effectors such as smooth muscle, cardiac muscle, exocrine glands (glands with ducts), and endocrine glands (glands without ducts). The

autonomic nervous system is the structural pathway linking the control centres in the brain with the internal organs. Physiological control of visceral functions includes sensors and nerve pathways as well as central control.

Specialized nerve cells, found in both external and internal receptor organs, are the organism's principal means of obtaining information. External sense organs are those stimulated by environmental changes. They include organs for the senses of sight, hearing, smell, taste, touch, pressure, temperature, and pain. Internal sense organs, affected by stimuli arising within the body, include proprioceptors (sensors of body movement and position), deep pain receptors, interoceptors—for such internal sensations as hunger, thirst, and nausea—and sensors of salt balance and carbon dioxide level in the blood plasma. Receptor organs are capable of responding to stimuli by initiating impulses that are transmitted by nerve fibres to the central nervous system for interpretation as sensations. The central nervous system may then send appropriate signals to muscles, glands, or other effector systems for response to the stimuli.

Feedback mechanisms. Since homeostatic mechanisms act to minimize the difference between the actual and optimal response of a system, they may be considered as biological examples of negative feedback control. In systems of this type, the level of the controlled variable is sensed, and action is taken to oppose change from the desired level. If the response increases, a negative, or inhibitory, signal is fed back to an effector mechanism so that the subsequent response is reduced. On the other hand, a decrease in response elicits a subsequent increase.

The interaction between the anterior pituitary gland and the thyroid gland is an example of negative feedback control. Thyrotropic hormone, released by the pituitary when the concentration of thyroxine in the blood is low, stimulates increased production of the thyroxine by the thyroid. When the rising concentration of thyroxine in the blood reaches a certain level, the secretion of more thyrotropic hormone by the pituitary is inhibited. There is thus a feedback of information from the thyroid to the pituitary. The pituitary exerts control over the thyroid, and the thyroid in turn exerts some control over the pituitary. A similar relationship exists between the pituitary and the other endocrine glands.

The nervous control of the rate of the heartbeat provides a good example of autonomic nervous system feedback control. When the heart is engorged with blood, stretch receptors in the wall of the upper right portion (right atrium) send impulses to the accelerating centre in the brain. The impulses from the stretch receptors stimulate the accelerating centre to send, via the autonomic nervous system, excitatory impulses to the S-A (sino-atrial) node, or "pacemaker," thereby causing the heart to beat faster. As the blood pressure rises, however, pressure receptors in the wall of the aorta (the blood vessel that conveys blood to the body) begin sending impulses to the inhibiting centre in the brain, stimulating it to send inhibitory impulses, via the autonomic nervous system, to the S-A node, with the effect of slowing the heart.

DEVELOPMENT OF ORGAN SYSTEMS

Germ-layer concept

The embryonic origins of individual systems. During embryonic development the cells making up the so-called germ layers—ectoderm, endoderm, and mesoderm—are said to be undifferentiated; *i.e.*, they do not possess distinctive or individual characteristics. Further development of an embryo entails, among other things, a differentiation or specialization of various groups of cells to form the several types of tissues and organs that make up an organism. Growth of the embryo and differentiation of organs from germ layers begin even before the germ layers are completely formed. The entire process of development is gradual, and, although the various phases occur in a step-by-step order, they are not in themselves distinct but are merged imperceptibly with each other. In all vertebrates the different germ layers and structures arising from them are considered to be homologous, or similar in origin. The germ-layer concept is of importance chiefly because it furnishes a convenient method of classifying organs according to their embryonic derivation.

The ectoderm gives rise to three main structural groups and their derivatives: (1) the epidermis of the skin and its derivatives, including the skin glands, hair, feathers, nails, claws, hoofs, horns, epidermal scales, and the coverings of external gills; (2) the lining of the mouth and related structures, including enamel of teeth, glands of the mouth, covering of the tongue and lips, and anterior and intermediate lobes of the pituitary gland; (3) the nervous system, consisting of the brain and spinal cord, cranial and spinal nerves, autonomic portion of the peripheral nervous system, sensory parts of all sense organs, inner region (medulla) of the adrenal gland, infundibulum (structure connecting the pituitary gland to the brain), and the posterior lobe of the pituitary gland.

The endoderm develops into the epithelial lining of the following: (1) the alimentary canal, which includes pharynx, esophagus, stomach, intestine, liver, pancreas, and most of the cloaca (a cavity at the posterior end of the body); (2) the pharyngeal derivatives, consisting of the larynx, trachea, lungs, gills of the internal type, middle ear, eustachian tube, tonsils, thyroid, parathyroids, and thymus; (3) miscellaneous structures, including the urinary bladder, its canal to the outside of the body (urethra), and two embryonic structures—allantois and yolk sac.

The mesoderm differentiates into (1) the muscles, including smooth, skeletal, and cardiac; (2) the skeleton, composed of cartilage, bone, and other connective tissue; (3) the excretory organs, including the kidneys and their ducts; (4) the reproductive organs; (5) the circulatory system, composed of the heart, blood vessels, blood, spleen, lymphatics, and blood-forming tissues; and (6) miscellaneous tissues, including dentine of teeth, dermis of skin, outer region (cortex) of the adrenal glands, lining of body cavities, mesenteries (tissues supporting the viscera), and portions of the eye.

Appearance of organ systems during development. Organ systems of the vertebrate body make their appearances at precise stages of embryonic development. The human ovum is normally fertilized in the upper fallopian tube, which leads from the ovary to the uterus. At the end of the first week, it becomes implanted in the uterine wall.

Summary of embryonic development

After two weeks the embryo consists of a flat disk, in the centre of which are the primitive streak, which later becomes the mesoderm, and the first rudiments of the nervous system, the neural plate, and the neural groove. A head process and rudiments of the heart may be present. No endodermal derivatives are yet present.

By the end of the third week the neural groove is complete and closes to form the neural tube; the optic vesicles (rudimentary eyes), auditory placodes (rudimentary ears), and ganglia are present; and the oral membrane may rupture to form the mouth. In addition, the separation of the body from the yolk sac produces the foregut and hindgut, antecedents of the pharynx and lower parts of the digestive tract. The visceral pouches and visceral arches—which both develop into parts of the jaw and throat—and the lung and liver pouches appear. Blood vessels containing blood cells develop and connect with an S-shaped heart, in which pulsations begin. The pronephros, an early stage of the kidney, and its ducts grow posteriorly, or downward.

After four weeks the body has a C shape. The neural tube has entirely closed, the three primary brain sections have formed, and the spinal and cranial nerves are developing. In the endoderm, paired lung buds grow posteriorly; the stomach, liver, pancreas, and intestines are defined. Paired limb buds and primitive vertebral organization are apparent. Abundant blood is forming and circulating, and the heart is tubular. The pronephros degenerates and is replaced by subsequent stages of the embryonic kidney, the mesonephros and the metanephros.

At the end of the fifth week the embryo has a temporary tail. The face is assuming a characteristic appearance with jaws, and the brain has five sections. The trunk and

appendage muscles develop, and bone-forming centres arise. The circulatory system is extensively developed, and the heart begins its final divisions.

After six weeks of development, the heart is the four-chambered organ of the adult; undifferentiated gonad primordia are prominent. By the seventh week, the eye and pituitary gland are well developed, as are the lungs and the components of the digestive system. The body muscles are becoming organized; cartilage formation is extensive; and the jaws, vertebrae, and ribs are becoming bony.

After eight weeks, the embryo, now termed a fetus, has a recognizable form. Muscles of the body are differentiated and innervated to allow movement. The blood-vascular system is complete; the mesonephros degenerates as the metanephros grows; and sex differentiation begins. During the ninth week nails and hair follicles form, and the metanephric kidney becomes functional. By the tenth week, the brain is essentially developed, the lungs are almost complete, and the smooth muscle of the entire gut is organized.

By the 14th week, the head, until now large with respect to the body, and the body approach normal proportions. Most of the bones are present in some degree, allowing spontaneous movements. After 18 weeks, the body proportions approach those of the newborn. The cerebral hemispheres of the brain become convoluted and creased; the retina of the eye becomes sensitive to light, and the sex organs have developed.

EVOLUTION OF ORGAN SYSTEMS

Organelles. Unicellular animals are found only among the Protozoa; all other animals are multicellular. The individual protozoan, however, should not be regarded as equivalent to a single cell of a more complex animal but as a complete organism with the same properties and characteristics of a multicellular animal. Protozoans lack tissues and organs, since such parts are defined as aggregations of differentiated cells, but they do have functionally equivalent subcellular structures called organelles, which are analogous to multicellular organs. Some protozoans digest food in food vacuoles. No special organelles exist for gas exchange, the general cell membrane serving as the gas exchange surface. Many freshwater protozoans possess contractile vacuoles, which function primarily in the regulation of osmotic pressure, thus controlling the internal concentration of salts and water; nitrogen-containing wastes may also be expelled through these vacuoles. Locomotion is performed by formation of pseudopodia or with beating cilia or flagella. Although an individual protozoan has many cilia, cilial action is coordinated by a system of fibrils connecting so-called basal bodies of the cilia.

Multicellularity. Members of the animal kingdom possessing a multicellular plan of structure constitute the subkingdom Metazoa. The body structure of a typical metazoan involves more than just a multicellular condition, however; the specialization of cells for different functions produces an interdependency of cells. Cell specialization has in turn led to the development of tissues consisting of similar cells organized in sheets and layers. In lower metazoans, development of tissues is relatively primitive; in higher groups tissues have become organized to form organs.

Nature of ancestral forms

Zoologists agree that metazoans evolved from unicellular organisms. There are three theories concerning the nature of the ancestral form and the mode of origin. The first is the syncytial theory, which holds that multicellular animals arose from a primitive group of unicellular, ciliated animals having more than one nucleus. The ancestral metazoan, at first syncytial (*i.e.*, with more than one nucleus within a single mass of cytoplasm) in structure, later became compartmented or cellularized by the development of cell membranes, thus producing a typical multicellular condition. Because many ciliates tend toward bilateral symmetry—that is, similar left and right sides comprise the body—proponents of the syncytial theory maintain that the ancestral metazoan was bilaterally symmetrical and gave rise to the acoel flatworms (lacking a digestive cavity with definite walls). The fact that these flatworms, thought to be the most primitive living metazoans, are about the same size as the ciliates, have bilateral symmetry and cilia, and tend towards a syncytial condition is regarded as evidence in support of the syncytial theory. That it requires acoels to be the most primitive living metazoans is regarded as an objection to the theory; a ciliate ancestry, moreover, does not explain the general occurrence of flagellated sperm in metazoans. Opponents of the syncytial theory further point out that the developmental patterns among the lower metazoans are not comparable.

Colonial theory

The second theory of the origin of metazoans is the colonial theory, which states that the origin of multicellular animals is from a spherical, hollow, colonial, flagellated organism. Proponents of the colonial theory regard the following as evidence: (1) flagellated body cells commonly occur among lower metazoans; (2) sperm and eggs as clearly definable entities have evolved in the phytoflagellates, organisms with flagella and the ability to photosynthesize; (3) in some forms of flagellates that live in colonies, a differentiation between reproductive cells and somatic (nonreproductive) cells has occurred. The colonial theory holds that through the migration of cells into the interior of the colony, the originally hollow sphere became a solid, ovoid mass, with similar parts arranged radially around a central axis (radially symmetrical) and with the exterior cells flagellated. Since this hypothetical organism is very similar to an immature coelenterate form called a planuloid larva, it is called the planuloid ancestor of the lower metazoans. The bilateral symmetry of the flatworms would then represent a later modification in symmetry. The principal weakness of the colonial theory is that the extant colonial phytoflagellates are plantlike and possess cellulose walls with chlorophyll. An alternative explanation is that metazoans arose from some group of extinct flagellates that did not have these typical plant features.

The third theory of the origin of metazoans, the theory of polyphyletic origin, proposes that the sponges and coelenterates evolved from colonial flagellates, ctenophores, and flatworms by way of the ciliates.

The colonial theory, despite certain weaknesses, appears to be the most compatible with evidence relating to the subsequent evolution of metazoans. The problem of the planuloid ancestor shifting from a radial to a bilateral symmetry is not difficult if it is assumed that the ancestral planuloid stock lived on the ocean bottom and, as a result, developed a creeping mode of movement over rocks. This type of movement would lead to a differentiation between dorsal and ventral (top and bottom) surfaces; bilateral symmetry would result. The free-living flatworms demonstrate the transition from a bilateral planuloid ancestor to a more complex form.

A hypothetical ancestral flatworm may be described from the various primitive features of living flatworms. A marine animal, it was dorsal-ventrally flattened, with a single layer of ciliated epidermal cells, the bases of which contained contractile extensions that formed a muscle layer. An otocyst, or hearing organ, and a light-sensitive pigment spot may have been present. The otocyst, near the front end, may have been covered by a delicate network of nerve cells. Possibly a midventral mouth opened into a syncytial network of nutritive cells and a meshwork of reproductive cells.

In subsequent evolution of the flatworms, a number of fundamental changes appear that foreshadow the structure exhibited by most higher bilateral animals: (1) the separation of the contractile function from the epidermis to form a distinct muscle layer; (2) the reorganization of a nerve net contiguous with the body surface into a series of deeper, radially arranged, longitudinal nerve cords; (3) the concentration of nervous tissue around the statocyst (an organ of equilibrium) to form a brain; (4) the development of a cup-shaped eye with pigment cells from flattened pigment spots; (5) the formation of a rudimentary digestive system with inturning of the epidermis around the mouth to form a short pharynx; (6) the formation from mesenchyme of a rudimentary reproductive

system with structures for conducting, transmitting, and receiving sperm.

BIBLIOGRAPHY. C.K. WEICHERT, *Anatomy of the Chordates*, 4th ed. (1970); and R.D. BARNES, *Invertebrate Zoology*, 2nd ed. (1968), are detailed references describing animal organ systems and their evolution. The former deals primarily with vertebrates, the latter with invertebrates. See also W.T. KEETON, *Biological Science* (1967), a thorough biology textbook, with several chapters concerning the interrelationships of organ systems; E.E. SELKURT (ed.), *Physiology*, 2nd ed. (1966), a complete account of mammalian physiology; and R. RUGH, *Vertebrate Embryology* (1964), which presents the normal sequence of events of vertebrate development.

(C.K.W./K.W.K./M.S.K)

Organs and Organ Systems, Plant

Plant organs include such structures as individual leaves, stems, roots, flowers (or parts of flowers, such as the male or female reproductive structures), fruits, and other units of vegetation depending upon the level of organization chosen to delimit them. Organ systems are groups of mutually functioning organs, such as the aerial shoot system or the entire underground complex of roots, the root system of a plant. Such structures and interrelated systems are the means by which plants exploit the environment; they are the products of successful evolutionary advances over geological time.

The organs and organ systems of plants are most essential to the well-being of man. In fact, man is utterly dependent upon plants for the essentials of his existence. The food of man comes from plants, either directly or via the flesh of animals that feed upon plants. Much of man's clothing and shelter are made from plant products. Paper is made from wood, and many drugs used to cure man's diseases are prepared from plant materials. The coal used to generate electric power is simply fossilized plants that lived millions of years ago. Plants generate oxygen, modify the climate, and aid immensely in the control of erosion and floods. Many kinds of trees, shrubs, and herbs are used as ornamentals in gardens and in landscaping. Plants have exerted enormous influence upon the social and economic life of peoples from earliest times.

IMPORTANCE TO MAN

The shoot system, or portions of it, is used in numerous ways. The leaves of cabbage, lettuce, and watercress are eaten in green salads; those of spinach, kale, collards, and turnips are consumed as a cooked vegetable. Spices such as basil, bay, sage, savory, tarragon, and thyme are leaves, as are spearmint and peppermint used in flavouring drinks, chewing gum, and candies. Sisal fibre used in making twine and rope comes from leaves. Tea, the favourite beverage of millions of people, and tobacco, a popular stimulant, are prepared from leaves. The pain-killer cocaine is made from coca leaves. Celery and rhubarb are leaf petioles (stalks).

Uses of stems

Stems are also utilized in many ways. Gourmets relish the young, juicy, succulent shoots of asparagus; many people enjoy the large, spherical, juicy stems of kohlrabi. Both the spice cinnamon and the medicine quinine come from the bark of certain woody plants. The narcotic peyote, used by certain American Indians in religious ceremonies, comes from the fleshy stem of a small cactus. Wood from trees is used for building material, papermaking, fuel, and furniture. Phloem fibres of flax are used to weave linen cloth; those of hemp and jute are woven into ropes and bags. Both stems and wood strips are made into baskets and matting. The large, hollow stems of bamboo are extensively used in constructing furniture, fishing rods, and many other articles. A favourite vegetable of many people is Brussels sprouts, the axillary (stem) buds of a relative of cabbage.

Inflorescences (flower clusters) consisting of immature flowers and flower stalks comprise the vegetables cauliflower, artichoke, and broccoli. The spices known as cloves and capers are the unopened flower buds of certain flowering plants, and saffron, a flavouring material, is made from flower stigmas and style tips (the upper extremities of the female flower parts). Marijuana is the dried upper portion of female plants; the resinous hairs on the flowers and the flower stalks contain most of the active narcotic principle. The characteristic bitter flavour of beer is caused by the flower buds of the hop plant used during the brewing process.

Underground stems have many food uses. Potato tubers feed millions of people in many parts of the world. In Southeast Asia and Polynesia, the tubers of taro and dasheen are an important starchy food. Onion bulbs also are consumed by millions of people.

Throughout the world, the fleshy storage roots of numerous plants are eaten. Sugar beets constitute an important source of sugar, and radishes add variety to meals. The roots of beets, turnips, parsnips, and rutabagas are cooked and eaten as a starchy food by people in temperate regions. The roots of yams, cassavas, and sweet potatoes provide starch for many people, especially in tropical areas. Gourmets relish the flavours imparted to foods by ginger, turmeric, and horseradish.

Products from fruits and seeds

The reproductive system of plants, including fruits and seeds, provides numerous products of great economic importance to mankind. The entire inflorescence of broomcorn is used in making brooms. The fruits of many species throughout the world are eaten raw or cooked, including such plants as breadfruit, cucumber, eggplant, pumpkins, squash, tomato, apple, pear, watermelon, grape, cranberry, strawberry, orange, date, olive, pineapple, and many others. Spices such as allspice and pepper are fruits, as are the "seeds" (actually one-seeded fruits) of caraway, celery, dill, coriander, and fennel. Two narcotics, opium and heroin, are made from the juice of the unripe ovaries of poppy plants.

There are a multitude of ways in which seeds are utilized. Such beverages as coffee and chocolate are enjoyed by many people. Cereal grains such as rice, wheat, and corn (maize) are staples in the diet of hundreds of millions of people throughout the world. Other cereals such as oats, barley, rye, and millet are important food plants in various regions. Wild rice and buckwheat are delights fancied by many gourmets. Other dietary mainstays of millions of people are the legumes, including beans, peas, soybeans, peanuts, and lentils. Nuts such as pecan, hickory, walnut, pine, almond, and chestnut provide a variable diet for many people. Betel nuts and kola nuts are chewed and enjoyed as mild narcotics.

A fibre obtained from the seed coat of the cotton plant has been used in cloth making for centuries.

DEVELOPMENT OF PLANT ORGANS

Origin of stem, leaf, and root

Very early in its development, the body of a plant exhibits a polarized (*i.e.*, directional with definite shoot and root ends) differentiation into an axis composed of three organs: stem, leaf, and root. This differentiation occurs as the fertilized egg develops into the embryo and, in the seed plants, is evident long before the seed is mature (see DEVELOPMENT, PLANT). Although each of these organs differs significantly in structure and function, all are similar in important ways. They have a common origin in the young embryo. The tissues of each organ merge imperceptibly, an arrangement facilitating the performance of joint functions—*e.g.*, the movement of water, dissolved minerals, and food longitudinally from one part of the plant body to another (see TISSUES AND FLUIDS, PLANT). In addition, the stem and root actually constitute one continuous structure, exhibiting basically a cylindrical or rodlike form. These two organs also possess similarities in their method of growth from the meristems, the regions of cell division in shoot and root tips (see DEVELOPMENT, PLANT). During plant evolution, the stem, leaf, and root have arisen from a common structure.

The stem. *Description.* The stem is normally a cylindrical, rodlike axis that supports the food-producing leaves and connects them with the anchoring and absorbing roots. The stem also produces the cones, flowers, fruits, and other reproductive structures. In addition, the growth pattern and structure of the stem determine the habit or form of the plant.

At its apex, the stem possesses a bud, a juvenile and incompletely developed structure. The bud contains the

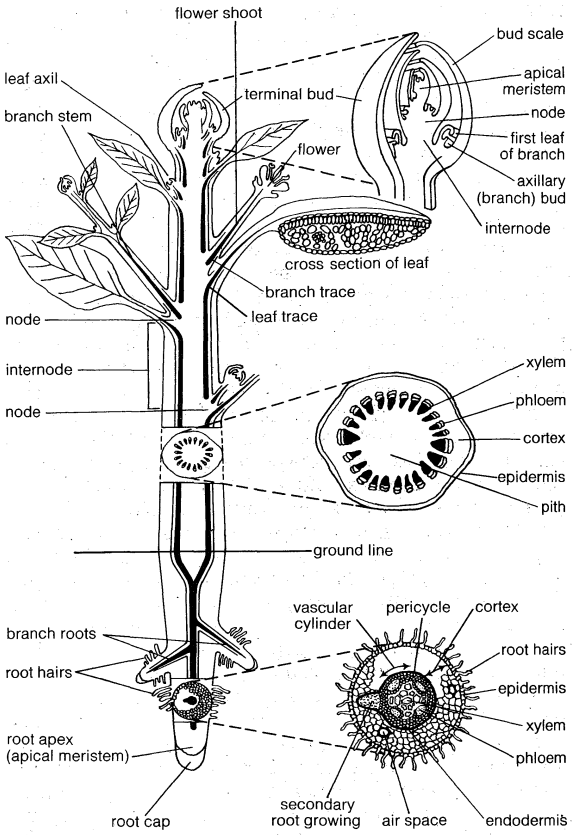


Figure 1: Principal organs and tissues of the body of a seed plant.

(Left, centre right) from W.W. Robbins, T.E. Weir, and C.R. Stocking, *Botany, An Introduction to Plant Science*, 3rd ed. (1964); John Wiley & Sons, Inc. (Top right) *Biology*, 3rd ed., by Willis H. Johnson, Richard A. Laubengayer, Louis E. DeLanney and Thomas A. Cole; copyright © 1966 by Holt, Rinehart and Winston, Inc.; copyright © 1956 and 1961 by Holt, Rinehart and Winston, Inc., under the title *General Biology*; reprinted by permission of Holt, Rinehart and Winston, Inc. (Bottom right) Kimball, *Biology*, 2nd ed. (1968); Addison-Wesley Publishing Company, Reading, Mass.

microscopically small apical meristem, which gives rise by cell division to new cells and tissues, thereby increasing the height of the stem. During seasons of the year unfavourable to plant growth (winter or dry seasons), the delicate meristem tissue is often protected from desiccation by small modified leaves known as bud scales. During the growing season, the apical meristem is usually protected by a covering of young leaves. Some buds give rise only to new stem and leaves, while others produce only flowers; still others develop leaves and flowers. When the bud develops into a new stem and leaves, the bud scales fall away, leaving several closely crowded bud scale scars on the stem. Embryonic leaves formed by the apical meristem develop at regions of the stem called nodes; the intervening portion of the stem between two nodes is the internode. Depending upon the species, the internodes may be long and distinct or very short and indistinct. In trees and shrubs, the division of the stem into nodes and internodes eventually becomes obscured because of growth in thickness of the stem caused by the cambium, a layer of dividing cells just under the bark. Eventually the leaves are shed from the plant, leaving leaf scars on the stem. Within each leaf scar may be seen one or more vascular-bundle scars produced at the location of the water- and food-conducting tissues (vascular bundles) that enter the leaf from the stem. Each node on the stem, therefore, may have a variety of features: there may be one or more leaves, each usually with a bud in the leaf axil, the upper angle between the leafstalk (petiole) and the stem. After the leaves have been shed, there is left on the stem a leaf scar containing vascular-bundle scars, and a dormant bud is left in the axil.

Structure and contents of buds

Tissue organization. After being produced by the apical meristem, the cells of the young stem differentiate gradually into the three main tissues: protective (epidermis, the outer "skin"), fundamental or ground (cortex, pith), and conducting or vascular (phloem and xylem). There is considerable diversity among plants in the patterns exhibited by these tissues.

Arrangement of vascular tissue

Among the lycopods (club mosses), ferns, and some aquatic angiosperms, the centre of the stem is occupied by a solid rod of xylem, specialized water-conducting tissues made of hollow cells. In most plants, however, the vascular tissue is arranged in a cylinder or as a complex of interconnected strands arranged either as a cylinder or scattered in bundles throughout nearly all of the stem. The region between the vascular tissue and the epidermis is occupied by the cortex, one of the fundamental tissues. The interior of stems in which the vascular tissue is arranged as a cylinder is filled with the pith.

Each strand of the primary (*i.e.*, derived from the apical meristem) vascular tissue is a vascular bundle. Within each vascular bundle, the xylem and phloem are arranged in characteristic patterns. In the collateral pattern, the phloem lies on only one side of the xylem, usually toward the stem exterior, whereas in bicollateral vascular bundles, the phloem is located on both sides of the xylem. In concentric bundles, one type of vascular tissue completely encloses the other, with the phloem surrounding the xylem in some species and the xylem surrounding the phloem in others. All of the vascular bundles and the associated fundamental tissues are interpreted by some plant anatomists as constituting a single unit, the stele.

According to the stele concept, the simplest and most primitive type of stele is the protostele, consisting of a solid column of vascular tissue without a pith. The protostele varies considerably, depending upon the species. Among the earliest known vascular plants and in the stems of young ferns, the protostele consists of a central core of xylem surrounded by a cylinder of phloem. In more complex protosteles, the contour of the xylem core is star shaped (as in many angiosperm roots), or the phloem and xylem may be in the form of separate strands of tissue (as in lycopods).

A type of stele that is more complex and more advanced evolutionarily is the tubular stele, or siphonostele, consisting of a cylinder of vascular tissue enclosing a central column of pith. The siphonostele, with several modifications, is present in most species of vascular plants. In some ferns the xylem cylinder is bounded on both sides by phloem. The xylem cylinder may lack leaf or branch gaps—interruptions evident in cross sections—caused by portions of the vascular cylinder leaving the stem to enter leafstalks. If leaf and branch gaps are present, they may be so large as to dissect the xylem cylinder into a netlike form; the result is a modified siphonostele called a dictyostele. In most plants, however, the phloem is located on the outer side of the xylem cylinder. Dissection of such a stele by the leaf and branch gaps results in a modification termed the eustele. The eustele is present in many seed plants. In monocotyledons, in which the stele consists of a system of scattered vascular strands, the result is the atactostele.

Leaf traces

At each node, one to several vascular bundles depart from the stem vascular system and enter the leaf; each such bundle is a leaf trace. The xylem and phloem of the leaf traces are continuous with the xylem and phloem in the stem. In the region of the vascular cylinder, immediately above the point of departure of the leaf trace out into the leaf, is a gap in the vascular tissue. Within this gap, parenchyma (cells of an undifferentiated, roughly globular form, usually with thin walls) tissue develops instead of vascular tissue, producing a region known as a leaf gap. Leaf gaps are often very conspicuous, especially in the stems of herbaceous plants from which the cortex, phloem, and pith have rotted away. The xylem tissue, more resistant to decay, remains, and the leaf gaps can be seen as holes through the xylem cylinder.

Axillary buds are also served by strands of vascular tissue departing from the vascular system of the stem. The bundles leading to the buds are called bud traces. Just as with the leaf trace, immediately above each bud trace is a region of parenchyma tissue, the bud gap. Bud traces are conspicuous parts of the stem anatomy and also appear as large holes in the vascular cylinder of rotted herbaceous stems.

All plants increase the height of their stems through the production of new cells and tissues, called primary tissues, by the apical meristem. In the young plant, only a small amount of conducting and strengthening tissue is present, the stem and roots being thin and fairly delicate. Many species, especially among the dicotyledons (broad-leaved flowering plants such as roses, woody trees and shrubs, members of the pea family, and others) and gymnosperms, possess the ability to produce additional conductive, fundamental, and protective tissues (secondary tissues) by means of lateral meristems—the vascular cambium and the cork cambium (see TISSUES AND FLUIDS, PLANT). The great bulk of a tree or shrub, therefore, consists of secondary tissues, especially of xylem, produced by the lateral meristems. The xylem formed by the vascular cambium accumulates year after year, with each year's accumulation forming a layer, the annual, or growth, ring. The phloem produced by the vascular cambium usually functions for only one or two seasons and is progressively crushed and destroyed as the new xylem presses outward. The tree does not become devoid of phloem, however, because the cambium is constantly producing new phloem cells during each growing season.

Production of annual growth rings

The vascular cambium in the stems of most gymnosperms and dicotyledons originates as a cylinder between the primary xylem and phloem. As long as the plant is alive, the vascular cambium occupies this same relative position, producing secondary xylem toward the stem centre and secondary phloem toward the stem surface. Secondary growth by the vascular cambium modifies the primary body of the plant in various ways. The secondary tissues usually cover the primary xylem and the pith, and the cells in these tissues eventually die. The primary phloem soon becomes nonfunctional and is destroyed as the stem increases in diameter. The cortex and epidermis are eventually replaced by the development of periderm (bark).

Periderm usually forms in the stems (and roots) of trees and shrubs and may be produced by the older parts of herbaceous dicotyledons. The outer portion of the periderm, the outer bark or cork, may be relatively thin, giving the trunk a smooth surface, or it may be thick, producing a cracked and fissured surface. The bark may develop as wings, projecting from the surface, as large scales or plates, or as long strips. Clearly defined, wartlike lenticels (pores) occur here and there in the cork. Each lenticel is a mass of loosely arranged cells protruding above the stem surface through a crack in the periderm. Each lenticel usually forms just beneath a stoma-guard-cell apparatus in the epidermis. A stoma, or stomate, is a tiny opening or pore in a leaf or stem surface, usually flanked by two liplike cells called guard cells. Like the stomata, lenticels are believed to function in the movement of gases into and out of the stem.

Types of stems. *Life-span categories.* The type of stem is related to the life-span of the plant. Many long-lived plants increase the rigidity of their stems (and often their roots also) by the production in the vascular cambium of hard, mechanically strong xylem tissue; such plants are called either trees or shrubs. Trees are characterized by a stem that grows to become a conspicuous trunk; shrubs usually have several stems, which arise near the ground. Many species of plants are relatively short-lived and consequently possess weak stems, with only a little woody tissue at the base or without any woody tissue at all. Such herbaceous plants are of diverse types. In shrubby herbs, the upper portions of the stems are herbaceous, and the lower parts are woody; the herbaceous portions die during the winter. Perennial herbs possess a very short woody crown stem, which may continue to produce new herbaceous stems for many years. Biennial herbs live only two years, the lower portion of the stem persisting through the winter and giving rise the next spring to an erect stem. After producing flowers and seeds the second year, the entire plant dies. Finally, annual herbs live for only a single growing season, completing their entire life cycle during this brief period.

Perennial, biennial, and annual plants

Woody stems. When the trunk of a tree is sawed in half, the bark and the wood are easily distinguished. The bark, usually much darker in colour than the wood, can be separated fairly easily from the wood, especially in the spring and early summer, because the cells of the vascular cambium have thin delicate walls that are readily broken. Microscopically, the bark can be seen to be composed of all of the tissues outside the vascular cambium, including the phloem and the periderm.

To the interior of the vascular cambium lies the wood or secondary xylem. A conspicuous feature of the wood is the concentric layers, the annual increments, or growth rings, especially in trees of temperate climates. The outer portion of the trunk may be a light-coloured zone of sapwood; the inner part may be differentiated into a generally darker region of heartwood. The sapwood is the younger tissue, and it functions in the conduction of water and dissolved minerals (*i.e.*, "sap") and in the storage of food. As the tree ages, the sapwood is gradually transformed into heartwood. During this process, many by-products of cell metabolism (*e.g.*, oils, tannins, resins, pigments, phenols) are transported away from the living cells of the phloem, cambium, and xylem in the direction of the inner sapwood, where they are stored in the cells. The formation of heartwood in trees is the result of such deposition of the plant's waste products. As the waste materials accumulate in the cells of the inner sapwood, a toxic level is finally reached, the xylem parenchyma cells die, and another cylinder or ring of heartwood is produced. In this manner, the sapwood–heartwood boundary moves outward as the diameter of the tree increases.

Herbaceous stems. The stem anatomy of herbaceous plants varies considerably. Some perennial herbs possess small amounts of secondary growth, especially in the lower portion of the stem and in the root. The stems of many biennial and annual herbs may lack secondary growth entirely because of the absence of a vascular cambium. Externally, the stems of these herbs resemble closely those of the young twigs of a woody plant, except that typical winter buds are not formed. Apical growth, nodes, internodes, and lateral branches are usually present in herbaceous stems. Internally, the vascular bundles are usually arranged in a single ring. This basic pattern of stem anatomy may be variously modified, especially in those stems functioning as food- and water-storing organs. In such stems, extensive parenchyma is often present.

Plants without secondary growth

The stems of most monocotyledons (flowering plants with parallel-veined leaves, such as grasses, orchids, and palms) lack a vascular cambium and do not have secondary growth. Palms and other treelike monocotyledons do, however, possess a means of thickening the trunk through the activity of a primary thickening meristem located beneath the embryonic leaves at the stem apex. In herbaceous monocotyledons such as the grasses, the stem-tissue arrangement features numerous vascular bundles scattered throughout the stem, as in corn (maize), or in two circles near the periphery, as in wheat. Just beneath the epidermis is (usually) a continuous cylinder of sclerenchyma (thick-walled cells) that functions in support of the stem. In some grasses, the stem centre is occupied by a pith, which may break down in the internodes during growth, leaving a hollow pith cavity. At the base of each internode in grasses is a region of cell division called the intercalary meristem, which adds new cells to the internode.

The leaf. The leaf is the principal photosynthetic (food-producing) organ of the plant. Although the leaf normally differs significantly, especially in form, from the stem, both of these organs are actually part of the same unit, the shoot. Both the stem and the leaf arise in the bud from the apical meristem. When fully developed, the tissues of the stem merge with those of the leaf. The leaf possesses basically the same tissues as the stem, with an epidermis forming the outermost layer, vascular tissues arranged in veins, and photosynthetic tissue occupying the same region as does the cortex in the stem.

Form and structure. The leaf is highly variable in its form and internal structure, and, accordingly, several kinds of leaves can be distinguished. The most familiar type is the foliage leaf, usually consisting of a thin, flat-

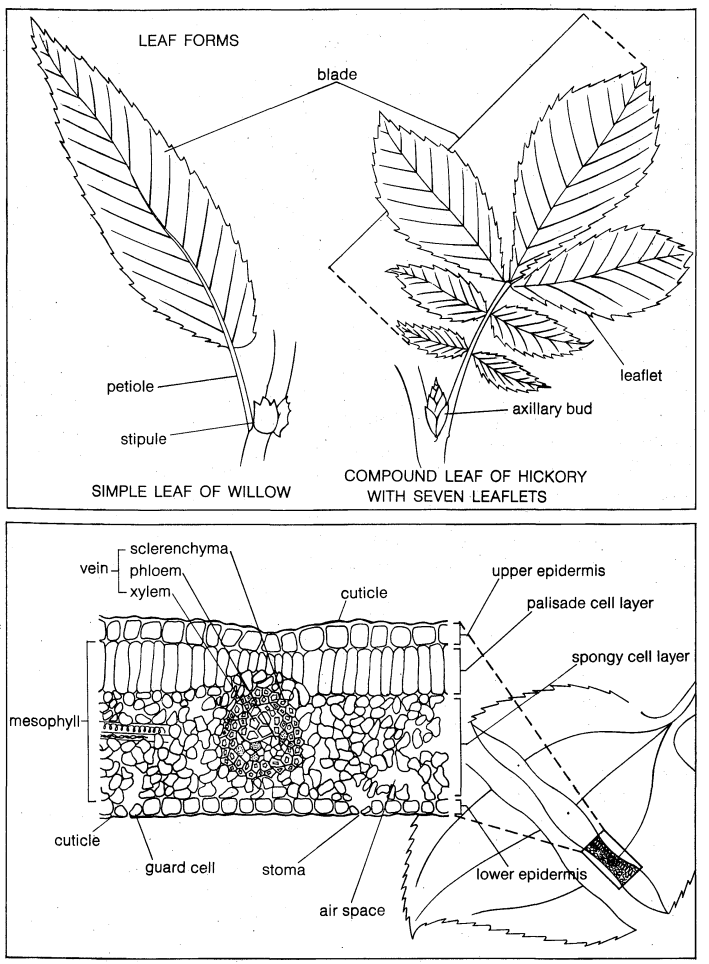


Figure 2: (Top) Leaf forms. (Bottom) Internal organization of a leaf shown in cross section.
From (top) Adams, Baker, and Allen, *The Study of Botany* (1970), Addison-Wesley, Reading, Massachusetts; (bottom) Kimball, *Biology*, 2nd ed. (1968), Addison-Wesley

tened, green blade, which, in many species, is attached to the stem by a stalk, the petiole. The base of the leaf in many monocotyledons and some dicotyledons is differentiated into a sheath, which encircles the stem. The leaf base of some species possesses leaflike outgrowths, the stipules, which function in protecting the young leaf before it unfolds. The stipules are often inconspicuous and fall away early in leaf development. If only a single leaf blade is present, the leaf is simple. The simple leaves of species such as maples may be deeply lobed. A compound leaf has two or more blades, the leaflets. The leaf margins of some species are smooth, or entire, while the margins of others may be finely or coarsely indented. Only one vascular bundle or vein may be present, as in conifers, but leaves of most ferns and angiosperms have numerous veins. In many dicotyledons, the veins form a reticulate, or netted, pattern. The leaves of many monocotyledons have veins that run essentially parallel throughout most of the blade. In both types of venation, the veins join with one another at intervals, forming a closed system. The venation of many ferns, *Ginkgo*, and a few dicotyledons consists of repeatedly forked, or dichotomous, veins.

Leaf venation

Internal structure. Internally, foliage leaves are constructed of three main tissues: epidermis, mesophyll, and vascular tissue. The epidermis, a single layer of living cells, forms a protective layer on both surfaces of the leaf. The outer walls of the epidermal cells are covered by a delicate film, the cuticle, composed of a waxy material, cutin, which enables the leaf to resist excessive evaporation of water. In many plants the waxy cuticle forms a light-gray "bloom" that can be easily wiped off, as in the leaves of red cabbage. A distinctive feature of the epidermis is the conspicuous pairs of usually bean-shaped cells with a small opening between them. Since the size of the opening can vary through changes in the shape of the pair of cells, they are known as guard cells. The pair of guard cells and the opening between them constitutes a stoma (plural: stomata). The stomata provide a connection between the external atmosphere and the system of intercellular spaces inside the leaf, thereby facilitating the exchange of gases between the atmosphere and the leaf interior. The stomata are usually much more abundant on the lower surface of the leaf than on the upper. Many woody plants typically lack stomata on the upper surface of the leaf. The floating leaves of aquatic plants, however, usually have stomata only on the upper surface.

Arrangement of cells inside leaves

Enclosed within the epidermis is the mesophyll, a tissue composed of parenchyma cells rich in chloroplasts (small structures containing chlorophyll, the photosynthetic pigment). This is the principal photosynthetic tissue of the plant. In many species, the mesophyll is differentiated into regions called the palisade parenchyma and spongy parenchyma. The palisade cells are more elongated at right angles to the epidermis and are separated from each other by spaces that facilitate the aeration of these photosynthetic cells. The spongy parenchyma occupies the lower half of the mesophyll and possesses a conspicuous system of air spaces that connect with the stomata, thus facilitating the movement of gases. The leaf, therefore, possesses an aerating system of impressive proportions, the spaces between the mesophyll cells making possible a thorough exchange of water vapour, carbon dioxide, and oxygen between the photosynthetic cells and the outside atmosphere.

The vascular tissues of the leaf are contained within the veins, which form either a network or a parallel system coursing throughout the mesophyll. Each of the larger veins is enclosed within a layer of tightly packed parenchyma cells, the bundle sheath, which extends to the ends of the veins, completely enclosing each one. The bundle-sheath cells participate in the movement of materials between the vascular tissue inside the veins and the mesophyll cells. In addition to containing the conducting xylem and phloem tissues, the veins serve as a strengthening framework for the leaf. Additional support for the leaf blade is provided by the pressure of the water inside the mesophyll cells. The epidermis also affords much support for the leaf because of the compact arrangement of its cells and the strength of the cuticle layer. Collenchyma cells (cells with thickened walls) and sclerenchyma fibres may also be present beneath the epidermis or in close association with the veins, an arrangement that provides additional support for the leaf blade.

Appearance and behaviour of cotyledons

Kinds of leaves. In addition to the more familiar foliage leaves, plants usually possess one or more of the following kinds of leaves: cotyledons, scale leaves, and bracts. Cotyledons are the first leaves to be formed, arising early in the development of the embryo inside the seed. During germination the cotyledons in some species remain enclosed within the seed coat and hidden beneath the soil surface. In other species the cotyledons break through the seed coat and appear on the young seedling above the ground surface, where they become green and carry on photosynthesis for a time, just as do foliage leaves. Scale leaves occur as bud scales, where they function in protecting the young developing leaves and apical meristem inside the bud. Scale leaves are the only type of leaf produced by underground stems, or rhizomes; they are usually small and short-lived. Bracts are often associated with the flowers, where they serve to protect the young developing flower. Sometimes the bracts are brightly coloured (as in the poinsettia) and function as petals in the attraction of prospective pollinators (bees, butterflies, birds, etc.; see POLLINATION).

Leaf fall. Leaves are relatively short-lived organs and sooner or later are shed from the plant, leaving leaf scars on the stem marking their former locations at the nodes. In evergreen species, such as conifers, and many broad-leaved angiosperms, the leaves may be retained on the tree for two or three years, after which they fall irregularly. In deciduous plants, the leaves function for only a single growing season and fall completely, usually in the autumn or at the beginning of the dry season. The fall of leaves of perennial and woody plants is a complex phe-

nomenon involving the separation of the leaf from the stem without undue damage to the newly exposed living tissues. Preceding the fall of the leaves from the tree, a special group of cells known as the abscission layer differentiates across the petiole near its base. Through chemical changes in their walls, the cells in this abscission layer become softened until the leaf finally breaks away and falls to the ground. On the stem side of the abscission layer, a healing layer develops, closing the wound with a corky tissue, which forms the leaf scar. The abscission process involves complex biochemical processes that are, in turn, influenced by various environmental factors such as day length, temperature, availability of oxygen, water, and mineral nutrients, and even attacks by insects or fungi. Various chemical sprays are often used in agriculture and in warfare to promote leaf abscission.

Leaf modifications. Leaves of many species have undergone various modifications during the course of evolution. Their form and structure have become conspicuously adapted in varying degrees to their environment. The leaves of plants in very dry habitats may be very small in blade area, or they may quickly die and in such plants the photosynthetic function is performed mainly by the stem, as in the cacti. In species such as *Acacia*, the leaf blade has disappeared, and the petiole has become flattened and leaflike (a phyllode). Many plants of deserts and other dry habitats possess pointed, awl-shaped spines, which are, in many cases, evolutionary derivatives of leaves. The leaves of other desert plants are very thick and fleshy, with abundant water-storage capacity; such plants are known as succulents.

Adaptations to dry habitats

Climbing plants often possess elongated, threadlike, branched or unbranched organs known as tendrils, which wrap themselves around nearby stems, thereby anchoring the plant. In many species tendrils are evolutionary modifications of leaves or leaf parts. In cucumbers and melons, for example, the tendrils have evolved from the midribs of leaves, whereas in peas the tendrils originated from a leaflet of a compound leaf.

Highly bizarre leaf modifications are present in carnivorous plants. The leaf of the sundew plant (*Drosera*) has tentacles resembling the "horns" of a snail. These tentacles excrete a sticky fluid attractive to insects, which become firmly caught and are eventually digested by the leaf. The leaves of the Venus's-flytrap function as mechanical traps, capturing insects with amazing rapidity. The leaves of *Sarracenia* and *Nepenthes* are modified into hollow, trumpet- or jug-shaped, brightly coloured structures containing a watery fluid attractive to animals. Upon entering the leaf, the animals (usually insects) slide into the fluid and are unable to escape. Subsequently, their bodies are digested and the dissolved materials used in the plant's own nutrition.

The root. *Form and growth.* The root is generally the underground portion of the plant body and functions primarily as an anchoring and absorbing organ. Some subterranean structures are actually modified stems. The root, however, differs from such underground stems by always lacking leaves and buds and by having its growing tip protected by a root cap. The root cap consists of short-lived parenchyma cells. A short distance behind the root tip, where the cells have reached their maximum length, root hairs occur. Each root hair, a tubular outgrowth of a single epidermal cell, greatly increases the surface area of the root, thereby facilitating absorption of water and dissolved minerals from the soil. Root hairs usually live only a few days but are rapidly replaced by new ones, which form just back of the growing root apex. Root hairs are usually lacking in plants living in aquatic habitats.

Primary roots

The first root of a plant, the primary root, develops from the root end of the embryo (the radicle) during seed germination. In gymnosperms and dicotyledons, the primary root usually develops into the root system of the plant through the formation of branch or secondary roots from the pericycle tissue, a special layer of cells surrounding the vascular bundle, or stele, in the interior of the root. In monocotyledons, however, the primary root commonly dies while the plant is still very small. In these plants adventitious roots, roots that grow in other than the usual place on a plant, develop from the stem, usually close to the nodes. Adventitious roots are also produced on the creeping and underground stems (rhizomes) of many dicotyledons.

Internal structure. Internally, the root possesses a definite pattern of tissues. Three main regions are discernible: epidermis, cortex, and vascular cylinder. The epidermis is composed of tightly packed cells and typically lacks a cuticle. Root hairs occur in the region near the root tip. The cortex consists mostly of parenchyma cells that are characteristically separated from each other by spaces of various sizes. The cortex functions in the movement of water and dissolved minerals across the root from the epidermis to the xylem and also stores food molecules transported downward from the leaves. The innermost layer of the cortex usually differentiates as an endodermis, composed of a cylinder of cells, each of which develops a narrow, waterproof band, the Casparian strip, around all but the innermost- and outermost-facing cell walls. The Casparian strip serves as a barrier to the free movement of water and minerals between the cells, thus requiring all such materials to enter the central stele through the membranes of the living endodermis cells.

Interior to the endodermis is the vascular cylinder, which begins with one or more layers of parenchyma cells, the pericycle, immediately inside the endodermis. Cells of the pericycle retain the ability to undergo cell division, thereby giving rise to branch or secondary roots. On the inner surface of the pericycle are the vascular tissues, which, in many species, are arranged in a star-shaped pattern. Comprising the core of this star is the xylem tissue. The phloem tissue is located in small groups between the points of the star. Usually the phloem is separated from the xylem by a cambium layer, which, in those species whose roots undergo an increase in thickness, produces both secondary xylem and phloem. In most monocotyledons and some herbaceous dicotyledons, the central core of the stem consists of pith tissue.

Secondary growth in roots

In many plants the roots undergo increase in thickness through the production of secondary growth by the vascular cambium, just as does the stem. In fact, the old roots of trees may reach several inches in diameter and exhibit numerous annual rings similar to those in the wood of the stem. As the root increases in thickness, a secondary covering of periderm (bark or cork) similar to that of the stem replaces the epidermis as a protective covering. Externally, therefore, the older root can hardly be distinguished from an older stem, especially when erosion has removed the soil cover, exposing the root to weathering. Internally, the woody root can hardly be distinguished from the woody stem.

Modifications of the root. During the course of evolution, roots of many species have become modified and specialized with reference to particular functions. The root may function as a food-storage organ, becoming enlarged and swollen, as in the carrot, turnip, radish, or sweet potato. Roots may serve as supporting structures, as in mangrove and corn (maize) plants. In plants of swamps and marshes, branch roots may grow above the mud into the atmosphere and function as aerating organs (pneumatophores) for the entrance of oxygen into the root system. In some palms the roots become modified into thorns. Many species of vines and epiphytes (plants that are supported on the branches of trees rather than being rooted in soil) form aerial roots that hang freely in the air. In many trees and shrubs, the absorptive roots are short and thick and are enclosed within a dense network of fungus hyphae (filaments). The fungus enters the root cortex, producing a condition known as mycorrhiza. The mycorrhizal association of root and fungus takes the place of root hairs in the absorption of water and dissolved minerals. The roots of the mistletoe plant, a partial parasite on trees, enter the stem cortex of the host, where they absorb water and dissolved minerals (see SANTALALES). In Spanish moss, an epiphyte, roots are absent, and the leaves function in the absorption of water and minerals by means of special absorptive hairs.

Mycorrhizal fungi

PHYSIOLOGY OF THE PLANT ORGANS

The external and internal morphology of the organs of the plant body can be understood simply by viewing them as adaptations facilitating the way the green plant makes its living as an autotrophic (*i.e.*, self-feeding—plants manufacture their own food by photosynthesis) organism. Although it is convenient for botanists to describe the plant in terms of three organs (stem, leaf, root), the plant must be viewed as a whole functional organism in which the cells, tissues, and organs work together to carry out an integrated series of activities. Only when the individual organs of the plant body are viewed in relation to the whole can the intricate mechanisms by which the living plant maintains itself begin to be understood.

During the course of plant evolution, there has been considerable selective value in the development by the green plant of light-accessible surfaces: the erect stem with many branches carrying numerous flattened leaves. It has also been of great selective value for the green plant to develop an underground portion highly efficient in both anchoring the plant and in absorbing water and dissolved minerals from the soil; hence, the root with its many branches and numerous root hairs has developed.

Green plants require three classes of materials from their environment: water, minerals, and the gases carbon dioxide and oxygen. To be used by the plant, however, these materials must be moved from the environment into the plant body.

Water absorption by roots

Nearly all of the water for the plant is absorbed by the roots, especially through the root hairs. In the soil surrounding the root hairs, water is usually present in high concentration. Inside the root cells, water is generally in much lower concentration because of the dissolved sugars, salts, and other substances present in living cells. Water molecules, therefore, move into the root cells by the process of osmotic diffusion, the movement of water across cell membranes (which prevent the passage of sugars, salts, and many other substances) from regions of high water concentration to regions of lower water concentration. There is also evidence suggesting that some of the water absorption by the root requires the expenditure of energy by the root cells.

Considerable amounts of water may enter the root and move across the epidermis and cortex by diffusing through the cell walls, never actually entering the living cells. To move into the vascular cylinder, however, the water molecules must move across the membranes of the endodermis cell layer. The water is channelled in its movement through the endodermal-cell membranes by the bandlike layer of waxy thickening (Casparian strip) in some of the walls of the endodermis.

In addition to water entering the root through the root-hair region of young roots, considerable absorption of water has been found to take place in roots old enough to possess a corky bark. Since as much as 95 percent of the root surface of a tree or shrub consists of bark-covered roots, it is most likely that the major portion of the water and mineral absorption occurs not through the root hairs but through lenticels and cracks in these older roots.

Roots also absorb the mineral elements necessary for plant growth. These elements enter the root cells as ions; *i.e.*, atoms or groups of atoms bearing an electric charge. Phosphorus, for example, is absorbed as phosphate ions (PO_4^{-3}), nitrogen as nitrate ions (NO_3^-) or ammonium ions (NH_4^+), and potassium in the form of potassium ions (K^+). There is good evidence that the root cells must expend energy to absorb mineral ions from the soil; that is, the absorption process is an active transport phenomenon. Water and minerals may also be absorbed through the stems and leaves. It is common practice, for example, to apply iron and other minerals to shrubs and trees by spraying them on the leaves.

Gaseous exchange in leaves

The gases, carbon dioxide and oxygen, enter the plant by diffusion whenever the stomata of the leaves are open. Experiments have demonstrated that, while the open stomata constitute only 1 to 2 percent of the epidermal surface, they possess, nonetheless, the capacity to permit the passage of gases from the atmosphere into the interior of the leaf to a degree far greater than that required by the plant. It seems very likely that the stomatal complex of guard cell, stoma, and intercellular-space system of the leaf evolved primarily as an adaptive device facilitating the absorption of the carbon dioxide required for photosynthesis and the oxygen necessary for respiration. Once inside the intercellular-space system of the leaf mesophyll, carbon dioxide and oxygen become dissolved in the thin film of water that encloses the mesophyll cells and then diffuse into the interior of the cells.

It has been known at least since the late 17th century that plants not only absorb but also lose large amounts of water from their leaves. The corn plants on a single acre, for example, have been shown to lose some 324,000 gallons of water during one growing season. When glass containers are placed over a leafy shoot, water will soon condense on the container walls. The process of water loss involves evaporation from the mesophyll cells into the intercellular spaces. The diffusion of this water vapour to the exterior of the leaf is controlled by the behaviour of the guard cells enclosing the stomata. Since the loss of water in gaseous form from the leaves differs in this respect from ordinary evaporation, the term transpiration is used for this process.

Transpiration, the loss of water from the leaf in vapour form, is, therefore, one of the normal functional processes of plants. The plant can do little to stop this water loss. The stomata are an adaptation for the movement of carbon dioxide and oxygen from the atmosphere into the leaf. Both gases are absolutely necessary for the continued life of the plant. A structure that is an excellent adaptation for the entrance of gases is also an efficient device for the loss of water vapour from the leaf interior. In fact, the entire leaf structure is highly favourable for the loss of water by transpiration. Whenever the stomata are open, there is a direct pathway from the surface of the mesophyll cells in the leaf interior out through the stomata into the atmosphere. There is some evidence that the loss of water from the leaf by transpiration is of some value in cooling the plant. On the other hand, when plants are transpiring rapidly, as in the middle of a hot, dry summer day, they may lose so much water that the cells of the leaves and young stems lose turgor (internal water pressure), causing these structures to wilt. If the wilting is severe, the leaves and young stems may droop, and the leaves become rolled. Little harm is done to the plant if these tissues recover their turgidity at night, but, if the loss of water has been too great, the plant may not recover and may eventually die from dehydration.

Causes of plant wilting

If water is being lost from the plants by transpiration from the leaves and if water is being absorbed from the soil by the roots, then there must be some mechanism for the transportation of water from the root tissues upward through the xylem of the stem into the veins of the leaves. Many explanations have been proposed to account for this water movement upward in plants; none is completely satisfactory. Various experiments have clearly demonstrated that the water is actually being pulled upward in the plant by forces acting in the leaves. It has been shown that the water in the xylem cells is actually under tension; *i.e.*, it is being pulled much in the way a rubber band is stretched.

In addition to the transport of water and dissolved minerals in the xylem, there is also movement through the phloem of sugars, hormones, and vitamins. Much is known about the phloem-transport process, yet no single explanation yet proposed has been able to account for all the various observations about conduction of materials through the phloem. It is known that the phloem cells are alive, unlike the cells of the xylem, which are dead at functional maturity. The rate of movement of materials through the phloem is relatively rapid, although much slower than movement through the xylem.

THE ORGAN SYSTEMS OF VASCULAR PLANTS

Variability of growth forms. The body of a vascular plant consists basically of the three organs: stem, leaf, root. Any individual plant, however, may possess several to many stems (and stem branches), numerous leaves, and many roots. It is convenient for descriptive purposes

to speak of the stem with its many branches and numerous leaves as comprising the shoot system and all of the roots as the root system. The shoot system and the root system form one continuous structure and function together in a remarkably integrated, harmoniously interacting organization, the whole plant body.

Extremes of plant size

Vascular plants exist in a remarkable array of diverse forms and sizes. Among the angiosperms, the plant body ranges from giant *Eucalyptus* trees of over 300 feet in height to the minute floating aquatic herb *Wolffia*, with a body so tiny that it can be more easily felt between the fingers than seen with the unaided eye. This amazing diversity presented by vascular plants is the result of the working-out in detail during evolution of a scheme of construction that is in itself unlimited. Any plant has the fundamental architectural potential to attain indefinite size, even the smallest herbs (through the activity of the meristems).

The unlimited scheme of growth exhibited by plants becomes evident very early in the development of the embryo. As the embryo grows, the production of new cells and tissues gradually becomes limited to certain regions of the embryonic axis, the meristems. The meristems remain permanently embryonic. Thus, unlike an animal, which usually completes the construction of all its tissues and organs during an early period of embryonic development, a plant body is partially embryonic and partially adult throughout its entire life. In a plant body, therefore, additional organs (*e.g.*, leaves) are formed throughout the life-span of the individual, while, in the animal body, the number of organs is fixed at a certain definite number early in embryo development.

The shoot system. *Branching habits.* The shoot system consists of the stem, its branches, and the leaves. During its development, the shoot system may attain considerable complexity due to branching. In lycopods and some ferns, the growing shoot subdivides or dichotomizes into two more or less equal apices. Each of these sister apices grows individually for a time and then may dichotomize once again. Dichotomous branching is considered to be the most primitive type of ramification in the vascular plants.

The shoot system in the vast majority of vascular plants consists of a main shoot axis or stem from which lateral branches develop from buds in the leaf axils; this is monopodial or lateral branching. The lateral branches may themselves branch and rebranch, producing an intricate shoot system. In some species, one of the axillary buds near the apex of the main stem may grow vigorously, soon overtopping the main stem, producing an apparent monopodial condition that actually consists of a series of lateral branches. This is sympodial branching.

The lateral branches of the main stem may grow into long shoots with widely spaced internodes and numerous leaves. In some species, however, most of the lateral branches are short shoots with crowded internodes and only a few leaves. In pines, most of the leaves are borne on short shoots. The flowers and fruits in many plants occur only on short shoots. Short shoots are apparently caused by growth-inhibiting hormones produced by the long-shoot tissues.

The tree—a model shoot system. Perhaps the tree is the most familiar type of shoot system. Trees vary considerably in both external form and in internal construction. In the tree ferns, much of the trunk thickness is caused by a thick sheath of intertangled adventitious roots that arise close to the shoot apex and grow downward through the trunk. At the base of the trunk, a dense mat of these roots forms a buttress that aids in supporting the tree trunk. Another type of tree construction occurs in palms and other, arborescent (treelike) monocotyledons. In these plants, the stem apex of the seedling enlarges rapidly just below the apical meristem. Cells in the region below the embryonic leaves function as a primary thickening meristem, which enlarges the stem of the young palm to almost the full diameter it is destined to have. Only after this dramatic increase in thickness does the stem begin to increase markedly in length. Thus the stout trunk is a primary construction of the stem apex and involves no secondary tissues of the kind produced by the vascular cambium of dicotyledons.

Increase in stem thickness among palms

The most common type of tree construction is that characteristic of the conifers and angiosperms. In these plants the original stem is produced by the apical meristem. Later, the stem becomes thicker by the activity of a lateral meristem, the vascular cambium. The vast bulk of a coniferous or angiospermous tree, therefore, is a secondary construction of the cambium, not the apical meristem.

In addition to variations in their internal anatomy, trees also exhibit a wide range of external forms. At one extreme is the short, unbranched or sparsely branched tree with a stout trunk supporting an umbrella-like crown, or rosette, of large compound leaves. Some of these umbrella, or rosette, trees have soft trunks with only weak secondary thickening; *e.g.*, the papaya (*Carica*). Other species—*e.g.*, the cycads and palms—exhibit the same general form but possess much more secondary thickening and much stouter trunks. At the other extreme is the lofty canopy tree with an immense crown consisting of thousands of small leaves borne on hundreds of twigs on the lateral branches. The canopy tree is well developed in the tropical rain forests. Between these extremes are various types of trees with diverse combinations of branching pattern, leaf types, sizes, and wood and bark development.

The shrub and other shoot systems. Another familiar type of shoot system is the shrub. The shoot system of a shrub is characterized by extensive branching, usually from or near the ground level. The buds in the lower portion of the main stem tend to develop instead of those higher up the principal axis. The shrubby shoot system generally is much shorter than that of the tree, and the life-span of the shrub is usually much less than that of the tree.

The shoot system of some species is well adapted for climbing up on other plants, thereby bringing the leaves into the sunlight. Such climbing plants are called vines or lianas and are very abundant in the tropical rain forests, often weaving the canopy into a continuous web. The stem of many species of vines is a long, flexible woody structure resembling a rope or cable. Internally, the secondary xylem is often arranged in bands, furrows, or discrete cylinders or bundles. Many species possess special structures that facilitate climbing. The shoots of roses (*Rosa*) and blackberries (*Rubus*) possess sharp, curved epidermal outgrowths called prickles. Other plants (*e.g.*, *Bougainvillea*) have hooklike thorns. Climbing plants such as grape (*Vitis*) have tendrils that wrap around the stems of other plants.

Lianas

Shoot adaptations. The shoot systems of many species exhibit various modifications enabling the plant to be better adapted to its environment. Perennial herbs, for example, may have slender stems that grow horizontally along the soil surface or just beneath the ground level; these horizontal stems are known as stolons or runners. In other plants the underground horizontal stem may be a thick structure with short internodes and much-reduced, colourless scale leaves. These are rhizomes, the "root" of *Iris* being a familiar example. Adventitious roots usually arise from the nodes. These subterranean stems or rhizomes function as food-storage organs. Rhizomes usually continue growth for several years, giving rise each growing season to new, aboveground portions of the shoot. Other plants produce short thick stems or parts of stems in which food is stored but that persist only from one growing season to the next. These structures, known as stem tubers, may be formed entirely from the hypocotyl (the root end of the embryo), as in the radish and beet, or they may arise at the tips of underground stolons, as in the potato. The depressions ("eyes") on a potato tuber are the locations of lateral buds and small scalelike leaves (visible in the young potato tuber). After the tubers are produced, the parent plant dies. In such species as crocus and gladiolus, the lower portion of the main shoot itself may develop into another kind of thickened underground stem known as the corm. Thin, membraneous scale leaves cover the solid, essentially round stem. Another

Stem tubers

type of shoot modification is the bulb, consisting of an underground, very short stem from which arise thick, fleshy scale leaves in which food is stored. Bulbs are characteristic of onions, tulips, hyacinths, and leeks.

The root system. Just as the shoot system becomes more complex through the production of lateral branches, the root system increases in surface area by branching. In the lycopods, dichotomous branching of the root occurs. In most vascular plants, root branching is lateral. Unlike the lateral branches of the shoot, however, the lateral roots arise at some distance back of the growing apex in a region where the tissues are fully matured. In addition, lateral roots originate from tissues deep within the parent root, usually the pericycle of the vascular cylinder in the seed plants, and they must penetrate the cortex tissue of the parent root in their growth outward into the soil. Usually the lateral roots are more slender than the parent root, and they are usually arranged more or less at right angles to the main root. Most trees and shrubs and many herbs have a root system consisting of a radially symmetrical taproot that grows straight downward into the soil. From this taproot, lateral or branch roots develop in sequence, the youngest near the growing root tip, and spread out horizontally or obliquely into the soil. In turn, these branch roots give rise to smaller lateral roots. The result is an extensively branched root system that is especially effective in both the absorption of water and dissolved minerals and in anchoring the plant. The taproots of many species may be especially adapted for the storage of food; *e.g.*, the carrot. The lateral roots of some species (*e.g.*, *Dahlia*) develop into thick, fleshy tubers, which, like the stem tubers they resemble, serve as food-storage structures. The root system of grasses, sedges, and most ferns is a fibrous system composed of numerous roots arising from the stem, all of about the same size. Although fibrous root systems do not anchor plants in the ground as well as do taproot systems, they are well adapted for absorbing water near the surface. In ferns and in many monocotyledons, most of the root system is adventitious, with the roots developing from the stem nodes. This type of root system is well illustrated by corn and grasses.

Root systems of grasses

Transition between the root and shoot systems. Usually there is a transition between the root system and the shoot system near the soil level. This transition is marked externally in the young plant by a rapid abrupt change in coloration from the pale colour of the root to the green colour of the stem. In woody plants with secondary growth, the change in external appearance is less pronounced since the bark extends over both the stem and older roots. In the embryo, the connection between the root and shoot systems is accomplished by the embryo axis or hypocotyl. Internally, the transition between the two parts of the plant body involves changes in the arrangement of the primary vascular tissues. The vascular tissue, which in the root is essentially a single central core or bundle with discrete strands of phloem lying adjacent to the central core of xylem, is gradually changed so that, in the stem, the xylem and phloem tissues assume characteristic positions within the several vascular bundles. Once cambial activity begins, however, the secondary xylem and phloem form continuous tissues between the root and the stem, thus obscuring the initial differences in the structure of the root–shoot transition region.

The reproductive system. Both the shoot and the root systems of the plant may participate in reproduction. The shoot system is involved in the formation of flower clusters or inflorescences; the inflorescence is a part of the branching system of the stem with flowers at the tips. Inflorescences vary greatly in size, form, and in flower number; various types have been distinguished. For more information on flowers, see REPRODUCTIVE SYSTEMS, PLANT; for seeds and fruit, see SEED AND FRUIT.

Asexual reproduction.

Both the shoot and the root systems are involved in vegetative or asexual reproduction. The shoot system exhibits diverse modifications facilitating reproduction. Some plants—*e.g.*, the aquatic plant *Anacharis canadensis* (elodea)—multiply largely by fragmentation of the plant body. Gardeners and horticulturists propagate plants by inducing cuttings from the parent plant to produce adventitious roots. In many plants, specialized portions of the plant body known as propagules are produced. The propagule may be a stolon or runner, which gives rise to new plants at the nodes, as in the strawberry. In species such as iris and canna, the propagule is a rhizome, which forms a new plant at the growing tip. The corm is utilized as a propagule in species such as crocus and gladiolus, while bulbs serve the same function in onions, tulips, and many other plants. Stem tubers are used to propagate potatoes; the tuber is cut into pieces, each of which gives rise to a new plant. Many aquatic plants, such as *Utricularia*, develop special buds called turions that overwinter in the bottom mud and germinate the following spring to produce new plants. Small, fully formed plantlets are formed on the leaf margins of many ferns and angiosperms such as *Bryophyllum*. These eventually fall to the ground and grow into new plants.

Like the shoot system, the root system also participates in asexual reproduction through the production of a diversity of structures. Species of *Cyrilla* and *Hypericum* develop new plants at irregular intervals along roots that grow just beneath the soil surface; these are "root sprouts." Many perennial and biennial herbs form root tubers; *e.g.*, *Dahlia* and many terrestrial orchids. See also REPRODUCTIVE SYSTEMS, PLANT.

THE ORGANS OF NONVASCULAR PLANTS

Differentiation of the plant body into organs has occurred also in various kinds of multicellular nonvascular plants, including mosses, liverworts, and algae. Much of the vegetative body of mosses and most liverworts is constructed very similarly to that of the vascular plants; that is, there is an aerial, shootlike portion attached to a subterranean rootlike portion. The leafy shootlike gametophyte (*i.e.*, one of the two main phases in the life cycle of lower plants—in mosses and liverworts it is the conspicuous leafy growth form) of mosses exhibits differentiation into three tissue regions: epidermis, cortex, and central conducting strand. While the conducting tissue is composed of phloemlike cells, there are no tracheids (elongate overlapping water-conducting cells that communicate by lateral pores) or vessel elements (more advanced tubular hollow cells with open ends) such as those of vascular plants. The leaves are usually one-cell-layer thick, except over the midrib, and typical mesophyll is lacking. The rootlike underground portion has filamentous structures called rhizoids, but these serve mainly to anchor the plant in the soil. The gametophyte body of some liverworts consists of a flat, ribbonlike, often dichotomously branched, green structure with filamentous rhizoids on its lower surface. Both mosses and liverworts are largely restricted to moist habitats.

Lower plant structures compared to vascular plants

Among the algae are many species that exhibit bodies remarkably similar to those of vascular plants, at least in external appearance. The stoneworts *Chara* and *Nitella* have an upright, green, shootlike portion with whorls of branches at the nodes and very long internodes. The plants are anchored to the mud in the bottom of ponds and streams by colourless, branched filamentous rhizoids that arise from the nodes near the base. Among the brown algae (Phaeophyta) are many species with bodies composed of leafy shootlike portions anchored to the rocks by means of rootlike holdfasts; *e.g.*, the giant kelps *Nereocystis* and *Macrocystis*. Some of these large brown algae even have food-conducting cells remarkably similar in structure and function to the sieve cells (a type of phloem cell) of vascular plants. Various members of the red algae (Rhodophyta) also display leaflike and stemlike bodies very similar in external appearance to the leaves and stems of vascular plants. The internal anatomy of these highly organized algal bodies is much less complex than that of the vascular plants, however. See ALGAE. Among the green algae (Chlorophyta), the genus *Fritschiella* has a multicellular body consisting largely of parenchyma tissue differentiated into an erect, aerial system anchored to the mud by a prostrate system of rhizoids. In fact, these features make *Fritschiella* an attractive model for a possible ancestor of green land plants.

EVOLUTION OF PLANT ORGANS AND ORGAN SYSTEMS

Functional evolution. The great diversity of organs and organ systems among the numerous species of plants is the result of evolution by natural selection. These structures enable the plant to be better adapted to its environment. The origin and diversification of multicellular plants with complex organs and organ systems have long been a subject of intense interest among plant evolutionists. The fossil record indicates that simpler plants originated earlier in geological time than the more complex plants. It is consistent with the concept of evolution by natural selection to assume that plants received some selective advantage in becoming more complex.

Advantages of complex multicellular structure

What advantages may have accrued to plants by the adoption of the more complex multicellular state? The multicellular condition enables cell and tissue specialization, a division of labour among the parts of the whole plant body. Some cells and tissues can become specialized for capturing sunlight and synthesizing food, others anchor the plant body, while still others play a role in reproduction. The result is a corresponding increase in the ability of the plant to exploit its environment and increase its chances for survival.

The fossil record. The time of origin of the multicellular plant body can be estimated from the earliest occurring fossils presently known. Filamentous structures identified as blue-green algae have been discovered in rocks calculated as being about 2,000,000,000 years in age. Multicellular bodies differentiated into organs had evolved among the green algae and possibly also among the red algae and brown algae by the beginning of the Ordovician Period, some 500,000,000 years ago.

Proposed conditions leading to land-plant evolution

The earliest known fossil land plants (*i.e.*, vascular plants with an erect aerial photosynthesizing system and a horizontal underground anchoring and absorbing system) are found in rocks formed during the Upper Silurian and Early Devonian periods, around 395,000,000 years ago. Geological evidence suggests that, during this time, many regions of the Earth's surface may have been subjected to prolonged seasonal droughts. During some portions of the year, rainfall would be abundant (as today in some tropical regions); at other times, the rains would cease, and the water level in the larger bodies of water would decrease drastically, while that of many smaller ponds and lakes might dry up altogether. Under these conditions of intermittent drought, the evolutionary migration of plants to the land is thought to have occurred.

Once the transition to full-time land existence had been accomplished, the stage was then set for the relatively rapid adaptive radiation of plants into all suitable habitats. This radiation produced several evolutionary lines of vascular plants as well as of the mosses and liverworts. The invasion of the land by photosynthetic plants also created a new environment—one in which fungi and animals could compete successfully for the necessities of life. Indeed, clear evidence that fungi were in existence during the Early Devonian Period is provided by the discovery of fungal hyphae in the fossilized remains of vascular plants that lived during that time. Fossil remains of land-dwelling invertebrate animals, which lived during this time, have also been discovered.

BIBLIOGRAPHY. E. STRASBURGER, *Lehrbuch der Botanik für Hochschulen,* 28th ed. (1962; Eng. trans., *Textbook of Botany,* 1965), a comprehensive college-level botany textbook and reference work (a classic since 1894, periodically revised by outstanding authorities); K. ESAU, *Plant Anatomy,* 2nd ed. (1965), a splendid, well-illustrated, college-level textbook of seed plant anatomy by a foremost plant anatomist; A.J. EAMES, *Morphology of the Angiosperms* (1961), an excellent college-level textbook that brings together both factual material and theories on the morphology and phylogeny of the angiosperms; T.E. WEIER, *et al., Botany,* 4th ed. (1970), an introductory college-level botany textbook with excellent coverage of plant morphology and well illustrated with many photographs and drawings; E.J.H. CORNER, *The Life of Plants* (1964), a highly original and refreshing story of important events in the evolution of plants from single cells to forest trees (well illustrated, especially with examples from the tropics); H.P. BANKS, *Evolution and Plants of the Past* (1970), a short, well-illustrated, college-level book on fossil plants by one of the foremost researchers in Devonian paleobotany; A.S. FOSTER and E. M. GIFFORD, JR., *Comparative Morphology of Vascular Plants* (1959), a well-written, attractively illustrated college-level textbook of vascular plant morphology; P. ADAMS, *et al., The Study of Botany* (1970), a college-level botany textbook, with emphasis on the thought processes of plant scientists, containing several chapters on the evolutionary aspects of plant anatomy and morphology; A.F. HILL, *Economic Botany* (1937), an old but still very useful college-level textbook describing the uses of plants and plant products; F.O. BOWER, *Plants and Man* (1925), an old but still excellent, relatively nontechnical discussion by a recognized authority on the life of plants, especially those aspects of interest to the thinking layman; P.R. BELL and C.L.F. WOODCOCK, *The Diversity of Green Plants* (1968), a well-illustrated college-level textbook on the morphology of autotrophic plants; F.B. SALISBURY and R.V. PARKE, *Vascular Plants: Form and Function,* 2nd ed. (1970), a well-written college-level discussion of plant anatomy in relation to plant physiology; M. RICHARDSON, *Translocation in Plants* (1968), a short, succinct discussion of the movement of water, minerals, and metabolites within the xylem and phloem tissues of the plant body; A. CRONQUIST, *The Evolution and Classification of Flowering Plants* (1968), a unique, advanced level book presenting major new concepts about flowering plant taxonomy and evolution; A. TAKHTAJAN, *Flowering Plants: Origin and Dispersal* (1969; Eng. trans. from the 2nd Russian edition, 1961), an advanced level treatment of the problems of the origin and dispersal of the flowering plants by a recognized Russian authority in plant evolution.

(P.Ad.)

Origen

Origen (Origenes Adamantius), the most influential and seminal theologian and biblical scholar of the early Greek Church, provoked questions and controversies that for centuries were to absorb the attention of churchmen.

Early life and education. He was born *c.* 185, probably in Alexandria, Egypt, of pagan parents, according to the Neoplatonist philosopher Porphyry, but of Christian parents, according to the ecclesiastical historian Eusebius of Caesarea, whose account is probably more accurate. Eusebius stated that Origen's father Leonides, was martyred in the persecution of 202, so that Origen had to provide for his mother and six younger brothers. At first he lived in the house of a wealthy lady. He then earned money by teaching grammar and lived a life of strenuous asceticism. Eusebius added that he was a pupil of Clement of Alexandria, whom he succeeded as head of the Catechetical school under the authority of the bishop Demetrius. Eusebius also alleged, without supporting documents, that Origen, as a young man, castrated himself so as to work freely in instructing female catechumens; but this was not the only story told by the malicious about his extraordinary chastity, and thus it may merely have been hostile gossip. Writing in 248, Origen deplored the fanaticism of those who literally interpreted Matt. 19:12 ("... there are eunuchs who have made themselves eunuchs for the sake of the kingdom of heaven"). Eusebius' account of Origen's life, moreover, bears the marks and embellishments of legends of saints and thus needs to be treated with this in mind.

Head of the Catechetical school

According to Porphyry, Origen attended lectures given by Ammonius Saccas, the founder of Neoplatonism. A letter of Origen mentions his "teacher of philosophy," at whose lectures he met Heraclas, who was to become his junior colleague, then his rival, and end as bishop of Alexandria refusing to hold communion with him.

Career as an educator and polemicist. Origen invited Heraclas to assist him with the elementary teaching at the Catechetical school, leaving himself free for advanced teaching and study. During this period (from *c.* 212), Origen learned Hebrew and began to compile his *Hexapla.*

A wealthy Christian named Ambrose, whom Origen converted from the teachings of the heretical Valentinus (who, as a Gnostic, advocated the dualism of spirit and matter) and to whom he dedicated many of his works, provided him with shorthand writers. A stream of treatises and commentaries began to pour from Origen's pen. At Alexandria he wrote his lost *Miscellanies* (*Stromateis*), *On the Resurrection* (*Peri anastaseos*) and, above all, *On First Principles* (*De principiis*). He also began his im-

mense commentary on St. John, written to refute the commentary of the Gnostic follower of Valentinus, Heracleon. His studies were interrupted by visits to Rome (where he met the theologian Hippolytus), Arabia, Antioch, and Palestine. Because of his reputation, he was much in demand as a preacher, a circumstance that provoked the disapproval of Demetrius, bishop of Alexandria, who was anxious to control this free lay teacher and especially angry when Origen was allowed to preach at Caesarea Palestinae. In about 229–230 Origen went to Greece to dispute with another follower of Valentinus, Candidus. On the way he was ordained presbyter at Caesarea. The Valentinian doctrine that salvation and damnation are predestinate, independent of volition, was defended by Candidus on the ground that Satan is beyond repentance; Origen replied that if Satan fell by will, even he can repent. Demetrius, incensed at Origen's ordination, was appalled by such a doctrinal view and instigated a synodical condemnation, which, however, was not accepted in Greece and Palestine. Thenceforth, Origen lived at Caesarea, where he attracted many pupils. One of his most notable students was Gregory Thaumaturgus, later bishop of Neocaesarea, who praised his master according to the contemporary fashion by writing a eulogy and claimed for Origen a position equal to that of the great pagan philosophers.

Ordination as a presbyter

From Caesarea, Origen continued his travels. In 235 the persecution of Maximinus found him in Cappadocia, from which he addressed to Ambrose his *Exhortation to Martyrdom*. During this period falls the "Discussion with Heracleides," a papyrus partially transcribing a debate at a church council (probably in Arabia) where a local bishop was suspected of denying the pre-existence of the divine Word and where obscure controversies raged over Christological issues and whether the soul is, in actuality, blood. During the persecution under the emperor Decius (250), Origen was imprisoned and tortured but survived to die in about 254 at Tyre. His tomb there was held in honour, and its long survival is attested by historians of the period of the Crusades.

Writings. Origen's main lifework was on the text of the Greek Old Testament and on the exposition of the whole Bible. The *Hexapla* was a synopsis of Old Testament versions: the Hebrew and a transliteration were followed by the Septuagint (an authoritative Greek version of the Old Testament), the versions of Aquila, Symmachus, and Theodotion and, for the Psalms, two further translations (one being discovered by him in a jar in the Jordan Valley). The purpose of the *Hexapla* was to provide a secure basis for debate with rabbis to whom the Hebrew alone was authoritative. The Greek churches generally accepted the Septuagint text and canon. Origen frequently mentioned textual variants among New Testament manuscripts but undertook no study of them comparable to the *Hexapla*, which is extant only in parts.

The *Hexapla*

Exegetical writings. His exegetical writings consist of commentaries (scholarly expositions for instructed Christians), homilies for mixed congregations, and scholia (detached comments on particular passages or books) which are almost entirely lost. None of his commentaries survives complete; quite apart from doubt about his orthodoxy, their length militated against preservation in that they were too long to copy. All extant manuscripts of the commentary on St. John, which extended to 32 books, depend on a codex preserved in Munich (Germany) containing only a few of the books. This codex and a related manuscript at Trinity College, Cambridge (England), are the sole witnesses for the Greek original of books 10–17 of his commentary on St. Matthew. Greek fragments of this, as of most of Origen's exegetical works, survive in writings known as *catenae* ("chains"; *i.e.*, anthologies of comments by early Church Fathers on biblical books). Commentaries on the Song of Solomon and on Romans survive in a drastically abbreviated Latin paraphrase by the Christian writer Tyrannius Rufinus (*c.* 365–410/411). The homilies on Genesis through the Book of Judges (except Deuteronomy) and Psalms 36–38 survive in a Latin translation by Rufinus. Jerome, the great Christian scholar (*c.* 347–*c.* 420), translated homilies on the Song of Solomon, Isaiah, Jeremiah, Ezekiel, and Luke. These Latin homilies were widely read in medieval monasteries and have a rich manuscript tradition. The Greek original of homilies on Jeremiah survives in a single manuscript in the Escorial (Spain), and that of a homily on the witch of Endor (which provoked early criticism for its thesis that Samuel really was conjured up) in a manuscript in Munich and on papyrus.

Doctrinal writings. Prior to 231 Origen wrote *De principiis*, an ordered statement of Christian doctrine on an ambitious scale, based on the presupposition that every Christian is committed to the rule of faith laid down by the Apostles (the Creator as God of both Old and New Testaments, the incarnation of the pre-existent Lord, the Holy Spirit as one of the divine Triad, the freedom of rational souls, discarnate spirits, the noneternity of the world, judgment to come) but that outside this restriction the educated believer is free to speculate. Origen was writing long before the conciliar definitions of Chalcedon (451) concerning the Trinity and the Person of Christ and at a period when a far larger area of doctrine could be regarded as open for discussion and argument than was the case by 400. *De principiis* diverged in its speculations from later standards of orthodoxy. The original was consequently lost and can only be reconstructed from the *Philocalia* (an anthology compiled by Basil the Great and Gregory of Nazianzus illustrating Origen's biblical interpretation), from Rufinus' Latin paraphrase (which avowedly rewrites heterodox-sounding passages), and from later writers, notably Jerome and Justinian I (who quote especially compromising passages to prove Origen a heretic). The polemical anti-Origenists, however, need to be read with care since they were not above misquoting Origen and ascribing to him the words of later Origenists.

Speculative theology

Polemical writings. Origen's great vindication of Christianity against pagan attack, *Contra Celsum*, written (probably in 248) at Ambrose's request, survives in its entirety in one Vatican manuscript, with fragments in the *Philocalia* and on papyruses. Paragraph by paragraph it answers the *Alēthēs logos* ("The True Doctrine" or "Discourse") of the 2nd-century anti-Christian philosopher Celsus and is therefore a principal source for the pagan intelligentsia's view of 2nd-century Christianity as well as a classic formulation of early Christian reply. Both protagonists agree in their basic Platonic presuppositions, but beside this agreement, serious differences are argued. Celsus' brusque dismissal of Christianity as a crude and bucolic onslaught on the religious traditions and intellectual values of classical culture provoked Origen to a sustained rejoinder in which he claimed that a philosophic mind has a right to think within a Christian framework and that the Christian faith is neither a prejudice of the unreasoning masses nor a crutch for social outcasts or noncomformists.

Devotional writings. The tract *On Prayer*, preserved in one manuscript at Cambridge, was written about 233; it expounds the Lord's Prayer and discusses some of the philosophical problems of petition, arguing that petition can only be excluded by a determinism false to the experience of personality, while the highest prayer is an elevation of the soul beyond material things to a passive inward union with Christ, mediator between men and the Father.

Theological system. Origen's experience as a teacher is reflected in his continual emphasis upon a scale of spiritual apprehension. Christianity to him was a ladder of divine ascent, and the beginner must learn to mount it with the saints in a never-ceasing advance. Uneducated and elementary Christians may entertain strange misunderstandings and may even believe things of God that would not be believed of the most savage and cruel men. But the truth at the higher level is plain: that God's nature is essentially goodness and that he desires of his creatures a love that is free. Everything in Origen's theology ultimately turns upon the goodness of God and the freedom of the creature. The transcendent God is the source of all existence and is good, just, and omnipotent. This omnipotence is never mere power emptied of moral quality; one cannot appeal to it to rationalize absurdity or the merely extraordinary. In overflowing love, God created rational and spiritual beings through the Logos (Word); this creative act involves a degree of self-limitation on God's part.

Nature of God

The cosmos. In relation to the created order, God is both conditioned and unconditioned, free and under necessity, since he is both transcendent to and immanently active. In one sense, the cosmos is eternally necessary to God since one cannot conceive such goodness and power as inactive at any time. Yet in another sense, the cosmos is not necessary to God but is dependent on his will, to which it also owes its continued existence. Origen was aware that there is no solution of this dilemma. The rational beings, however, neglected to adore God and fell. The material world was created by God as a means of discipline (and its natural catastrophes such as earthquakes and plagues remind man that this world is not his ultimate destiny). Origen speculated that souls fell varying distances, some to be angels, some descending into human bodies, and the most wicked becoming devils. (Origen believed in the pre-existence of souls, but not in transmigration nor in the incorporation of rational souls in animal bodies.) Redemption is a grand education by providence, restoring all souls to their original blessedness. For none, not even Satan, is so depraved and has so lost rationality and freedom as to be beyond redemption. God never coerces, though with reformative intention he may punish. His punishments are remedial; even if simple believers may need to think of them as retributive, this is pedagogic accommodation to inferior capacity, not the truth.

The role and work of Christ. The climax of redemption is the incarnation of the pre-existent Son. One soul had not fallen but had remained in adoring union with the Father. Uniting himself with this soul, the divine Logos, who is the second *hypostasis* (Person) of the Triad of Father, Son, and Spirit (subordinate to the Father but on the divine side of the gulf between infinite Creator and finite creation), became incarnate in a body derived from the Virgin Mary. So intense was the union between Christ's soul and the Logos that it is like the union of body and soul, of white-hot iron and fire. Like all souls Christ's had free will, but the intensity of union destroyed all inclination for change, and the Logos united to himself not only soul but also body, as was apparent when Jesus was transfigured. Origen, influenced by a semi-Gnostic writing, the *Acts of John*, thought that Jesus' body appeared differently to different observers according to their spiritual capacities. Some saw nothing remarkable in him, others recognized in him their Lord and God. In his commentary on St. John, Origen collected titles of Christ, such as Lamb, Redeemer, Wisdom, Truth, Light, Life. Though the Father is One, the Son is many and has many grades, like rungs in a ladder of mystical ascent, steps up to the Holy of Holies, the beatific vision.

The Logos and revelation

The union of God and man in Christ is pattern for that of Christ and the believer. The individual soul, as well as the church, is the bride of the Logos, and the mystery of that union is portrayed in the Song of Solomon, Origen's commentary on which was regarded by Jerome (in the period of his enthusiasm for Origen) as his masterpiece. Thus redemption restores fallen souls from matter to spirit, from image to reality, a principle directly exemplified both in the sacraments and in the inspired biblical writings, in which the inward spirit is veiled under the letter of law, history, myth, and parable. The commentator's task is to penetrate the allegory, to perceive within the material body of Scripture its soul and spirit, to discover its existential reference for the individual Christian. Correct exegesis (critical interpretation) is the gift of grace to those spiritually worthy.

Destiny of man and the world. Both the biblical revelation and the spiritual life of the believer are progressive. The church is the great "school of souls" in which erring pupils are disciplined: elementary education in this life, higher education in the world to come, where the atoning and sanctifying process will continue in a purging baptism of fire, burning up the wood, hay, and stubble. Hell cannot be an absolute since God cannot abandon any creature; because of his respect for freedom it may take time, but God's love will ultimately triumph. Christ's work remains unfinished until he has subdued all to himself. Heaven is not necessarily absolute because freedom is an inalienable characteristic of the rational creature. "If you remove free will from virtue, you destroy its essence." Because the redeemed remain free, when all souls have been restored the whole drama may begin again. The Stoics believed in world cycles determined by fate. Origen thought them possible for the opposite reason, because freedom means that there is no ultimate finality.

Hell and heaven

Influence. If orthodoxy were a matter of intention, no theologian could be more orthodox than Origen, none more devoted to the cause of Christian faith. His natural temper is world denying and even illiberal. The saintliness of his life is reflected in the insight of his commentaries and the sometimes quite passionate devotion of his homilies. The influence of his biblical exegesis and ascetic ideals is hard to overestimate; his commentaries were freely plagiarized by later exegetes, both Eastern and Western, and he is a seminal mind for the beginnings of monasticism. Through the writings of the monk Evagrius Ponticus (346–399), his ideas passed not only into the Greek ascetic tradition but also to John Cassian (360–435), a Semi-Pelagian monk (who emphasized the worth of man's moral effort), and the West. Yet he has been charged with many heresies. In his lifetime he was often attacked, suspected of adulterating the Gospel with pagan philosophy. After his death, opposition steadily mounted, respectful in the Greek Christian Methodius of Olympus' criticism of his spiritualizing doctrine of the Resurrection (*c.* 300), offensive in Epiphanius' (375), a refuter of Christian heresies, violent in Jerome's anti-Origenist quarrel with Rufinus (*c.* 393–402). Origen had his defenders, especially in the East (Eusebius of Caesarea; Didymus the Blind, the head of Catechetical School of Alexandria [*c.* 313–398], Athanasius, bishop of Alexandria [*c.* 293–373], to some degree; and especially the Cappadocian Fathers—*i.e.*, Basil the Great, Gregory of Nazianzus, and Gregory of Nyssa); but in the west Rufinus' translation of *De principiis* (398) caused scandal, and in the East the cause of Origen suffered by the permanent influence of Epiphanius' attack. In the 6th century the "New Laura" (monastic community) in Palestine became a centre for an Origenist movement among the monastic intelligentsia, hospitable to speculations about pre-existent souls, universal salvation, and spherical resurrection bodies (a belief not countenanced by Origen himself). The resultant controversy led Justinian I to issue a long edict denouncing Origen (543); the condemnation was extended also to Didymus and Evagrius by the fifth ecumenical council at Constantinople (553). Nevertheless, Origen's influence persisted, such as in the writings of the Byzantine monk Maximus the Confessor (*c.* 550–662) and the Irish theologian John Scotus Erigena (*c.* 810–877); and since Renaissance times, controversy has continued concerning his orthodoxy, Western writers being generally more favourable than Eastern Orthodox.

Ecclesiastical opposition

The chief accusations against Origen's teaching are the following: making the Son inferior to the Father and thus being a precursor of Arianism, a 4th-century heresy that denied that the Father and the Son were of the same substance; spiritualizing away the resurrection of the body; denying hell, a morally enervating universalism; speculating about pre-existent souls and world cycles; dissolving redemptive history into timeless myth by using allegorical interpretation, thus turning Christianity into a kind of Gnosticism, a heretical movement that held that matter was evil and the spirit good. None of these charges is altogether groundless. At the same time there is much reason to justify Jerome's first judgment that Origen was the greatest teacher of the Early Church after the Apostles.

BIBLIOGRAPHY. J. DANIELOU, *Origène* (1948; Eng. trans. 1955), the best biography; CHARLES BIGG, *The Christian Platonists of Alexandria*, 2nd ed. (1913, reprinted 1969), a sensitive and profound study of Origen's thought; HENRY CHADWICK, *Early Christian Thought and the Classical Tradition: Studies in Justin, Clement and Origen* (1966), on Origen's critique of Greek philosophical ideas; R.P.C. HANSON, *Allegory and Event* (1959), the best study of Origen's biblical interpretation.

Translations: Origen on First Principles, by G.W. BUTTERWORTH (1936); *Contra Celsum*, by H. CHADWICK, 2nd ed. (1965), both with notes and introduction.

(H.Cha.)

Orinoco River

The Orinoco River (Río Orinoco) and its tributaries constitute the northernmost of South America's three major river systems. Bordered by the Andes mountains to the west and the north, the Guiana Highlands to the east, and the Amazon watershed to the south, the river basin covers an area of about 366,000 square miles (948,000 square kilometres). It encompasses approximately four-fifths of Venezuela and one-fourth of Colombia. The Orinoco River itself flows in a giant arc for 1,337 miles (2,151 kilometres) from its source in the Guiana Highlands to its mouth on the Atlantic Ocean. Throughout most of its course it flows through Venezuela, except for a section where it forms part of the frontier between Venezuela and Colombia. The name "Orinoco" is derived from Guarauno words meaning "a place to paddle"—*i.e.*, a navigable place.

For most of its length, the Orinoco flows through impenetrable rain forest or undeveloped grasslands; however, the cities of Ciudad Bolívar (Bolívar City) and Ciudad Guayana (also known as Santo Tomé de Guayana or Guayana City) are located on its lower course, and this region is fast developing into one of the most industrialized areas of South America. The river forms a waterway used in the exploitation of the vast mineral wealth of the Venezuelan interior. (For an associated physical feature, see LLANOS.)

The natural environment. *The course of the river.* The western slopes of the Sierra Parima (Parima Mountains), which form part of the boundary between Venezuela and Brazil, are drained by spring-fed streams that give rise to the Orinoco River. The source is placed in Venezuela at 63°21′ W and 2°19′ N, at an elevation of 3,523 feet (1,074 metres). From its headwaters the river flows west–northwest, leaving the mountains to meander through level plains known as the Llanos. The volume of its waters increases as the river receives numerous mountain tributaries, including the Río Mavaca on the left bank and the Manaviche, Ocamo, Padamo, and Cunucunuma rivers on the right.

The Brazo Casiquiare

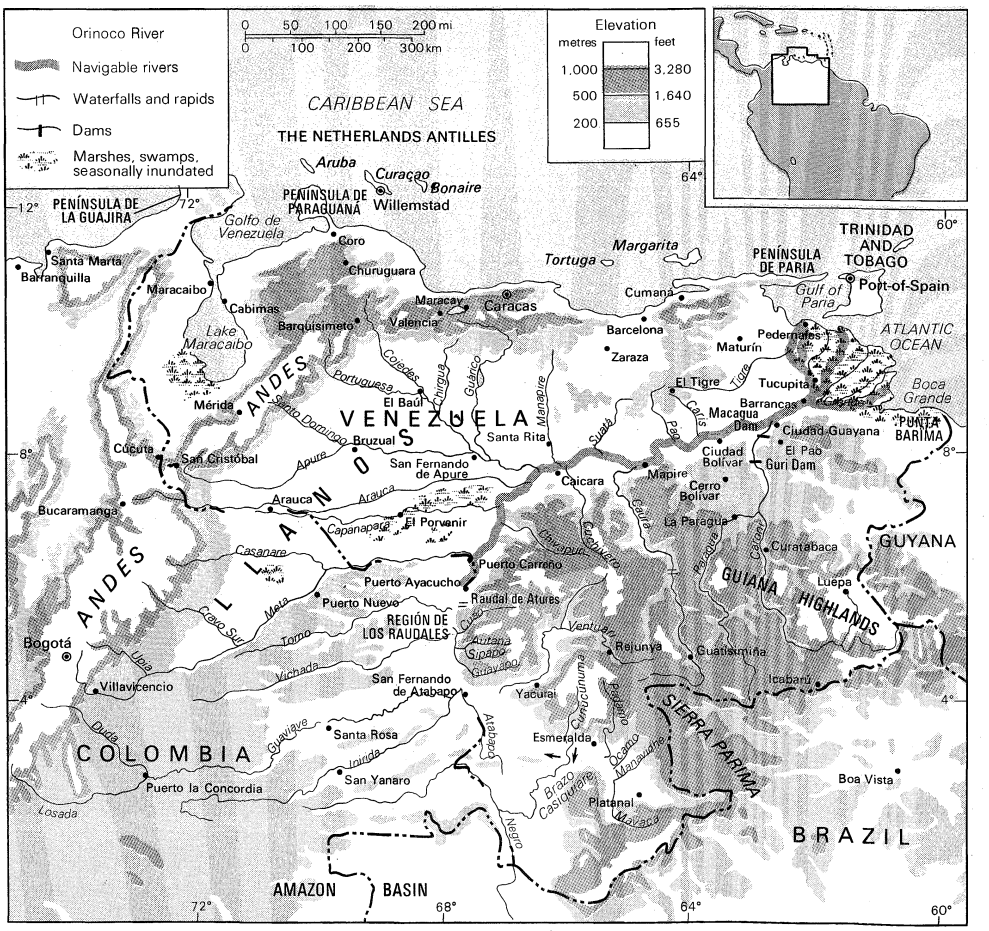
Below the town of Esmeralda some of the waters of the Orinoco flow south into the Brazo Casiquiare (Casiquiare Channel). This channel, a feature peculiar to the Orinoco River system, is a natural passage that flows for more than 220 miles to the Río Negro (Black River), linking the Orinoco and Amazon river systems.

After its bifurcation in the Casiquiare, the river bends to the northwest and flows in great meandering curves to its confluence with the Río Ventuari. There the river turns to the west to run between high alluvial banks; its course is marked by numerous extensive sandbars. Near San Fernando de Atabapo, the Atabapo and Guaviare rivers join join the Orinoco, marking the end of the Upper Orinoco.

Downstream from San Fernando de Atabapo, the river flows northward and forms part of the border between Venezuela and Colombia. It passes through a transitional zone, the Región de los Raudales (Region of the Rapids), where the Orinoco forces its way through a series of narrow passes among enormous granite boulders. The waters fall in a succession of rapids, ending with the Raudal de Atures (Atures Rapids). In this region, the main tributaries are the Vichada and Tomo rivers from the Colombian plains, and the Guayapo, Sipapo, Autana, and Cuao rivers from the Guiana Highlands.

The Raudal de Atures marks the beginning of the Lower Orinoco Basin, in which the river makes its great bend to the east. In this section, the river flows slowly through the lowest level of the plains and increases to about five miles in width. Along the bend, it receives the largest number of affluents of its entire course, including the Meta, Arauca, and Capanaparo rivers. The Río Apure contributes waters from numerous Andean streams, which form a swampy maze in their lower courses.

From its junction with the Río Apure, the Orinoco meanders eastward over gently sloping plains. Shoals and alluvial islands are abundant; some of the islands are



The Orinoco River Basin and its drainage network.

large enough to divide the channel into narrow passages. Tributaries include the Guárico, Manapire, Suatá, Pao, and Caris rivers, which enter on the left bank, and the Río Cuchivero and the Río Caura, which join the main stream on the right. So much sediment is carried by these rivers that they often form islands at their mouths. The Río Caroní, one of the Orinoco's largest tributaries, joins the river on its right bank after plunging over the Saltos del Caroní (Caroní Falls) above Ciudad Guayana. Many lagoons, including the Mamo, Amana, and Colorada, are located on the banks of the Orinoco west of its confluence with the Río Caroní and east of Ciudad Bolívar. At the town of Barrancas, the river begins to form its great delta.

The delta region

The delta extends for about 275 miles along the Atlantic coast, from Pedernales on the Gulf of Paria in the north to Punta Barima in the south on the Boca Grande (literally the "Great Mouth"). Scores of islands are connected by innumerable *caños*, or canals, which constitute an intricate network. The main channel of the Orinoco, known in the delta as the Río Grande, flows eastward from Barrancas to discharge into the Boca Grande.

Climate. The climate is tropical, with the seasons marked by differences in rainfall rather than in temperature. The year is divided into two seasons, the rainy and the dry (locally known as winter and summer), the former extending from April to October or November, the latter most marked from November through March or April. Temperature differences, on the other hand, are slight throughout the year; and no month averages more than 69° F (21° C) or less than 64° F (18° C). Whatever the average temperature, there is little difference from month to month. The only marked variation is from day to night, being greater than that from month to month.

Rainfall varies considerably from district to district. The northeast trade winds blow across the coastal districts without losing much precipitation, in some places leaving less than 20 inches per year. Areas lying behind topographic barriers also get little rain, while windward slopes are generally well watered. In some areas enough rain falls to support a lush jungle growth and, in others, enough for a true selva (rain forest). The Llanos suffer severely from drought from about January to April, and then suffer equally from flooding of the whole countryside from June to October.

Hydrology. The river basin, as a geomorphological feature, dates from the Quaternary Period (from 2,500,000 years ago to the present). The enormous quantities of material produced by highlands are carried down by torrential rains to the rivers. The rivers, unable to hold the excessive material, overflow or break their banks, producing periodic floods that submerge the lowlands.

The drainage pattern

Under these conditions, drainage presents an unstable and indefinite pattern, marked by the shifting of rivers, lagoons, and swamps over the lower lands. The Orinoco Delta is rapidly extending into the ocean, but the tremendous amounts of sediment that accumulate are accelerating the subsidence occurring in the entire delta region.

Wide fluctuations in the river's flow reflect the seasonal rainfall regime. During the dry season or "low-water" period from October to March, the average river depth is about 49 feet in the lower basin near Ciudad Bolívar. The rise of the Orinoco begins with manifest regularity in April at the beginning of the rainy season. The "high water" period from April to October reaches its maximum in July. The depth of the river at this period is about 165 feet at Ciudad Bolívar. From June to August the lowlands of the basin are flooded and in some places are 65 feet under water. At the end of August the waters gradually recede and reach their lowest point in October.

Vegetation. The Llanos are grass-covered plains with isolated stands of palms and chaparro (scrub oak). The tributary streams that cross the plains have deposited alluvial soils, and strips of forests line their banks. Some of this natural tree cover, however, has been reduced by deforestation. The Guiana Highlands are covered with high, dense forest that is interrupted by small patches of savanna (grassy parklands). The tropical rain forest of the Upper Orinoco Valley contains hundreds of species of trees. Rain forest also covers the delta region.

Animal life. More than a thousand species of birds frequent the Orinoco region; among the more spectacular are the scarlet ibis, the bell bird, the umbrella bird, and numerous parrots. The numerous fish include the voracious caribe (piranha) and the *laulao*, a catfish that often weighs more than 200 pounds. The Orinoco crocodile is the longest of its kind in the world, reaching a length of more than 20 feet. The array, or side-necked turtle, which grows to a shell-length of about 30 inches, nests on the sandy islands of the river.

The human imprint. *The population.* Except for the Guajiros of Lake Maracaibo, all of the Venezuelan aboriginal population lives within the Orinoco Basin. The most important indigenous groups include the Guaicas (Waicas), also known as the Guaharibos, and the Maquiritares (Makiritares) of the southern uplands, the Guarauno (Warrau) of the delta region, and the Guahibos and the Yaruros of the western Llanos. These peoples live in intimate relationship with the rivers of the basin, using them as a source of food as well as for purposes of communication.

Indigenous peoples of the basin

The important towns, with the exception of Ciudad Bolívar, are built on high ground to avoid recurrent flooding. Town plans reflect Spanish influence: streets are arranged in a chequered pattern with a central plaza. Most of the original colonial buildings have vanished, although some old homes still remain. In the small villages, thatched huts are the usual dwellings.

History of exploration. European exploration of the Orinoco River Basin began in the 16th century. A series of expeditions sponsored by the banking house of Welser of Augsburg penetrated the Llanos southward across the Apure and Meta rivers. From the east, several Spanish expeditions ascended the river from its mouth without much success. In 1531 the Spanish explorer Diego de Ordaz voyaged up the river, and that same year another Spanish explorer, Antonio de Berrio, descended the Casanare and Meta rivers and then descended the Orinoco to its mouth.

In 1744 Jesuit missionaries discovered the Brazo Casiquiare. Alexander von Humboldt, the German naturalist, travelled over 1,700 miles of the river basin in 1800. By 1860 steamships were navigating the Orinoco. The source of the river remained in dispute, however, until a Venezuelan expedition finally identified it in 1951.

Navigation and river crossings. The Orinoco and its tributaries have long served as vast waterways for the indigenous inhabitants of the Venezuelan interior. Especially during the floods of the rainy season, the canoe with outboard motors is the only means of communication throughout large areas of the river basin. Large river steamers travel upriver for about 700 miles from the delta to the Raudal de Atures. Dredging has allowed large oceangoing vessels to navigate the Orinoco from its mouth to its confluence with the Río Caroní—a distance of 226 miles—in order to tap the iron-ore deposits of the Guiana Highlands. The Llanos and the Guiana region were connected in 1967 with the completion of the 5,507-foot bridge across the Orinoco at Ciudad Bolívar. In 1961, the Venezuelan government bridged the mouth of the Río Caroní to connect the new industrial town of Puerto Ordaz with the old Orinoco port of San Félix, thereby creating the urban unit of Ciudad Guayana.

Economic development. The Guiana Highlands are rich in mineral deposits. Iron ore, which is more than 60 percent pure, is mined at Cerro Bolívar and El Pao. Numerous other mountain deposits constitute proved reserves totalling about 1,567,000,000 tons.

Other minerals include deposits of manganese, nickel, vanadium (a metallic element used to form alloys), and chrome. There are also deposits of gold and diamonds. Petroleum and natural gas are exploited in the northern and northwestern parts of the basin.

The Saltos del Caroní has the largest hydroelectric potential in South America, amounting to 10,000,000 kilowatts. A vast project to tap this power source includes the Macagua Dam, in use since 1960, and the Guri Dam, planned as one of the world's largest hydroelectric plants, with a capacity of 6,000,000 kilowatts.

The Llanos, along the great bend of the Orinoco, have long formed one of Venezuela's major cattle-raising areas. Seasonal subsistence farming is carried out along the river from the confluence with the Apure to Ciudad Bolívar. West of the Apure junction, cotton is grown on a commercial scale. Land reclamation and flood control projects in the delta region are planned in order to open vast agricultural lands.

The Ciudad Guayana complex

Industrial development of the river basin is concentrated at Ciudad Guayana. The semi-autonomous government body Corporación Venezolana de Guayana (Venezuelan Guiana Corporation) is conducting a careful development plan based on the resources of the nearby Guiana Highlands and the Río Caroní. The dams at Saltos del Caroní supply power for industry, including a steel mill, an aluminum plant, and a paper factory. Power is also supplied by natural gas, which is piped from the oil fields north of the Orinoco River. There are plans for a timber and furniture industry, as well as for other industries that would make use of electrochemical processes.

(Me.F.G.)

Orissa

Orissa, a state of the Indian Union, is situated in eastern India. It is bounded by the Bay of Bengal in the southeast; by West Bengal in the northeast; by Bihār in the north; by Madhya Pradesh in the west; and by Andhra Pradesh in the south. Its area of 60,178 square miles (155,860 square kilometres) and population of about 22,000,000 (1971), represent about 5 percent of the area and about 4 percent of the population of India, respectively. Many people speaking Oriya (the language of Orissa) also live in the adjoining areas of neighbouring states. Before India became independent in 1947 Orissa's capital was at Cuttack. The present capital was subsequently built at Bhubaneswar, in the vicinity of its historic temples.

The land corresponding roughly with modern Orissa, but at times much larger in area, passed under the names of Utkala, Kaliṅga, and Oḍra Deśa in ancient and medieval times. In origin, these names were associated with peoples. The Okkalā or Utkala, the Kaliṅgā, and the Oḍra or Oḍḍakā were mentioned in literature as tribes. Ancient Greeks knew the latter two as Kalingai and Oretes. Ultimately the names became identified with territories. For centuries before and after the birth of Christ, Kaliṅga was a formidable political power, extending from the Ganges River to the Godāvari. The great war of the Indian emperor Aśoka (*c.* 265–238 BC) was called the Kaliṅga War; Khāravela's extensive empire (1st century BC) was known as the Kaliṅga Empire; and even the ancient Hindu colonies in Malaya and the East Indies were named Kaliṅga. Approximately between the 11th and 16th centuries the name fell into disuse; instead, the name Oḍra Deśa was gradually transformed into Uḍḍiśa, Uḍisā, or Oḍisā, which in English became Orissa. The language of Oḍisā came to be known as Oriya or Oria.

The British did not conquer the entire Oriya-speaking area at the same time; nor did they administer the whole area, when conquered (1803), as one unit. Consequently, there arose a demand for unification on a linguistic basis, and as a result Orissa was constituted a separate province on April 1, 1936. Even then 26 Oriya princely states remained outside the provincial administration, but after 1947 all those states, except Saraikela and Kharsāwān (which merged with Bihār), were included in Orissa. The Constitution of India (1950) recognized Orissa's statehood.

Linguistic basis of unification

History. At the dawn of Indian history Kaliṅga was already famous. Buddhist sources refer to the rule of King Brahmadatta in Kaliṅga at the time of Buddha's death. In the 4th century BC the first Indian empire builder, Mahāpadma Nanda, conquered Kaliṅga, but the Nanda rule was short-lived. In 261 BC Aśoka invaded Kaliṅga and fought one of the greatest wars of ancient history. He then renounced war, became a Buddhist, and preached peace and nonviolence in and outside India. In the 1st century BC the Kaliṅga emperor Khāravela conquered vast territories. In the early centuries of the Christian Era Kaliṅga was a maritime power. Its overseas activities culminated in the 8th century AD with the establishment of the Śailendra Empire in Java. During the 8th, 9th, and 10th centuries Orissa was ruled by the powerful Bhauma-Kara dynasty, and in the 10th and 11th centuries by the Soma dynasty. The Temple of Liṅgarāja at Bhubaneswar, the greatest Śaiva monument of India, was begun by the Soma King Yayāti.

The Gaṅga dynasty

Medieval Orissa enjoyed a golden age under the Gaṅga dynasty. Its founder, Anantavarma Cōḍagaṅgadeva (1078–1147), ruled from the Ganges to the Godāvari with Cuttack as his capital. He began the construction of the Temple of Jagannātha (Lord of the Universe) at Puri. Narasiṃhadeva I (1238–1264) built the Sun Temple of Konārak, widely acknowledged as the finest specimen of Hindu architecture. In the 13th and 14th centuries, when the whole of India was overrun by the Muslims, independent Orissa remained a citadel of Hindu religion, philosophy, art, and architecture. The Gaṅgas were succeeded by the Sūrya dynasty. Its first king, Kapilendradeva (1435–1466), won territories from his Muslim neighbours and greatly expanded the Orissan kingdom. His successor, Puruṣottamdeva, maintained these gains with difficulty. The next and the last Sūrya king, Pratāprudradeva, became a disciple of Caitanya, the great medieval saint, and became a pacifist. After his death (1540) Orissa declined, and 1568, when King Mukundadeva was killed by his own countrymen, Orissa lost its independence to the Afghān rulers of Bengal.

The Mughal emperor Akbar conquered Orissa from the Afghāns. When the Mughal Empire fell, part of Orissa remained under the Bengal Nawabs, but the greater part passed to the Marāthās. The Bengal sector came under British rule in 1757 after the Battle of Plassey; The Marāthā sector was conquered by the British in 1803. Meanwhile, the inaccessible areas of Orissa remained under princely rulers. Orissa assumed its present form with the inclusion of the princely states after 1947.

The landscape. Physiographically, Orissa is of a heterogeneous character. Its geological formations vary from the oldest rocks of the Earth's crust in the stable landmass of the Indian Peninsula to deltaic alluvium or littoral deposits and ridges of windblown sand on the seaboard.

The four geographic regions

Broadly, the state can be divided into four natural divisions: (1) the Northern Plateau, (2) the Eastern Ghāts, (3) the Central Tract, and (4) the Coastal Plains. The Northern Plateau is an extension of the Chota Nāgpur plateau covering the Mayūrbhanj, Keonjhar, and eastern Sundargarh areas. The upper reaches of the rivers Subarnarekha, Baitarani, and Brāhmani lie in the area. Its central portion contains many small hills and forests. The Eastern Ghāts form an undulating plateau that extends over Koraput, parts of Kalāhāndi, Phulbāni, and Ganjām. The Central Tract covers the districts of Sambalpur, Bolāngir, and Dhenkānāl, and parts of Kalāhāndi, Sundargarh, and Phulbāni. The rivers Mahānadi, Brāhmani, and Baitarani and their tributaries flow through the area. The landscape consists of a succession of plateaus, hills, uplands, and valleys. The coastal plains comprise the Balasore, Cuttack, and Puri districts, and parts of Ganjam. This fertile area runs parallel to the coast and contains the river deltas; it is formed of alluvium and silt.

The main rivers are the Subarnarekhā, Burābalang, Baitarani, Brāhmani, Mahānadi, Rushikulya, and Vansadhara. Notable mountain ranges are the Mahendra Giri (rising to 5,000 feet), the Malyagiri (3,896 feet), and the Meghasana (3,823 feet). Orissa's Chilka Lake is the biggest saltwater lagoon in India.

Climate. Orissa is situated in the hot climatic belt where the monsoon winds blow. The main seasons are summer from mid-February to June, the rainy season from July to October, and winter from November to mid-February. Places in Sambalpur and Mayūrbhanj districts have an extreme climate, whereas in places like Puri and Gopālpur the climate is equable. The average rainfall is about 60 inches.

Vegetation. The vegetation is of a tropophilous character—that is to say it is able to adapt itself to a rapidly

changing environment. Forests cover about 40 percent of the total land area. In forest wealth, Orissa occupies the third place in India. The forests are of three broad types: northern tropical semievergreen; northern tropical moist deciduous; and northern tropical dry deciduous. The first group includes the mango; the second group includes, for example, teak; while the third includes bamboo.

Animal life. Orissa has the same animal life as the rest of peninsular India. Monkeys are common. Carnivores include different types of tigers. The elephant, the wild buffalo, the blackbuck, and the four-horned antelope are found in some districts. The peafowl is one of the features of the Orissa forests. Lake Chilka is noted for its marine fauna.

Population. The population includes the Australoid, the Alpinoid, and the Mediterranean racial groups. The Australoids are represented mainly by the Saora, Santāl, Khond, Gond, Munda, Ho, Juang, Gadaba, Bhuiyan, Koya, and Oraon tribes. These tribes are divided into three linguistic groups: the Munda-speaking (the Santāl, Saora, Juang, and others); the Dravidian-speaking (the Khond, Gond, Oraon, and others); and the Oriya-speaking (the Bhuiyan, and others). Most tribes live in the hill areas, but they are also found in the plains. The Alpinoid and Mediterranean groups are represented among the nontribal population, which is mainly Oriya-speaking and Hindu.

The tribes have for a long time been going through the process of Hinduization. Tribal chieftains have claimed Kṣatriya (warrior) status, while many of the Khonds, who constitute the largest tribe, have abandoned their Kui language (Dravidian) and speak Oriya. Many tribes are bilingual. Some have become almost indistinguishable from the Hindus and have lost their original language. The Oriya language does not vary in its written form, but it differs in the spoken form from place to place. The purest Oriya is spoken in the Cuttack and Puri districts. The Balasore, Sambalpur, and Ganjam districts have distinct local accents.

The caste structure

More than 97 percent of the people are Hindus. The caste structure is the same as in other states of eastern India, with some regional variations. Next to the Brahmins are the Karaṇas (the writer class), who claim Kṣatriya status (with the pen as their weapon rather than the sword). The Khandayats (literally, "swordsmen") are mostly cultivators but call themselves "Khandayat-Kṣatriyas." All castes look to Jagannātha (one of the forms of the Hindu god Viṣṇu) as the centre of their religious faith. For centuries Puri, the abode of Jagannātha, has been the only place in India where all castes, including the so-called untouchables, eat together.

Orissa has a predominantly rural population. In 1971, only 9 percent of its people lived in towns. The building of the new capital at Bhubaneswar, of a steel town at Raurkela, of a port at Paradip, of a dam at Hirākud, and of a power-generating station at Tālcher have necessitated new urban settlements, and a number of industrial towns are under construction. There were 79 towns including four cities—Cuttack, Raurkela, Bhubaneswar, and Berhampur—in 1971, and the number of villages exceeded 46,000

Administration. The head of the state is a governor appointed by the president of India. The actual administration is conducted by a council of ministers, headed by a chief minister and responsible to the elected legislature, which consists of only one chamber of 140 members elected at intervals of not more than five years through universal adult suffrage.

The districts of Orissa

There are 13 districts: Bālāsore, Bolāngir, Cuttack, Dhenkānāl, Ganjām, Kalāhāndi, Keonjhar, Koraput, Mayūrbhanj, Phulbāni, Puri, Sambalpur, and Sundargarh—grouped into three revenue divisions, each under a divisional commissioner. A board of revenue is in charge of revenue administration. The district administration is conducted by a deputy commissioner who is also the district magistrate. Each district has a superintendent of police. The districts are divided into *tahsils*, each having a *tahsildar* as its revenue officer. *Tahsils* comprise groups of villages, administered by *pañcāyats* (village committees), to which villagers elect their representatives. A *sarpañc* (elected president) heads the *pañcāyat*. The system represents a democratic decentralization of power for the benefit of the rural population. The towns are administered by municipalities.

Social conditions. *Health.* At one time the coastal belt was highly malarious and the whole state was subject to epidemics of cholera and smallpox. The incidence of filariasis (a disease caused by the presence of filarial worms in the blood and glands), leprosy, and tuberculosis was high. The high rate of mortality even led to a decrease of population in certain places. During the first three Indian five-year plans (1951–66) much attention was paid to health services, and excellent results were achieved by various programs.

By 1971, rural Orissa possessed 314 primary health centres; the number of hospitals and dispensaries had increased; and the three medical colleges, at Cuttack Berhampur, and Burla, had expanded considerably. For an estimated 22,000,000 population in 1970, the number of doctors in government service was 1,550, and the number of registered doctors about 3,400.

Education. The number of educational institutions increased considerably after 1947. In 1970 there were four universities and 73 colleges for general education, with an enrollment of 33,000 students; 5,200 secondary schools with 451,000 students; and 26,208 primary schools with 1,898,000 students. About 26 percent of the population was literate in 1971.

Welfare. Orissa has a Tribal Welfare Department. Schemes have been devised to promote the educational, cultural, economic, and social advancement of the tribes. The state Social Welfare Advisory Board, instituted in 1954, cares for the welfare of women and children through courses of instruction, urban-welfare-extension projects, and holiday camps for children.

Economy. About 80 percent of the rural population depends on agriculture, even though some areas are unproductive and some others are unsuitable for more than a single annual crop. The nonagricultural classes live by handicrafts, by trade, and by the rendering of services. About 6 percent of the rural families depend on village or cottage industries, including weaving.

More than 15 percent of the agricultural families engage in nonagricultural pursuits by way of subsidiary employment. Whatever their occupation, most rural people do not get continuous employment the year around. The amount of agricultural land available per person has been declining because of the growth of population. The per capita area under cultivation was only 0.83 acres in 1950. By 1970, it was further reduced, and consequently the search for new lands for cultivation, the partial abolition of large landholdings, and the distribution of available land among landless families became acute problems for the government.

Mineral deposits

The industrial resources of Orissa are considerable. As most mineral deposits are in the former feudatory states, their exploitation began only after 1947. In the early 1970s Orissa was ahead of all other Indian states in the production and export of iron ore and chromite and took second place in the production of manganese. It was producing annually nearly 4,000,000 tons of iron ore, 350,000 tons of manganese ore, 90,000 tons of chromite, 850,000 tons of coal, and more than 2,200,000 tons of limestone and dolomite. It also had rich deposits of graphite and bauxite. The reserves of Orissa iron ore were estimated at 5,900,000,000 tons, while an additional 4,000,000,000 tons also probably existed.

Large-scale industries in the early 1970s included: a steel plant at Raurkela with an annual production capacity of 1,000,000 tons of ingots; a ferromanganese plant at Joda in Keonjhar, with a capacity of 30,000 tons; a ferromanganese plant at Rāyagada (capacity 12,000 tons); a refractory at Rajgāngpur (capacity 80,000 tons); another refractory at Belpahar (capacity 124,000 tons); a tube works at Chowdwar (capacity 36,000 tons); a low-shaft furnace at Barbil, Keonjhar, and Chowdwar (capacity 100 tons per day each), for production of pig iron and ferroalloys; a cement factory at Rajgāngpur (capacity

725,000 tons); and an aluminum smelter at Hīrākud (capacity 10,000 tons).

Other industries included: paper mills at Brajrajnagar Chowdwar, and Singhpur; a fertilizer plant at Raurkela; a socket-manufacturing plant at Cuttack; a textile mill at Chowdwar; sugar factories at Rayagada and Aska; a chlorine and caustic soda factory at Brajrajnagar; and a glass and ceramic factory at Bārang. The Hīrākud dam project for the supply of electric power envisages an ultimate effective capacity of 200,000 kilowatts. The Machkund hydroelectric system is another source of power.

Transportation and communications. Communication facilities were undeveloped before 1947, but the merger of the feudatory states with Orissa and the discovery of mineral resources necessitated the construction of a network of good roads. During the five-year plans, bold construction programs were undertaken and bridges were built over most of the major rivers, but the state still lacks adequate railway communications. It has only 14 miles of rail track for every 1,000 square miles, against an all-India average of 33.

The port of Paradip

An all-weather, sheltered, deep-draft port has been constructed at Paradip at the mouth of Mahānadi River, whence it is intended to export annually 5,000,000 tons of iron ore. The port has an extensive hinterland with mineral deposits and a developing industrial belt.

Cultural life. Orissa has a rich artistic heritage and has produced some of the best examples of Indian art and architecture. Artistic traditions are maintained through mural paintings, stone carving, wood carving, icon paintings (known as *patta* paintings), and paintings on palm leaves. Handicraft workers are famous for their exquisite silver filigree ornamentation and decorative work.

In tribal areas Orissa has a wide variety of folk dances. The music of the *madal* and flute is common in the countryside. The classical dance of Orissa, known as the *orissi* dance, has survived for more than 700 years. Originally it was a temple dance, performed before gods. The modes, movements, gestures, and poses of the dance are depicted on the walls of the great temples, especially at Konārak, in the form of sculpture and in relief carvings. Modern exponents of the dance have made it popular outside the state. The *chhau* dance (performed by groups of masked dancers) of Mayūrbhanj and Saraikela regions is another feature of Oriya culture. For the promotion of dancing and music, the Kala Vikash Kendra centre was founded at Cuttack in 1952 with a six-year teaching course. The National Music Association serves a similar purpose. Other notable dance and music centres in Cuttack are the Utkal Sangit Samaj, the Utkal Smruti Kala Mandapa, and the Mukti Kala Mandir.

The Boita-Bandana ceremony

There are many traditional festivals. A unique festival is the ceremony of Boita-Bandana (worshipping of ships). In October–November for five consecutive days before the full moon, people gather near riverbanks or the seashore and float miniature boats, as a symbolic gesture that they will leave for the faraway lands (Malaysia and the East Indies) to which their ancestors once sailed. The greatest religious festival of Orissa is the Car Festival of Jagannātha at Puri, which attracts hundreds of thousands of people from all parts of India.

BIBLIOGRAPHY. General and historical works include R.D. BANERJI, *History of Orissa*, 2 vol. (1930–31); M.M. GANGULY, *Orissa and Her Remains* (1912); H.K. MAHATAB, *History of Orissa*, 2 vol. (1950); S. MISRA, *Economic Survey of Orissa*, 2 vol. (1960); G.C. MOHAPATRA, *The Stone Age Cultures of Orissa* (1962); S.N. RAJGURU, *Inscriptions of Orissa*, 4 vol. (1958–60); and the UTKAL UNIVERSITY, *History of Orissa*, vol. 1 (1964) and *Orissa, Past and Present* (1962). See also the *District Census Handbooks (Orissa)*, 13 vol. (1965), which contain detailed statistics relating to economic conditions, agriculture, administration, and education; and the *Techno-Economic Survey of Orissa* (1961), a publication of the government of India that includes socio-economic data.

(M.N.D.)

Orozco, José Clemente

Considered by many to have been the pre-eminent artist of modern Mexico, José Clemente Orozco was also one of the most powerful modern exponents of public-scale art. He helped found the mural movement in 20th-century Mexico and used that form of painting as an instrument of his social conscience. Orozco's art constituted a rare combination of intense, personal expressionism with simplified synthetic forms. His artistic style, as well as his aesthetic philosophy, influenced many artists of the Americas throughout the second quarter of the 20th century.

"Self-Portrait," tempera on cardboard, by Orozco, 1940. In the Museum of Modern Art, New York.
By courtesy of the Museum of Modern Art, New York

Early life and training. Orozco was born on November 23, 1883, in Ciudad Guzmán, also known as Zapotlán el Grande, in the state of Jalisco, Mexico. Believing themselves to be descendants of leading *conquistadores* of western Mexico, his family was prominent in Jalisco. Orozco first became obsessed with art in 1890, when his family moved to Mexico City. Going to and from school each day, he paused in the open workshop of José Guadalupe Posada, Mexico's first great printmaker, whose grotesque caricatures and illustrations appeared in sensational news sheets devoted to reporting lurid crimes and political scandals. Orozco was captivated by Posada's strong images and vivid style and for the rest of his life acknowledged the early influence of the master engraver.

Orozco's prodigious skill was immediately recognized, and he began night classes in drawing at the Academia de San Carlos. The future social critic tolerated no censure of his work, and his mother frequently had to defend him against charges of sacrilege and anticlericism. Toward the end of the decade, his pursuit of art was interrupted when he was forced to study for careers as an agronomist and, later, as an architectural draftsman. When he was 17, however, he lost his left hand in a laboratory accident, and he abandoned his architectural studies for painting. He re-entered the Academia de San Carlos in 1905 with a renewed passion for painting and set about assiduously to become a competent painter. Unlike many of his fellow students, he enjoyed the academy's rigorous discipline, which demanded that students draw a model repeatedly from the same view until a drawing was produced realistic enough to be favourably compared to a photograph of the subject.

Influence of Gerardo Murillo

More important to the ultimate meaning of Orozco's art was his acquaintance with a radical student named Gerardo Murillo. Violently opposed to the cultural anti-Mexicanism in vogue, Murillo assumed the Aztec name of Doctor Atl and urged artists to reject the cultural domination of Europe and to cultivate in their work Mexican traits. Accordingly, Orozco began conscientiously to explore Mexican themes and to draw more directly from scenes of daily life. In 1910 Orozco and other young artists working with Atl organized Mexico's first exhibition of Mexican artists, and the same year they formed an association called the Artistic Centre, which proposed to get permission to paint murals on the walls of government buildings. The organization was dissolved the same year, however, after disorders of the Madero Revolt made large-scale artistic activity impossible. With

the academy closed by student strikes, Orozco became a caricaturist for an opposition paper and haunted the barrios, or slums, of Mexico City, painting a series of watercolours dealing with the lives of prostitutes. The series, collectively titled "House of Tears," showed the mordant colours of his early Expressionist style and marked his initial use of the prostitute as a symbol of human degradation. When the academy reopened, it was dominated by would-be Impressionists, and Orozco left in disgust to continue developing his own style.

Again, in 1914, civil war broke out in Mexico. Orozco did not take an active part in the fighting but supported the forces of Gen. Venustiano Carranza as a satirical artist on the revolutionary paper *La Vanguardia* ("The Vanguard"), which was edited by Atl. Though a noncombatant, Orozco witnessed all the bestiality and excess that accompanies war, an experience that indelibly stamped his mind and art.

Mature work and later years. In 1917 the reaction of critics and moralists to the exhibition of his "House of Tears" paintings forced Orozco to leave Mexico for the United States, where he lived for several unhappy years in San Francisco and New York City. On his return to Mexico in 1920, he found the new government of Pres. Alvaro Obregón eager to sponsor his work, and, along with his colleagues Diego Rivera, David Alfaro Siqueiros, and others, he was commissioned to paint murals on the walls of the Escuela Nacional Preparatoria, initiating the so-called Mexican muralist movement. Orozco's early (1923–27) murals there, such as "Maternity" and "Christ Destroying His Cross," were derivative, and Orozco himself destroyed many of them. Those dating from 1926 however, show him coming into his own style, achieving in such works as "Cortés and Malinche" and "The Trench" (both in the Escuela Nacional Preparatoria) a monumentality unprecedented in Mexican art.

In 1927 government patronage and protection were withdrawn from Orozco and his fellow muralists, and the subsequent attacks of critics and moralists or conservatives again forced him to flee to the United States. Humiliated in his own country, he consciously strove, after settling in New York City, to forge an international reputation that would force his countrymen to recognize his value as an artist. By frequenting the salon of a prominent Manhattan art patron (whose circle included the poet Khalil Gibran), he slowly became known in American art circles and finally was commissioned in 1930 to paint a major mural in the refectory of Pomona College, Claremont, California. In choosing to do a mural of Prometheus, Orozco temporarily abandoned social criticism and historical subjects in favour of a more universal theme: the self-sacrificing titan from ancient Greek mythology, bringing man fire, which enlightens, liberates, and purifies but also consumes. Orozco also turned away from the relative stylistic repose of the Escuela Nacional Preparatoria murals. Recalling Atl's drawings and enthusiastic descriptions of the tortured figures in Michelangelo's "Last Judgment" in the Sistine Chapel, he portrayed Prometheus as a monumental pseudo-Michelangelesque giant, straining his powerful muscles against the burden of his fate. By contrast, his murals at the New School for Social Research in New York City, dealing with the themes of universal brotherhood and social revolution, suffer by the slavish use of "dynamic symmetry," a theory fashionable in the 1920s, which purported to represent the ancient Greek system of proportions.

Work in the United States

In 1932 Orozco made a brief trip to Europe, where he viewed the art of England, France, Spain, and Italy. Although he was impressed with the paintings of Picasso, his even deeper admiration of the Byzantine mosaics of Rome and Ravenna is reflected in his great series of murals at Dartmouth College, Hanover, New Hampshire. Just as the Byzantine mosaics illustrate the Christian concept of history, Orozco illustrated his world view in a vast scheme divided into two series of murals correlated to the two main scenes, "The Coming of Quetzalcoatl" and "The Return of Quetzalcoatl." This dichotomy contrasted the stages of man's progression from a primeval, non-Christian paradise to a Christian, capitalist hell. Byzantine mosaics also clearly influenced the pictorial style of "Modern Migration of the Spirit," but such scenes as "Stillborn Education" and the Quetzalcoatl murals achieve unique levels, respectively, of grotesqueness and of sweeping force.

Return to Mexico

His art enriched and matured and his reputation firmly established, in 1934 Orozco returned triumphantly to Mexico, where he completed the illustration of his view of history in the mural "Catharsis." This eschatological work displayed a laughing prostitute lying among the debris of civilization's last cataclysm, showing the constantly increasing pessimism that culminated in his Guadalajara murals. Murals painted in the lecture hall of the Universidad de Guadalajara, the Palacio de Gobierno (1937), and the chapel of the orphanage of Hospicio Cabañas (1938–39) recapitulate the historical themes developed at Dartmouth and in "Catharsis" but with an intensity of anguish and despair he never again attempted. Here, history is blindly careening toward Armageddon. The only hope for salvation is the self-sacrificing creative man, the "Man of Fire," painted in the Hospicio dome.

Orozco's subsequent murals, such as those in the Gabino Ortíz Library, in the Palacio de Justicia, and in his last great work, "National Allegory" (1947–48; Escuela Normal, Mexico City), neglect universal themes and dwell almost exclusively on nationalism. Canvases such as "Metaphysical Landscape" (1948; estate of Orozco, Mexico City), however, hint at a growing mysticism, and its abstract style indicates that Orozco may have been on the brink of nonfigurative painting when he died. His easel paintings, such as "Zapatistas" (1931; Museum of Modern Art, New York City), often attain the grandeur of his murals, which remain his definitive vehicles of expression, the touchstones of his genius.

Orozco, whose childhood and youth were filled with struggle and persecution, became a national hero in his later years, honoured as the leader among those who raised Mexican art to a position of international eminence. In 1947 the President of the Republic of Mexico awarded him a prize as the outstanding Mexican figure in the arts and sciences during the preceding five years. He died September 7, 1949.

MAJOR WORKS

Murals (1923–27; Escuela Nacional Preparatoria, Mexico City); murals (1926; Casa de los Azulejos, Mexico City); murals (1926; Escuela Industrial, Orizaba, Mexico); murals (1930; Pomona College, Claremont, California); murals (1930–31; New School for Social Research, New York); murals (1932–34; Dartmouth College, Hanover, New Hampshire); mural, "Catharsis" (1934; Palacio de Bellas Artes, Mexico City); murals (1936–39; Universidad de Guadalajara, Palacio de Gobierno, Hospicio Cabañas); murals (1940; Gabino Ortíz Library, Jiquilpan, Mexico); fresco panels, "Dive Bomber and Tank" (1940; Museum of Modern Art, New York); murals (1941; Palacio de Justicia, Mexico City); murals (1942–44; Hospital de Jesús Nazareno, Mexico City); murals (1947; Escuela Normal, Mexico City); mural (1948; "Juárez" Museo Nacional de Historia, Mexico City); murals (1949; Cámara Legislativa, Guadalajara, Mexico).

BIBLIOGRAPHY. JOSE CLEMENTE OROZCO, *Autobiografía* (1945; Eng. trans., 1962), essential to an understanding of Orozco; JUSTINO FERNANDEZ, *José Clemente Orozco: forma e idea* (1942), especially important for its insights into Orozco's art and aesthetic; MACKINELY HELM, *Man of Fire: J.C. Orozco* (1953, reprinted 1971), the earliest major work in English—still considered a classic; ALMA REED, *Orozco* (1956), a biographical account written by an American who had known the artist for many years.

(Ed.)

Orthopteran

Orthopteran has come to be regarded as a common name for several related groups of insects that exhibit considerable morphological, physiological, and paleontological diversity. Although sometimes these insects are combined into the order Orthoptera, generally, several orders are implied in the term orthopteran. Among the orthopterans, cockroaches and mantids are placed in the order Dictyoptera; the grylloblattids (order Grylloblattodea) and walking sticks (order Phasmida) are given ordinal

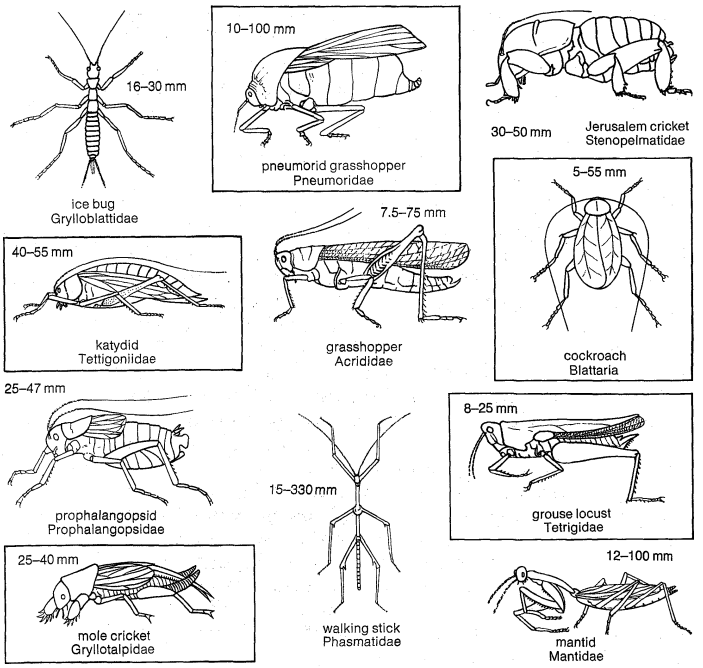


Figure 1: Diversity among orthopterans.

rank also. On the other hand, members of the suborders Ensifera (katydids, crickets, and camel crickets) and Caelifera (pygmy sand crickets, grasshoppers, and locusts) are considered to comprise the order Orthoptera. For completeness of discussion, all of these groups, handled here as four separate orders, are included in this article.

Orthopterans, abundant in tropical regions throughout the world in both numbers of species and individuals, are common in the summer months in temperate regions, when their relatively large size and chirping sounds attract considerable attention. Zoologists have long been interested in cockroaches, one of the oldest insect groups known. Most of the 24,000 species of orthopterans are plant feeders, with mouthparts adapted for chewing. Locusts, known as pests since biblical times, are very destructive to agricultural products.

GENERAL FEATURES

Orthopterans may be bizarre in appearance, unusually large in size, or show peculiar behaviour. They range in size from a few millimetres to more than 30 centimetres. Some tropical walking sticks resembling tree twigs are more than 30 centimetres long, and others, much smaller, resemble leaves of plants. The size range of present-day cockroaches is typical of the diversity of body size among orthopterans: tiny flightless cockroaches (*Attaphila*), living as commensals in the nests of ants, are only two millimetres long when mature, whereas a species of *Megaloblatta* found in South America reaches 10 centimetres in length with a wing span of almost 19 centimetres.

Distribution and abundance

Approximately 24,000 species of orthopterans have been identified. Throughout the U.S. there are about 1,300 species; not all of them inhabit any one place. There are more species (*e.g.*, 282 in Arizona) in southern and southwestern sections of the U.S. than in the North (*e.g.*, 103 in all the New England states). Of the 600 species found in Europe, Great Britain has only 35, including four established, introduced species (adventives).

The largest families of orthopterans are worldwide in range, although all have decreased numbers of species in cold temperate zones. Few mantids or walking sticks, for example, occur outside tropical or subtropical areas. There are about 20 mantid species and 27 walking stick species in the southern U.S., compared with 400 mantid species and 600 walking stick species in Central and South America. A few northern groups include the gryllobla ttids and several genera of grasshoppers.

Importance. Among the orthopterans are many species that are either harmful to agricultural products or are pests. Grasshoppers are capable of causing widespread devastation of the agricultural crops grown in many countries throughout the world. In cattle-growing regions there often is competition between grasshoppers and livestock for available forage. Mormon crickets (a common name for species of the genus *Anabrus* that originated during the early years of the Mormon settlement in Utah) are major pests of both crops and open rangeland in the western part of the U.S. during seasons that are favourable for their development. Cockroaches, known throughout the world as domestic pests, are a frequent nuisance, especially in warm-temperate to tropical areas. Although cockroaches occasionally carry organisms such as bacteria or parasites that produce intestinal diseases, they are more generally considered to be mechanical carriers of contaminating filth.

Mantids, predators on other insects, have become adapted to resemble the flowers, tree trunks, or grass stems on which they await their prey. Crickets, katydids, and grasshoppers are known for the songs they produce using stridulatory mechanisms, and research concerned with song production is an active field. The biology of migratory grasshoppers or locusts involves hormones that promote transformation of nonmigratory, solitary, shorthorned grasshoppers into gregarious hordes of locusts capable of causing great destruction. This transformation has been studied in attempts to control these pests.

NATURAL HISTORY

Life cycle. *General features.* Since orthopterans undergo simple metamorphosis and have externally developing wings, they are known as hemimetabolous insects. The grylloblattids are wingless, and all large orthopteran groups contain a few wingless species, even though the basic structure of the orthopteran thorax proves their relationship to winged insects. A typical orthopteran life cycle has three stages: egg, nymph, and adult. Usually eggs are deposited outside the body on the ground or on vegetation; however, in some cases (*e.g.*, viviparous cockroaches), eggs are incubated in a brood chamber within the body of the female, and nymphs are born alive. Nymphs resemble adults except for their smaller size and lack of development of reproductive organs and wings; there is no pupa, or resting stage. In most orthopteran groups, the hatching insect that wriggles from the egg is not a fully formed nymph with freely moving legs; actually it is little more than an active embryo and is still enclosed in a thin membrane. This stage is called a vermiform larva; shed-

Simple metamorphosis

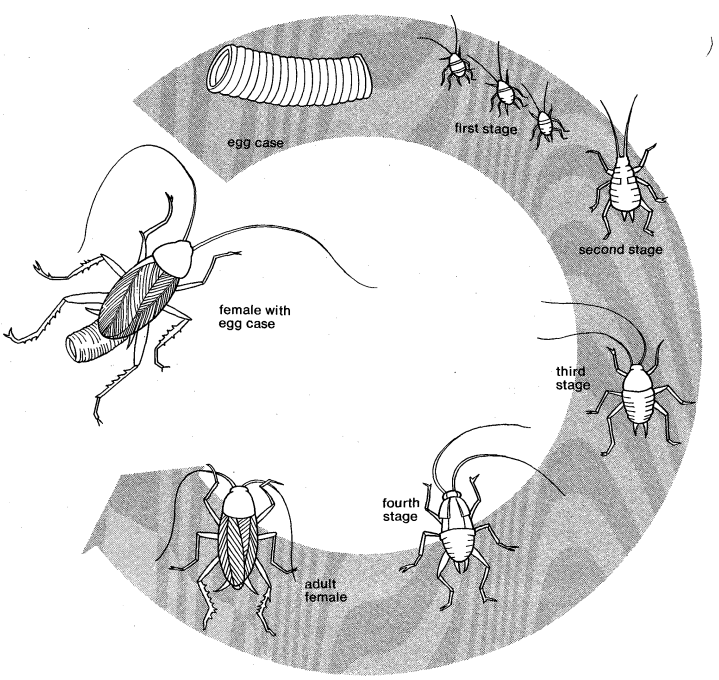


Figure 2: Orthopteran life cycle.
Courtesy of the U.S. Department of Agriculture

ding of the enclosing membrane occurs at the intermediate molt. The shapeless skins shed by young grasshoppers crawling from egg pods or by mantids leaving an egg case are examples of such exuviae (cast skins).

The number of nymphal stages between the intermediate molt and adulthood varies from about 4 to 13. Generalizations are approximately as follows: cockroaches, 5–13; mantids, 4–9; grylloblattids, about 8; crickets and katydids, 5–9; walking sticks, 4–6; and grasshoppers, 4–9, most often 5–7.

Egg cases

Egg laying. Egg-laying habits are distinctive in many orthopteran species. Among cockroaches, only one family (Blaberidae) is viviparous; the other four families contain species that have well formed egg cases (oothecae). Among these families, some species carry the oothecae protruding from the body until hatching time is near; others, however, deposit their egg cases within several days of formation. Usually oothecae contain from a few to more than 30 eggs arranged in two rows. Along one edge of the ootheca is a seam that bears a keel, or ridge; the shape of the ridge varies in species that carry the ootheca externally. Minute openings from the base of the keel to the interior of the ootheca are known to be a ventilating device in some species. The ootheca is first carried in the body with the keel uppermost; in certain groups, however, it is rotated prior to deposition so that the keel is on one side. Details concerning positioning of the ootheca and other aspects of egg laying were not correlated with behaviour patterns or group affinities until a basis for understanding their significance was established.

Mantids lay eggs in clusters of less than ten to more than 300. Usually they are laid in regular layers surrounded by a viscous quick-drying liquid that provides a light but tough protective covering. The egg masses of most mantids have a distinctive shape and size. Although most mantid species attach oothecae to vegetation, these egg cases may be attached to rocks in some environments or placed in grooves in the sand and covered over in the desert.

Grylloblattid eggs are laid in damp moss, decaying logs, or in pockets between broken rocks and wet soil. The eggs, about 3 mm long and usually black in colour, are laid loosely or inserted into the hatching site with the ovipositor.

Of the walking stick species studied, most have eggs that look like small seeds and are dropped loosely on the ground. At least one species, however, attaches its eggs to foliage, and a large, heavily bodied species of the southeastern U.S. (*Anisomorpha buprestoides*) scratches a depression in sandy soil with its front and middle legs, deposits eggs in it, and covers them with sand.

Crickets and katydids utilize their ovipositors to insert eggs into soil or plant material. The eggs of tree crickets (*Oecanthus*), for example, are inserted in rows in the canes of blackberries and various other stems; the eggs of field crickets (*Gryllus*) are laid in soil; the flattened eggs of certain katydids (*Scudderia*) are forced between the upper and lower epidermis layers at the margin of tree leaves; and the eggs of other katydids (*Microcentrum*) are laid in overlapping rows lengthwise on twigs of trees.

Most grasshoppers lay their eggs in soil; a few drill holes in dead wood or place their eggs at the bases of grass clumps or on the surfaces of leaves. Before laying the eggs, the female manipulates the valves of her ovipositor to make a hole in suitable soil (the type varies with the species). During the digging process, the abdomen is greatly extended, and the female manipulates the ovipositor valves to open and close and rotate on the long axis. Then she deposits several dozen eggs in the hole. The eggs are surrounded by a mucilaginous mass (the egg pod) that dries in a cylindrical shape. The number of egg pods laid by one female and the number of eggs in each pod vary according to the species and local conditions. The egg pods are laid over an interval of several weeks.

Parthenogenesis

A few orthopteran species have females only; therefore, reproduction occurs without fertilization (parthenogenesis). Only rarely are species that normally reproduce bisexually parthenogenetic; and when parthenogenesis does occur in bisexual species, it is usually only partially successful because the few nymphs that do hatch often are deformed and fail to reach maturity. In the laboratory, however, there have been a few cases in which several generations have been produced parthenogenetically, proving that there is an inherent capability in these bisexual groups for reproduction without males.

Growth and life span. Increases in the number of antennal segments occur during the development of some orthopterans. The German cockroach, *Blattella germanica*, has been studied in detail in this regard, and it has been shown that a newly hatched nymph has about 24 segments in each antenna. Each of the succeeding nymphal stages (there are usually six in this species) shows an increase in the number of segments until, by the time the adult stage is reached, the average number is 94. The two basal segments do not divide; the third segment, as well as certain other ones, is a growth centre that divides during molting. It is customary for grasshoppers to have 20–30 antennal segments when mature; this is about twice the number present in the first-stage nymph.

The life span of orthopterans depends in part upon whether or not there are long periods during winter or at other times (*e.g.*, a dry season) when the insects are quiescent. Some species habitually spend several months resting during unfavourable periods; such species have one generation per year, and the life span of an individual is approximately one year. The portion of the life span spent as an adult varies, but is likely to be about one or two months. In some species egg maturation is delayed several months after the final molt, and the females do not lay the eggs until they have matured. In other species eggs are not hatched until one or more years after they are laid; therefore, more than one winter or dry season is passed in the egg stage, and a single life cycle can occupy two or more full years. For instance, a walking stick commonly found in the U.S., *Diapheromera femorata*, often has some eggs that hatch the year following deposition and others that hatch after two winters have been passed amid dead leaves on the ground. There are some orthopterans that develop very slowly, and their life cycles require several years for completion; an example is a North American cockroach (*Cryptocercus punctulatus*) that inhabits rotting logs, feeds on decaying wood, and attains maturity after six or seven years. Grylloblattids live for five to seven years. The time required for domestic cockroaches to reach maturity varies with species and environmental conditions. The German cockroach completes nymphal growth in about 95 days, but the American cockroach (*Periplaneta americana*) needs about 225 days. Similarly, adults live from a month or two to several years, depending on conditions and species.

Reproduction. *General features.* Typically each female has paired ovaries consisting of tubes in which eggs develop, moving posteriorly into a single oviduct as they "ripen." The oviduct leads to a vagina and then to the exterior where there is either a simple or specialized ovipositor consisting of paired appendages called ovipositor valves. Attached to the oviduct or vagina is a sac (called the spermatheca) for storage of male sperm; as eggs move down the oviduct, they are fertilized by the sperm. The typical male contains paired testes that produce vast numbers of slender active sperm; these are stored in enlargements of the tubes leading posteriorly from the testes. Accessory gland secretions provide not only the medium for carrying the sperm but also a material that solidifies to form a thin-walled sac, or reservoir, containing some sperm and fluid. This reservoir, called the spermatophore, is almost universally found among orthopterans.

Spermatophores

Grasshopper spermatophores consist of a bladderlike reservoir and a spermatophore tube. The spermatophore is formed during the first two minutes or so of copulation, after which the tube extends from the male genital organs to the spermathecal duct of the female. The spermathecal duct opens at the base of the ovipositor valves. Sperm pass to the female during copulation; after sperm transfer is complete, parts of the spermatophore may remain attached to both male and female. In some orthopterans, particularly crickets and katydids, the entire spermatophore is attached to the female.

Courtship behaviour. All orthopteran groups have species that show definite courtship behaviour prior to actual mating. Male cockroaches are attracted particularly by females that are virgin and in a receptive condition. Such females frequently secrete pheromones. Pheromones, chemical substances secreted by certain insects, influence the behaviour of other individuals of the same species. Antennae of males of a domestic roach, *Periplaneta americana*, have specialized sense organs that detect the odour of female *P. americana* pheromones; upon detection of the odour, the male initiates searching movements, first with the antennae, then with the palpi. Finally, the male, with folded wings raised and fluttering, actively searches out the female. If the female is still receptive when he finds her, the male protrudes his posterior abdominal segments, pushes under the end of the female, and grasps the terminal ventral segments of the female with his genital hooks; then he expels the spermatophore, which becomes attached to the spermathecal opening of the female. The entire process lasts up to an hour.

Mating calls

Among the grasshoppers, species with coloured hind wings and the habit of making sounds during flight use hovering and other special flight patterns to attract the attention of females. Crickets and katydids have the most dramatic courtship displays because "songs" enter into the precopulatory behaviour. Females of some species are receptive only to the specific song of a male of the same species; in others, however, mating calls are not necessary, and a female will mate with a male who is unable to sing because his wings have been removed. Here, as in grasshoppers, a variety of mating positions are assumed.

A striking sequel to mating occurs frequently in mantids when the female eats the male. There is a popular opinion that mantid males always are eaten, but many escape under natural conditions. But in the close confines of a small cage cannibalism of the male is more common.

Ecology. In a broad sense, ecology represents the sum total of interrelations between organisms and their environment. In the case of orthopterans, the basic requirements of food and moisture; shelter, including protection from weather and from enemies; favourable habitats, involving special niches such as caves or deserts; as well as preferred seasons and conditions conducive to successful reproduction, are involved.

Food. Mantids are the only orthopteran group that feeds almost entirely on insects, but some members of primarily plant-feeding groups also capture and devour insects. For instance, tree crickets (*Oecanthus*) regularly eat flowers, leaves, and other plant parts, in addition to many aphids and other weak insects. Some katydids are active predators of insects. Most cockroaches are typical scavengers, but some are specialized feeders. *Cryptocercus*, for example, digests cellulose in decaying logs with the aid of symbiotic protozoans in its intestines. All walking sticks and a majority of grasshoppers are plant feeders. Although many grasshoppers feed on a wide variety of plant species, some are restricted to a single plant species or one group of plants. Some orthopterans consume only certain parts of plants; for example, coneheaded katydids of the genus *Neoconocephalus* feed mainly on seeds of grasses. Plant preferences among some leaf-eating orthopterans change with the seasons of the year; in other leaf-eating groups feeding habits are dependent on the stage of the life cycle—*i.e.*, nymphs do not eat the same plants as adults of the same species. Moisture requirements of orthopterans vary, as evidenced by the habitats they occupy. Some can absorb water from a drop on the cuticle (skin); others obtain it from water vapour in the air if the relative humidity is sufficiently high.

Predators, plant feeders, scavengers

Habitat. Shelter utilized by orthopterans ranges from general hiding places amid living plant foliage or dead leaves on the ground to special structures such as subterranean galleries in soil or the recesses in caves occupied by various crickets. Some Gryllacrididae are leaf rollers and produce silk to maintain the rolled shape of their hiding places. The loose bark of trees and logs and the water-filled leaf bases of bromeliads often shelter certain genera of cockroaches, some of which are semi-aquatic in their habits. In Africa a few cockroaches (*Cyrtotria*), of elongated and cylindrical body shape, are adapted to enter round holes in hollow plant stems where they sometimes live. With the exception of cockroaches, most conspicuous orthopterans are active by day (diurnal), although there are nocturnal species in every group. Although grylloblattids are essentially nocturnal, they are sometimes active on cloudy days or in winter. The majority of crickets and katydids are nocturnal, as are many walking sticks and some mantids; however, many mantids prey on insects that visit flowers by day.

The degree of moisture, types of vegetation, and altitude above sea level influence the location of orthopteran communities. Grasshoppers breed in the Himalayan Mountains at altitudes as high as 6,000 metres (about 18,000 feet), and each mountain altitudinal zone has distinctive species; fully winged, actively flying species are usually not restricted to a single zone but are found in adjacent ones. On the other hand, at high altitudes there are proportionately greater numbers of grasshoppers with short, nonfunctional wings or none at all.

Caves as habitats

Caves are a special habitat occupied by orthopterans on all continents. The long-horned grasshoppers and the crickets are the principal orthopteran representatives; nearly 200 species of these two groups have been found in caves. In addition, more than 30 cockroach species inhabit caves; and a third group, the grylloblattids, has at least one cavernicolous species in Japan and three in the U.S. Some of these orthopterans inhabit lava tubes and fissures resulting from past volcanic action. Air currents and high humidity in these tubes and fissures produce an "ice cave" condition. In the U.S., the grylloblattids and a few dozen cave crickets (Gryllacrididae) are the principal cavernicoles. It is noteworthy that a bone from a bison skeleton, found in a French cave in the 1920s, bore a prehistoric carving that depicted *Troglophilus*, a European cave cricket.

Usually the orthopteran species found on a given continent are distinct from those of other continents, especially if the land masses are well separated. For example, there are about 2,000 species and 500 genera of grasshoppers in Africa; although several of the genera are found in North America, none of the species is. Some species in North Africa, however, also inhabit southern Europe and western Asia. Explanations for distinct continental species are found not only in the far, overwater distances involved but also in the long periods of isolation that have occurred and in the different conditions that have prevailed in past geologic periods.

Oceanic islands have been populated, in part, by species that were transported by hurricanes, floating debris, birds, and in recent centuries by human activities. The 85 species of Hawaiian orthopterans include four distinct genera of crickets and katydids comprising about 45 of the 85 species. Evidently some of these evolved following the establishment of a few parent species. The remainder are believed to have been established as a result of the activities of man. Since cockroaches are scavengers, they are often found where man is found; *i.e.*, in his buildings and campsites. Early man probably spread cockroaches as he moved about seeking food. Modern commerce has been even more helpful to these unwelcome travellers. An analysis of the distribution of 11 domiciliary species found in the U.S. and related species found elsewhere suggested that five reached America in shipping from West Africa; another African species might not have reached America directly; two probably came from Europe; two might have come from the Orient; and one was native to the West Indies. Thus man has played an important role in spreading cockroaches throughout the world.

Taxonomic significance of ecology

The familiar, large, black field crickets of the U.S. are good examples of ecological differences among similar species. There are six native species in the eastern U.S.; several others occur in the western states, and at least two introduced species have become established in the Gulf States. Using the general appearance of dead specimens, the native eastern species are very similar and difficult to distinguish. For many years, the taxonomy of these species was unsettled. Entomologists using behaviour rather than morphology as a major taxonomic criterion

have found that five of the six species have distinct songs; that four of them overwinter as nymphs, the other two chiefly as eggs; and that, to a considerable extent, the habitat preferences are different—*i.e.*, one species lives in abandoned fields, another in leaf litter of open woodland. Laboratory attempts to crossbreed males and females of different species have been unsuccessful; the pair either failed to copulate or, if they did, produced unhealthy hybrids. Ecological differences also are important for other groups of closely related orthopterans.

FORM AND FUNCTION

Adaptive features. *Movement.* Orthopterans exhibit various adaptations for movement; some are present in an entire family or suborder, others are peculiar to certain genera. The head of mantids is borne by the prothorax in such a way that it is easily turned to face in different directions. Since the mantid diet consists almost entirely of insects, vision is critically important and is unusually well developed. The best known orthopterans with specialized front legs are mantids; the principal leg segments are hinged and spined for seizing and holding prey. Some Orthoptera, especially certain groups of Tettigonioidea, also have front legs with long spines that enable them to hold other insects, although the hinging is not comparable to that in mantids.

Burrowers

Although some cockroaches burrow in soil, sand, or decomposing wood, the principal burrowers are found among the Orthoptera. In both groups the legs, especially the front tibiae, are short and strong, with heavy spurs. Mole crickets, false mole crickets, and sand crickets are accomplished burrowers. Small tunnels serve as shelter and as egg-laying locations, and roots or tubers encountered while burrowing are sometimes used as food. Several genera of camel crickets (Gryllacrididae) in the southwestern U.S. have conspicuous, sometimes basket-shaped, clusters of spurs on the hind legs. They often live on sand dunes and burrow chiefly for shelter. A few desert-living grasshoppers, some in the southwestern U.S., others in Africa, exhibit what has been called "self-burial." Instead of making an elongated cylindrical burrow, the grasshopper rests on the surface of the sand or moves forward and backward, manoeuvring its legs until it has submerged itself and covered its body with sand. The apparent purpose is protection.

Hind legs of Orthoptera, though useful in walking, are used primarily for leaping. Particularly important are the large muscle in the femur, the hinged attachment of tibia to femur, and the tendon extending within the leg from the femur to the end of the tarsus. In a few semi-aquatic Orthoptera, the hind tibia is broadened as a paddle or equipped with fringed spurs to permit effective swimming strokes in water. The ability to run swiftly is common among cockroaches and some mantids. Cockroaches escape enemies by running; mantids utilize their running ability both to escape predators and to catch prey. Some mantids that live on the ground, in deserts, or on tree trunks in the tropics are active runners; however, the majority of mantids stalk their prey slowly or wait quietly until an unsuspecting insect moves nearby.

Body shape is important to many orthopterans, either allowing them to live in places where adequate shelter from weather and enemies is provided or affording them concealment through camouflage. Most cockroaches have flat bodies that enable them to hide beneath stones, under other objects on the ground, or under the loose bark of logs. Examples of orthopterans whose camouflages resemble parts of plants are members of the Phasmida; some of them resemble leaves, others look like twigs or rough pieces of small limbs from trees. Several katydids and grasshoppers resemble leaves; some are green or brown, others have spots that resemble leaves affected by plant diseases. There are some slender grasshoppers that live among grasses, where they conceal themselves by clinging lengthwise to stems and remaining motionless or by quickly sidling around behind stems.

Colour

Camouflage. There are two basic types of insect colours. Structural colours occur when irregular cuticle or scale surfaces break up and reflect certain wave lengths of light. Metallic lustres of some orthopterans (*e.g.*, silvery patches on some grasshoppers) are examples. Most orthopteran colours are due to pigments; often they are located in the cuticle, but sometimes they occur in some deeper body layer. The pigments may be naturally occurring ones or, like melanin, dependent on an oxidation process or a hormonal balance that influences metabolism; these latter pigments are present in varying amounts in different individuals of the same species.

Among some orthopterans, especially grasshoppers, body colours tend to simulate the colour of the habitat background. This is particularly true of species inhabiting rocky or sandy environments. In some cases, colour changes occur rapidly; this was demonstrated by certain light gray African grasshoppers that became black after being caged a few days on dark burnt-over ground. In other cases more time is required. Colour changes usually involve the effect of bright light on integumentary pigments. Among some orthopterans, however, light must enter the eyes, and a rhythm related to some nervous-endocrine mechanism is apparently involved.

An unusual and rapid colour change occurs in an Australian alpine grasshopper (*Kosciuscola tristis*), which lives at above 5,000 feet elevation. The adult male, bright greenish blue on the upper part of its body at temperatures above 25° C, is dull and blackish below 15° C. At intermediate temperatures, correspondingly intermediate shades of colour occur. Detailed experiments by Australian entomologists prove that temperature, not light intensity, relative humidity, or degree of crowding is the controlling factor. The epidermal cells of the integument contain brown and blue granules; at warmer temperatures on sunny days the blue granules, in a discrete layer uppermost in the epidermal cells, are near the surface of the integument. At night or on cloudy days, the brown granules migrate from the bottom of the epidermal cells and change places with the blue granules. Thirty minutes is sufficient time for a colour change to take place.

Glandular secretions

Defense. There are no known stinging orthopterans but many have chemical mechanisms in the form of glands that produce irritating fluids or repugnant odours. The disagreeable smell of some cockroaches, especially when disturbed, is well known. Examples are several species of *Eurycotis* in Florida and tropical America; both sexes have a large gland in the hind part of the abdomen between the sixth and seventh segments. An acidic, milky fluid consisting of several chemical constituents is emitted either as an oozing liquid or as a three-foot spray. Another cockroach (*Diploptera*) has a defense gland that ejects a mixture of quinones from the second abdominal spiracles. Ants, beetles, and other predators become confused and avoid these cockroaches when they release their secretions; however, certain mantid predators are not affected.

Man may handle most walking sticks safely, but a large, heavily bodied species in the southeastern U.S. (*Anisomorpha buprestoides*) sometimes forcibly ejects a milky fluid that is extremely irritating if introduced into the human eye. This species has a pair of circular pores on the thorax leading to reservoirs of the fluid; each reservoir has circular muscles that permit ejection of fluid without the general body contraction characteristic of some grasshoppers. When handled, most grasshoppers and some other orthopterans regurgitate from the mouth a brown fluid that superficially resembles molasses. Release of the fluid from the forward part of the alimentary canal is triggered by a response of the nervous system to pressure on certain parts of the body, especially the sides of the thorax or the femurs. Some grasshoppers have other defense mechanisms (*e.g.*, some exude fluid through spiracles or from special glands opening on the body or even leg joints). Sometimes hissing sounds and blowing of bubbles from spiracles accompany secretion.

Physiology and biochemistry. *Body composition.* Several grasshopper species have been analyzed chemically. They consist of (by dry weight) roughly 50% to 75% crude protein, 4% to 18% fats, 4% to 16% carbohydrates, and 3% to 19% ash.

The tough and usually hard outer body wall (exoskele-

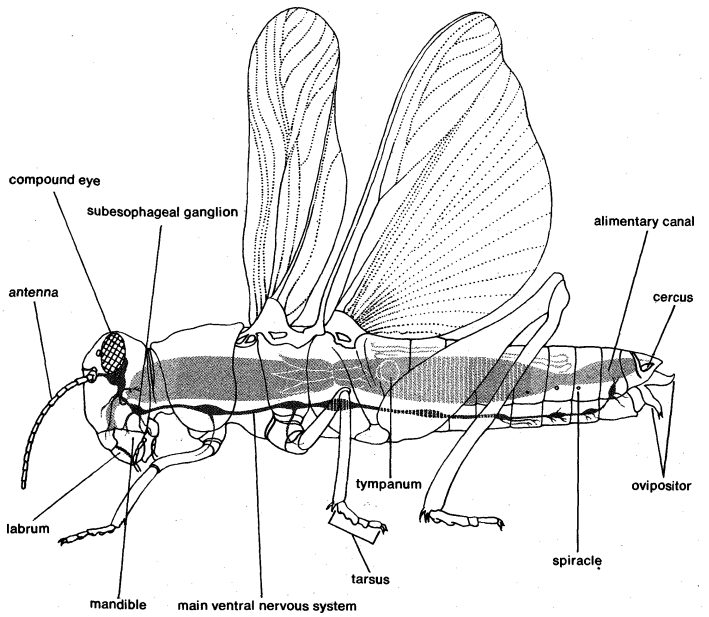


Figure 3: Internal and external body plan of adult orthopteran.
From H. Weber, *Grundriss der Insektenkunde* (1966); Gustav Fischer Verlag, Stuttgart

ton) of orthopterans is called the integument or cuticle; its most important component is chitin, a stable polysaccharide chemically similar to plant cellulose. Chitin makes the cuticle strong and flexible but does not provide rigidity. Sclerotin, the horny substance of the cuticle formed by a tanning-like process involving protein produced in the exoskeleton, is found in hard body plates (sclerites), leg spurs, and sharp tubercles; sclerotin is responsible for the rigidity of these structures. A heavily "sclerotized" cuticle is one that is hard and usually dark coloured.

Hormones. The importance of hormones in the biology of orthopterans has been revealed by research. Together with the related pheromones, which tend to coordinate individuals within the population of a species instead of regulating function within an individual, hormones are important in many activities of orthopterans related to mating and reproduction. Other activities involving hormones in grasshoppers include control of fat accumulation in metabolism, control of peristalsis in the malpighian tubules (excretory organs attached to the posterior part of the alimentary canal), secretion of an enzyme at hatching time for dissolving the cuticle that encloses the embryo, and control of the number of molts in nymphal growth.

Activities involving hormones

Detailed studies on the reproduction of cockroaches have disclosed an interrelated series of neurological and glandular functions that combine to control mating and egg production. Frequently, dorsal abdominal glands of the male aid in attracting the female to a mating position. In several cases, once a female has mated and an ootheca is being carried, mechanical pressure of the ootheca causes a stimulation to be transmitted to glandular bodies closely associated with the cerebral ganglia and called corpora allata; this in turn inhibits development of additional eggs in the ovarioles until laying and subsequent removal of pressure occur. In other cases, virgin females are receptive to mating just when yolk deposition is occurring in the first oocytes of developing eggs. Following mating, the mechanical stimulation of the inserted spermatophore inhibits further attraction of the female to the male abdominal glands until after the first group of eggs is deposited.

Phase polymorphism in locusts

Locust is a common name for several species of short-horned grasshoppers that often increase suddenly in numbers and undertake mass migration. Thus, a locust has both solitary and gregarious phases. Gregarious locusts outnumber solitary ones, migrate both as nymphs and adults, and travel in swarms. Swarming adults are tremendously destructive to crops. Typically, gregarious locusts have darker bodies and longer wings compared with solitary forms. Colour changes in adults are correlated with maturation of reproductive organs.

Hormones and pheromones are involved in many stages of locust development. Solitary locusts can transform into gregarious ones as a result of hormonal changes induced by crowding. The presence of mature male locusts under conditions of crowding stimulates a maturation hormone that causes females to mature rapidly. Head glands in the female are stimulated to release another hormone that speeds egg maturation. A favourable season followed by an unfavourable one may cause gregarious locusts to develop. In a favourable season with enough food, the population of solitary locusts increases. If the next season is a poor one, the solitary locusts are forced to crowd together where food is available. Crowding exposes the females to male secretions, females and their eggs respond by maturing rapidly, a population explosion occurs, and a locust horde results. In *Schistocerca gregaria*, the attainment of reproductive activity is sometimes synchronized with environmental contact with certain aromatic shrubs that produce terpenoids in season.

Sound production and hearing. Some orthopterans make conspicuous sounds, while others produce sounds that are outside the range of human hearing. In both cases sound production is important to behaviour necessary for success of the species concerned. Except for Grylloblattodea, in which sound production is unknown, all major groups of orthopterans produce some sort of sound, though sound production is widespread only in crickets, katydids, and grasshoppers.

Stridulation

The stridulatory mechanism of grasshoppers involves moving the hindleg across the folded front wing (tegmen). Serrations, or pegs, which vary in shape, number, and location among different species, are located on the inner surface of the femur and rub across special raised veins of the tegmen, creating a characteristic lisp; sometimes the serrations are on the tegminal veins. In the hindwings of other grasshoppers are stiff veins that make a crackling sound (crepitation) in flight.

Among male crickets and katydids, a front wing with an enlarged transverse vein near its base bears teeth that rasp when shuffled across a scraper on the other front wing. The row of teeth is called the file, and the membrane to which it is attached vibrates when the teeth move over the scraper. During stridulation the tegmina are lifted at an angle of 15° to 40° to the surface of the abdomen, then rapidly opened and closed (shuffled); sound is produced during the closure.

The best known auditory organs of orthopterans, the tympanic organs on each side of the abdomen, are found in both sexes of grasshoppers and on the front tibiae of most crickets and katydids. There are auditory nerves running from special cells beneath a tympanic membrane (a thin area of cuticle, backed by an air sac and free to vibrate) to a ganglion of the central nervous system. In addition to these evident tympanic structures, other less evident auditory organs occur in the orthopterans. Many orthopterans, however, have no conspicuous tympana and are entirely dependent for sound reception on sensory hairs located on cerci, the head, other parts of the body, and an auditory organ called Johnston's organ, which is widespread in the second segment of the antenna.

EVOLUTION AND PALEONTOLOGY

Cockroach fossils

Cockroaches are the most abundant and the earliest fossilized orthopterans found; fossils have been discovered at various localities in North America, France, and the U.S.S.R. Several hundred Carboniferous Period (about 280,000,000 to 345,000,000 years old) and at least a hundred Triassic cockroach species (about 190,000,000 to 225,000,000 years old) have been described; they differ from present Blattaria chiefly in details of wing venation. In addition, some early species had long ovipositors, unknown in recent species. Only a few fossil mantids are known, the oldest in Baltic amber of the Oligocene (about 26,000,000 to 38,000,000 years ago). Some phasmatid-like Jurassic species (about 136,000,000 to 190,000,000 years old) are believed to be primitive walking sticks.

Although fewer in number than the Blattaria, fossil Orthoptera have contributed to orthopteran classification. Ensifera occur from the Triassic to the present; apparently Ensifera and Caelifera separated as distinct evolutionary lines as early as the Carboniferous. The earliest fossils in the acrididoid line had long antennae. Shorter antennae, reduction of tarsal segments to three, and reduction in length of the ovipositor occurred by the Eocene Period (about 38,000,000 to 54,000,000 years ago).

By the late 19th century, all principal groups of orthopterans except Grylloblattodea were represented in collections; however, the order Orthoptera, broader in scope than it is at the present time, included earwigs and other groups. Gryllacridids were not placed in a separate family, and Phasmida were considered a family closely related to Mantidae because both are walking rather than jumping in habit. By the 20th century, however, basic morphological studies, as well as extensive reports on fossils, contributed new insights into fundamental relationships. The Grylloblattids were first reported in 1914, and numerous publications since then have analyzed their phylogenetic significance. In the late 1930s extensive studies of fossils were correlated with important work on current species, especially concerning the Orthoptera (restricted sense). Meanwhile, comparative studies of wing venation, the proventriculus, reproductive organs, and behaviour have steadily advanced the knowledge of group relationships. Additional details and supporting data, for example, were given in an extensive phylogenetic study in 1968. The rank of a suborder (Acridomorpha) for grasshoppers alone (Eumastacoidea through Acridoidea of this article), has not been evaluated sufficiently. Many of the earliest fossil orthopterans were different enough from any present ones to justify the recognition of separate though extinct families.

CLASSIFICATION

Distinguishing taxonomic features. Among the distinctive features of orthopterans are their wings, which, when present, usually number four. The two forewings, generally long and narrow, are many-veined and somewhat thickened. Among the Orthoptera, Dictyoptera, and Phasmida the forewings, hardened and of a leathery consistency, are known as tegmina. The hindwings, broad with many veins, usually are folded fanlike beneath the forewings when at rest. The females have ovipositors. Some are concealed by ventral abdominal segments; others are as long as the body. Orthopterans have mandibulate mouthparts adapted for chewing and undergo simple metamorphosis.

The classification of orthopterans into orders and families is based chiefly on comparative morphology, on indications of ancestry derived from fossils, and on relationships suggested by patterns of behaviour and the physiology of body systems. Similar anatomical structures that have been significant in deducing orthopteran relationships include: tarsal segments; hindlegs; wings; stridulatory and auditory organs; head capsule; thoracic sclerites; ovipositor; male genitalia; and proventriculus. In recent years correlations between behaviour and mating and reproduction, especially in cockroaches and crickets, have been used widely to support classification of families and subfamilies as well as distinctions between species.

In the grylloblattids, cockroaches, and mantids, the cerci are segmented; segments are lacking in katydids, crickets, grasshoppers, and walking sticks. The external male genitalia are sometimes concealed; in Grylloblattodea (grylloblattids), Dictyoptera (cockroaches and mantids), and Phasmida (walking sticks), the genitalia are asymmetric; in the Orthoptera (katydids, crickets, grasshoppers, and locusts) they are symmetrical. The Orthoptera have the femur of the hindleg enlarged for jumping; other groups have hindlegs similar in size to the middle legs. In mantids the front legs are modified for seizing prey, in mole crickets for digging. The tarsus consists of five segments in grylloblattids, mantids, and cockroaches; there are usually five segments in walking sticks, usually only three or four in Orthoptera. The antennal segments vary from fewer than seven to many long and setaceous ones.

Annotated classification.

ORTHOPTERAN

Common name for several orders of related insects; wings, when present, number 4; chewing mouthparts; mostly plant feeders; size range from 2 mm to 30 cm; more than 24,000 species; worldwide distribution.

Order Dictyoptera

Hindlegs similar to middle ones, adapted for running; tarsi 5-segmented; wing pads not reversed during nymphal stages (front wings remain above hindwings); antennae usually filiform, with more than 30 segments; cerci many-segmented; without auditory tympanum.

Suborder Blattaria (cockroaches)

Head usually concealed from above by shieldlike pronotum; two ocelli represented by pale areas (fenestrae); front legs adapted for running; proventriculus (gizzard) heavily armed on inner lining, with longitudinal folds between the teeth; members from Carboniferous to present; worldwide distribution; sizes 2 to 100 mm, average about 15 mm; more than 3,000 species.

Family Cryptocercidae. Wingless; blackish; cerci (sensory appendages) concealed by apical segments of abdomen; eyes small; size usually at least 15 mm; found in China, Manchuria, U.S.; 3 species.

Family Blattidae. Middle and hindfemurs with numerous strong spines similarly arranged on both ventral margins; ovipositor a plowlike (vavular) modification of apical ventral segment of abdomen; male subgenital plate symmetrical with widely spaced styli; size range about 10 to 35 mm; worldwide distribution; about 550 species.

Family Polyphagidae. Middle and hindfemurs mostly with few strong spines ventrally; spines on front and back ventral margins arranged differently; anal area of folded hindwing not plaited fanlike; postclypeus usually enlarged; range in size about 3 to 35 mm; worldwide distribution; about 200 species.

Family Blattellidae. Variable, mostly small species; mostly with numerous similar spines on ventral margins of middle and hindfemurs; ranging in size from 3 to 100 mm, with an average of about 12 mm; worldwide distribution; about 1,700 species.

Family Blaberidae. Mostly large species; spines of middle and hindfemurs variable; male subgenital plate asymmetrical; viviparous or ovoviparous; from 10 to 60 mm in size; worldwide in distribution; about 650 species.

Suborder Mantodea (mantids)

Head usually conspicuous anterior to a narrow pronotum, seldom concealed by a broad pronotum; 3 distinct ocelli; front legs adapted for seizing prey; proventriculus not heavily armed, inner lining with fine anastomosing ridges between teeth of moderate size.

Family Mantidae. Oligocene to present (fossils rare); size from 20 to 155 mm, average about 60 mm, worldwide distribution; about 2,000 species.

Order Grylloblattodea (Notoptera)

Legs similar, adapted for running; 5-segmented tarsi; wingless; eyes small or absent; no ocelli; antennae fairly long, filiform, about 25–40 segments; female with well-developed ovipositor, resembling Tettigoniidae (see below under *Orthoptera*); cerci long, with 8 to 9 segments.

Family Grylloblattidae (ice bugs). Recent (no fossils); size about 20–25 mm; found in northwestern North America, eastern Siberia, Japan; about 15 species.

Order Phasmida (Cheleutoptera or Phasmatoptera)

Legs similar, adapted for walking; tarsi nearly always 5-segmented; often wingless; when winged, tegmina often shorter than wings; wing pads not reversed as nymphs; ovipositor short, often concealed; male genitalia asymmetrical; cerci short, unsegmented; no tympanum or conspicuous stridulatory organs; usually very slender, elongated, sometimes broadened, even leaflike.

Family Phasmatidae (walking sticks, stick insects). Variable, rarely less than an inch long; a broadly conceived family; Triassic to present; 15 to 325 mm in size with average about 70 mm; worldwide distribution centred in warm countries; about 2,000 species.

Family Phylliidae (leaf insects). Depressed, leaflike, winged; tegmina short in males, covering most of abdomen in females; antennae long, pubescent in males, very short in females; all members recent; about 60 mm in size; Indo-Malayan distribution from Ceylon to Fiji; about 25 species.

Order Orthoptera

Hindlegs almost always enlarged and adapted for jumping; tarsi usually 3- or 4-segmented, occasionally fewer than 3 or as many as 5 segments; front wings more or less thickened; wing pads reversed (hindwing partly covering front wing)

during late nymphal stages; cerci nearly always unsegmented; specialized auditory and stridulatory organs often present.

Suborder Ensifera

Antennae usually long, with more than 30 segments; auditory organs, if present, consist of tympanum at base of front tibia; many species with stridulatory mechanism at base of tegmen; ovipositor usually present as rounded or flattened bladelike, or elongated cylindrical spearlike, structure.

Superfamily Prophalangopsoidea. Tarsi compressed or cylindrical, 3- or 4-segmented; tympanum present; simple stridulatory vein on each male tegmen.

Family Prophalangopsidae. Triassic to present. 25–47 mm in size; recent species found in India and northwestern North America; 3 species.

Superfamily Tettigonioidea. Tarsi usually depressed, 4-segmented, at least on middle and hindlegs; tympanum present; stridulatory organ specialized, with left tegmen uppermost and scraper (at edge of lower tegmen) often better developed on right tegmen.

Family Tettigoniidae (katydids, bush crickets). Jurassic to present; size range from 15 to 120 mm, average about 35 mm; distribution worldwide; about 3,000 species.

Family Phasmodidae (phasmodids). Very slender, resembling Phasmatidae (see above); head prognathous (*i.e.*, mouth directed forward instead of downward); hindfemurs not enlarged; all members recent; found in Australia; size range from 25 to 85 mm; 2 species.

Superfamily Gryllacridoidea. Tarsi depressed, usually 4-segmented; front wings without stridulatory mechanism; usually no tympanum; ovipositor usually flattened, bladelike; 12 to 50 mm in size with average 25 mm.

Family Stenopelmatidae (stenopelmatids, Jerusalem crickets). Heavy body, often wingless; front legs usually strongly armed for digging; ovipositor inconspicuous; members from Miocene to present; found in southeastern Asia, South Africa, North America, and Central America; about 35 species.

Family Gryllacrididae (gryllacridids). Usually winged, with folded wings surrounding abdomen at rest; found from Jurassic to present; chiefly tropical, most nocturnal, in trees and shrubs; one species in U.S.; about 550 species.

Family Schizodactylidae (schizodactylids). Winged or wingless; when winged, wing tips very long, rolled in a coil at rest; tarsal segments 2 and 3 with conspicuous lateral lobes; all members recent, found in southern Asia, South Africa; about 8 species.

Family Rhaphidophoridae (camel crickets and allies). Wingless; usually of humpback appearance; tarsi compressed; legs often very long and slender; found from Oligocene to present; worldwide distribution; about 300 species.

Family Lezinidae. Wingless; cerci with apical segmentation; found in Asia and Africa; about 10 species.

Family Henicidae. Usually wingless; front tibia with 1 or more dorsal spurs; front coxa usually with spine; represented in Old and New World tropics; about 150 species.

Superfamily Grylloidea. Tarsi 3-segmented; tegmina of males usually highly developed for stridulation; tympanum usually at base of front tibia; ovipositor most often round in cross section, spearlike, other times somewhat flattened or reduced; 2 to 50 mm in size with average 18 mm.

Family Gryllidae (crickets). Front legs usually not for digging; ovipositor usually elongated; species from Triassic to present; worldwide distribution; about 2,400 species.

Family Gryllotalpidae (mole crickets). Front legs highly modified for digging, with tibiae expanded and with fingerlike dactyls; ovipositor vestigial; represented from Oligocene to present; distribution worldwide; about 65 species.

Suborder Caelifera

Antennae short, with less than 30 segments; auditory organs, if present, at base of abdomen in form of tympanum; ovipositor usually consists of paired valvular appendages adapted for digging.

Superfamily Tridactyloidea. Tarsi 1- or 2-segmented; antennae short, with 12 or fewer segments.

Family Tridactylidae (pygmy sand crickets). Front tarsi 2-segmented; hindtibiae with unsegmented tarsus, no "claws," 2 long "paddles"; usually winged; 3 to 18 mm in size; worldwide distribution; about 60 species.

Family Cylindrachaetidae (false mole crickets). Body elongated, cylindrical, wingless; antennae 7- to 8-segmented; eyes reduced; ocelli absent; front tibia with 4 or 5 dactyls; cerci very short. All members recent; inhabit Australia and southern South America; size 40–60 mm; about 8 species.

Superfamily Tetrigoidea. Pronotum strongly elongated, extending backward over most or all of abdomen; tegmina short, scalelike; hindwings usually fully developed; no tympanum; front and middle tarsi 2-segmented; hindtarsus 3-segmented.

Family Tetrigidae (pygmy grasshoppers or grouse locusts). Miocene to present; 8–25 mm in size with average about 14 mm; worldwide distribution; about 700 species.

Superfamily Eumastacoidea. Abdominal spiracles situated in lateral membrane between terga and sternums; basal abdominal terga without auditory tympanum; basal segment of hindtarsus with serrated margins or at least with an external tubercle (all U.S. species wingless).

Family Eumastacidae (eumastacids or monkey grasshoppers). Sides of abdomen without stridulatory organ; antennae relatively short; frontal costa forked below median ocellus; represented from Miocene to present; worldwide in warm countries; about 200 species.

Family Tanaoceridae (tanaocerids). Specialized stridulatory organ on side of 3rd abdominal tergum, rubbed by hindfemur; antennae very long; frontal costa not forked; found in southwestern U.S.; 3 species.

Superfamily Pneumoroidea. Body robust; male fully winged, female short winged; hindlegs short; tympanum absent; male with stridulatory ridges on side of 3rd abdominal tergum.

Family Pneumoridae (pneumorids). All members recent; restricted to southern Africa; 11 mm to 107 mm in size; about 18 species.

Superfamily Proscopioidea. Body sticklike, superficially like Phasmatidae (see above under *order Phasmida*); head elongate, conical, usually oblique or vertical; wings rarely present, tympanum absent; no stridulatory organ.

Family Proscopiidae (proscopiids). All representatives recent; from Costa Rica to southern South America; size 30 to 180 mm with average about 60 mm; about 100 species.

Superfamily Acridoidea. Typically robust, with both pairs of wings and tympanum present, but very variable; pronotum not or briefly prolonged over abdomen; abdominal spiracles located in terga; tarsi 3-segmented, basal segment without serrated margins; size 8 to 120 mm with average about 28 mm.

Family Acrididae (grasshoppers). Many of unusual appearance; stridulatory mechanism often present, but not with organ on abdomen; vertex without a definite longitudinal furrow; tympanum normally present (absent in some, especially wingless genera); ocelli small; hindtibiae usually not expanded apically for swimming; Eocene to present; worldwide distribution; more than 5,000 species.

Family Pamphagidae (pamphagids). Usually large; wings variable; many wingless bulky species in arid regions; vertex of head with longitudinal furrow; lower basal lobe of hindfemur longer than upper lobe; members recent; in Africa and southern parts of Europe and Asia; about 300 species.

Family Pyrgomorphidae (pyrgomorphids). Head conical; longitudinal furrow of vertex present; lower basal lobe of femur longer than upper lobe; tympanum usually present; no stridulatory mechanism; recent; worldwide in tropical and subtropical areas, unknown in U.S.; about 350 species.

Each of the following 7 families contains only a few species: Pauliniidae, Xyronotidae, Ommexechidae, Trigonopterygidae, Charilaidae, Lathiceridae, Lentulidae. The Xyronotidae includes 1 Mexican species only; the other families are not represented in North America.

Critical appraisal. The arrangement of orders, suborders, superfamilies, and families presented above is a consensus of current opinion; however, some entomologists recognize different relationships and additional families too minor or too little understood to be included here. In most large groups, there are a few peculiar species that vary slightly from the characteristics described. Although an attempt was made in this classification to indicate the earliest geological period in which major groups were found, available information is not always clear as to whether the insect found belongs to a modern family or an ancestral one with more primitive features and a different name. The known species shown for each group represent living ones that have been named; it does not include fossils. In some cases modern catalog and monographs supply accurate counts; for others, estimates are given.

Several other orders of insects are orthopteroid in their general relationships. Among them, the Dermaptera (earwigs) differ from orthopterans by having short leathery front wings devoid of veins, hindwings with veins radiating from a central point midway of the anterior margin, and most cerci modified into pincer-like structures. Isop-

tera (termites) have similar front and hindwings; although in many ways they resemble some cockroaches, they differ in their elaborate caste system and their habits (*i.e.*, living in complex colonies consisting of reproductive individuals, sterile workers, and soldiers). Zorapterans show some morphological relationship to cockroaches but have two-segmented tarsi, peculiar wing venation, a primitive caste system, and other differences. Embiopterans (web spinners) are also orthopteroid in basic morphology, but are notably distinct from orthopterans by the much enlarged silk-producing basal segment of the front tarsus. Plecoptera (stoneflies) are also orthopteroid, but their front and hindwings are of a similar texture (unlike orthopterans), and their immature stages are specialized for an aquatic life.

BIBLIOGRAPHY. R.D. ALEXANDER, "Acoustical Communication in Arthropods," *A. Rev. Ent.*, 12:495–526 (1967), with many orthopteran examples; "Life Cycle Origins, Speciation, and Related Phenomena in Crickets," *Qt. Rev. Biol.*, 43: 1–41 (1968), a valuable summary of recent advances in cricket biology; and with D. OTTE, "The Evolution of Genitalia and Mating Behavior in Crickets (Gryllidae) and Other Orthoptera," *Misc. Publs. Mus. Zool. Univ. Mich.*, 133:5–62 (1967), well illustrated; E.D. BALL *et al.*, "The Grasshoppers and Other Orthoptera of Arizona," *Tech. Bull. Univ. Ariz.*, 93:257–373 (1942), well illustrated; W.S. BLATCHLEY, *Orthoptera of Northeastern America* (1920), an important guide, some parts out of date; I.J. CANTRALL, "The Ecology of the Orthoptera and Dermaptera of the George Reserve, Michigan," *Misc. Publs. Mus. Zool. Univ. Mich.*, 54:1–182 (1943), a survey of 78 species of orthopterans; R.F. CHAPMAN, *The Insects: Structure and Function* (1969), with many orthopteran examples; L. CHOPARD, "Dictyoptères," "Notoptères," "Chéleutoptères," and "Orthoptères," in *Traité de zoologie*, vol. 9 (1949), a general treatise on Orthoptera; V.M. DIRSH, "A Preliminary Revision of the Families and Subfamilies of Acridoidea," *Bull. Br. Mus. Nat. Hist. (Ent.)*, 10:351–419 (1961), the current classification of grasshoppers; "The Post-Embryonic Ontogeny of Acridimorpha (Orthoptera)," *EOS, Madr.*, 43:413–514 (1968), basic information on growth changes in grasshoppers; S.K. GANGWERE, "A Monograph on Food Selection in Orthoptera," *Trans. Am. Ent. Soc.*, 87:67–230 (1961), a summary of techniques for studying food preferences of orthoptera and a review of information on the food eaten by various orthopteran groups, chiefly U.S. species; A.B. GURNEY, "Praying Mantids of the United States: Native and Introduced," *A. Rep. Smithson. Instn. for 1950*, pp. 339–362 (1951), a semipopular review with photos of several species; D.M. GUTHRIE and A.R. TINDALL, *The Biology of the Cockroach* (1968), emphasis on the physiology of cockroach species commonly used as laboratory animals; P.T. HASKELL, *Insect Sounds* (1961), many references to orthoptera; M. HEBARD, "The Dermaptera and Orthoptera of Illinois," *Bull. Ill. St. Nat. Hist. Surv.*, 20:125–279 (1934), keys to species, illustrated; J.R. HELFER, *How to Know the Grasshoppers, Cockroaches and Their Allies* (1963), a semipopular handbook; A.D. IMMS, *A General Textbook of Entomology*, 9th ed. rev. by O.W. RICHARDS and R.G DAVIES (1957), an important but detailed text; F.A. MCKITTRICK, "Evolutionary Studies of Cockroaches," *Mem. Cornell Univ. Agric. Exp. Stn. 389* (1964), classification of cockroaches; K. PRINCIS, "Blattariae," in *Orthopterorum Catalogus*, 8 pt. (1962–71), a catalog of living cockroaches, of use to researchers in systematics and distribution; D.R. RAGGE, *Grasshoppers, Crickets and Cockroaches of the British Isles* (1965), a well-illustrated standard work; J.A.G. REHN, "Man's Uninvited Fellow Traveler, the Cockroach," *Sci. Mon.*, 61; 265–276 (1945), popular and factual; with H.J. GRANT, *A Monograph of the Orthoptera of North America (North of Mexico)*, vol. 1 (1961), detailed and authoritative; L.M. ROTH, "Oothecae of the Blattaria," *A. Ent. Soc. Am.*, 61:83–111 (1968), an illustrated review; "The Evolution of Male Tergal Glands in the Blattaria," *A. Ent. Soc. Am.*, 62:176–208 (1969), a review; "Evolution and Taxonomic Significance of Reproduction in Blattaria," *A. Rev. Ent.*, 15:75–96 (1970), a summary of reproductive habits; with T. EISNER, "Chemical Defenses of Arthropods," *A. Rev. Ent.*, 7:107–136 (1962), concerned with secretory glands that produce repugnant substances; and with E.R. WILLIS, *The Biotic Associations of Cockroaches* (1960), a documented summary; A.G. SHAROV, in *Trans. Paleont. Inst., Moscow*, 118:1-208 (1968; Eng. trans., "Phylogeny of the Orthopteroidea," pp. 1–251, 1971), evolutionary development; B.P. UVAROV, *Grasshoppers and Locusts*, vol. 1 (1966), the first of a 2-volume book on this subject; T.J. WALKER, "Cryptic Species Among Sound-Producing Ensiferan Orthoptera (Gryllidae and Tettigoniidae)," *Q. Rev. Biol.*, 39:345–355 (1964), emphasis on the large number of orthoptera species reproductively isolated by inconspicuous calling songs; F.E. ZEUNER, *Fossil Orthoptera Ensifera*, 2 vol. (1939), a review with excellent illustrations.

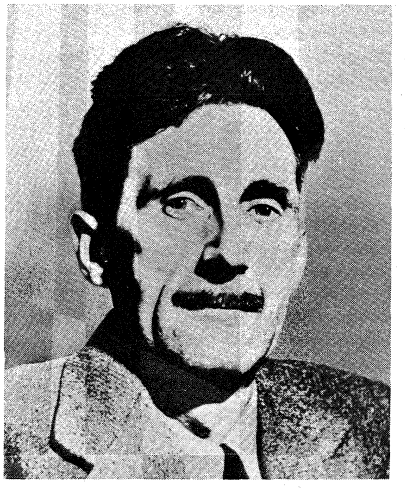
(A.B.G.)

Orwell, George

Journalist, satirist, autobiographer, and influential prophet of the evil shape of things to come, George Orwell, in his social and intellectual unorthodoxy, exemplifies an attractive recurrent strain in the British character.

Born Eric Arthur Blair, he never entirely abandoned his original name, but his first book (*Down and Out in Paris and London*, 1933) appeared as the work of George Orwell (the surname he derived from the beautiful River Orwell in East Anglia). In time his nom de plume became so closely attached to him in his daily life that eventually few people but relatives and his bank manager knew his real name was Blair.

By courtesy of the British Broadcasting Corporation



Orwell.

The change in name corresponded to a profound shift in Orwell's life-style, in which he changed from a pillar of the British imperial establishment into a literary and political rebel. He was born in 1903 at Motihāri, Bengal, into the class of sahibs. His father was a minor official in the Indian civil service; his mother, of French extraction, was the daughter of an unsuccessful teak merchant in Burma. Their attitudes were those of the "landless gentry," as Orwell later called people whose pretensions to social status had little relation to their income.

Early life

Orwell was thus brought up in an atmosphere of impoverished snobbery. After returning to England, he was sent in 1911 to a preparatory boarding school on the Sussex coast, where he was distinguished among the other boys by his poverty and his brilliance. The miseries of those years he told in his autobiographical essay, *Such, Such Were the Joys* (1953). He grew up a morose, withdrawn eccentric boy; but even then, as his schoolfellow Cyril Connolly remembered him, "Orwell . . . was a true rebel."

Orwell won scholarships to two of England's leading schools, Winchester and Eton, and chose the latter. He stayed from 1917 to 1921. Aldous Huxley was one of his masters, and it was at Eton that he published his first writing in college periodicals.

Instead of accepting a scholarship to a university, Orwell decided to follow family tradition and, in 1922, went to Burma as assistant district superintendent in the Indian Imperial Police. He served in a number of country stations, and at first appeared to be a model imperial servant. Yet from boyhood he had wanted to become a writer, and when he realized how much against their will the Burmese were ruled by the British, he felt increasingly ashamed of his role as an alien police officer. Later he was to recount his experiences and his reactions to imperial rule in his novel *Burmese Days* (1935) and in two brilliant autobiographical sketches, "Shooting an Elephant" and "A Hanging," classics of expository prose.

In 1927 Orwell, on leave to England, decided not to return to Burma, and on January 1, 1928, he took the decisive step of resigning from the imperial police. Already, in the autumn of 1927, he had started on a course of action that was to shape his character as a writer. Having felt guilty that the barriers of race and caste had prevented his mingling with the Burmese, he thought he could expiate some of his guilt by immersing himself in the life of the poor and outcast people of Europe. Donning ragged clothes he went into the East End of London to live in cheap lodging houses among labourers and beggars; he spent a period in the slums of Paris and worked as dishwasher in French hotels and restaurants; he tramped the roads of England with professional vagrants and joined the people of the London slums in their annual exodus to work in the Kentish hopfields.

Down and Out in Paris and London

These experiences gave Orwell the material for *Down and Out in Paris and London* (in which actual incidents are rearranged into something like fiction). Later experiences, teaching in private schools and working in bookshops, expanded his scope and contributed to his early novels, *A Clergyman's Daughter* (1935) and *Keep the Aspidistra Flying* (1936).

Orwell's revulsion against imperialism led not only to his personal rejection of the bourgeois life-style but to a political reorientation as well. Immediately after returning from Burma he called himself an anarchist and continued to do so for several years; during the '30s, however, he began to consider himself a Socialist, though he was too libertarian in his thinking ever to take the further step—so common in the period—of declaring himself a Communist.

Orwell's first Socialist book was an original and unorthodox political treatise, *The Road to Wigan Pier* (1937). It began by describing his experiences when he went to live among the unemployed miners of northern England, sharing and observing their lives; it ended in a series of sharp criticisms of existing Socialist movements. It combined mordant reporting with a tone of generous anger that from that time onward was to characterize Orwell's writing.

Role in the Spanish Civil War

By the time *The Road to Wigan Pier* was in print, Orwell was in Spain; he went to report on the Civil War and stayed to join the Republican militia, serving on the Aragon and Teruel fronts and rising to the rank of second lieutenant. He was seriously wounded at Teruel, damage to his throat permanently affecting his voice and endowing his talk with a strange, compelling quietness. Later, in May, 1937, after having fought in Barcelona against Communists who were trying to suppress their political opponents, he was forced to flee Spain in fear of his life. The experience gave him a lifelong dread of Communism, first expressed in the vivid account of his Spanish experiences, *Homage to Catalonia* (1938), which many consider one of his best books.

Returning to England, Orwell showed a paradoxically conservative strain in writing a novel (*Coming Up for Air*, 1939), full of nostalgia for the decency of a past England and fears for a future threatened by war and fascism. When war did come, Orwell was rejected for military service. For a time he headed the Indian service of the British Broadcasting Corporation, until in 1943 he became literary editor of *Tribune*, a left-wing Socialist paper associated with the British Labour leader Aneurin Bevan. At this period Orwell was a prolific journalist, writing many newspaper articles and reviews, together with serious criticism, like his classic essays on Dickens and on boys' weeklies and a number of books about England (notably *The Lion and the Unicorn*, 1941) that combined patriotic sentiment with the advocacy of a libertarian, decentralist socialism very much unlike that practiced by the British Labour Party.

In 1944 Orwell finished *Animal Farm*, a political fable based on the story of the Russian Revolution and its betrayal. At first he had difficulty finding a publisher for this small masterpiece, but when it appeared in 1945 *Animal Farm* made him famous and, for the first time, prosperous.

Animal Farm was probably Orwell's finest work, full of wit and fantasy and admirably written. It has, however, been overshadowed by the repute of his last book, *Nineteen Eighty-Four* (1949), a novel he wrote as a warning after years of brooding on its subject. In it he portrays the kind of society he believed could evolve if man allowed the state to assume more power and permitted politicians to establish and perpetuate totalitarian rule by a systematic distortion of the truth and a continuous rewriting of history.

Orwell wrote the last pages of *Nineteen Eighty-Four* in a remote house on the Hebridean island of Jura, which he had bought from the proceeds of *Animal Farm*. He worked between bouts of hospitalization for tuberculosis, of which he died in a London hospital in January 1950.

Orwell's ideal was to write "prose like a window pane," and he often succeeded. He recognized that he was not at his best as a novelist, and his best works of fiction are really fables. As a journalist and a writer of autobiography he was superb. He had little power of invention; it was life that moved him to his indignations and his finest writing.

MAJOR WORKS

NOVELS: *Burmese Days*, 1934; *A Clergyman's Daughter*, 1935; *Keep the Aspidistra Flying*, 1936; *Coming Up for Air*, 1939; *Animal Farm*, 1945; *Nineteen Eighty-Four*, 1949.

ESSAYS: *Inside the Whale*, 1940 (including also "Charles Dickens" and "Boys' Weeklies"); *Critical Essays*, 1946 (reprinting "Charles Dickens" and "Boys' Weeklies" and including also "The Art of Donald McGill," "Rudyard Kipling," and "In Defence of P.G. Wodehouse"); *Shooting an Elephant and Other Essays*, 1950 (including also "A Hanging," "How the Poor Die," "The Prevention of Literature," "Politics and the English Language," "Politics vs. Literature: An Examination of Gulliver's Travels," "Lear, Tolstoy and the Fool," and "I Write as I Please," articles from the *Tribune*); *England Your England and Other Essays*, posthumously published 1953 (including also "Inside the Whale" and the essays on Dickens and boys' weeklies, with "Looking Back on the Spanish War," "Anti-Semitism in Britain," "Poetry and the Microphone," "Why I Write," and part of *The Lion and the Unicorn* and *The Road to Wigan Pier*); *Such, Such Were the Joys*, posthumously published 1953 (except for the title essay, first published in this collection, and omission of extracts from *The Lion and the Unicorn* and *The Road to Wigan Pier*, containing the same essays as are collected in *England Your England*); *Collected Essays*, posthumously published 1961 (including most of the essays in *Shooting an Elephant*, *Critical Essays*, and *England Your England*).

OTHER PROSE (AUTOBIOGRAPHICAL): *Down and Out in Paris and London*, 1933; *The Road to Wigan Pier*, 1937 (account of conditions in the depressed areas of the north of England, observed on a tour commissioned by Victor Gollancz for the Left Book Club); *Homage to Catalonia*, 1938 (on his experiences in the Spanish Civil War). (POLITICAL): *The Lion and the Unicorn: Socialism and the English Genius*, 1941 (pamphlet); contributions to *Betrayal of the Left*, 1941; *Victory or Vested Interests?*, 1942; *Talking to India*, 1943 (broadcast talks). (MISCELLANEOUS): *The English People*, 1947 (text of volume in the "Britain in Pictures" series); introduction to vol. 1 of *British Pamphleteers*, 1948.

BIBLIOGRAPHY. The main cache of primary material is the Orwell Archive at University College, London; there are smaller collections at the University of Texas and the New York Public Library. All these archives are strong in letters and weak in other material since Orwell kept few papers. In accordance with a request in Orwell's will, no biography exists. The nearest thing, full of information about his life, with notes and a chronology, is SONIA ORWELL and IAN ANGUS (eds.), *Collected Essays, Journalism and Letters of George Orwell* (1968). Critical studies include: GEORGE WOODCOCK, *The Crystal Spirit* (1966); JOHN ATKINS, *George Orwell: A Literary Study* (1954); RICHARD REES, *George Orwell: Fugitive from the Camp of Victory* (1961); ROBERT A. LEE, *Orwell's Fiction* (1969); and KEITH ALLDRITT, *The Making of George Orwell: An Essay in Literary History* (1969).

(G.W.)

Ōsaka–Kōbe Metropolitan Area

The Ōsaka–Kōbe metropolitan area is the second largest urban and industrial agglomeration in Japan. Located on Ōsaka-wan (Ōsaka Bay) at the eastern end of the Inland Sea, it is usually defined to include the ancient city of Kyōto, once the national capital, which is located 25 miles northeast of Ōsaka.

Ōsaka, a city of 3,000,000, is the capital of Ōsaka Urban

Prefecture (Ōsaka-fu), an administrative division that includes Ōsaka City (Ōsaka-shi) and a large rural area. Kōbe, with a population of 1,300,000, is the capital of Hyōgo Prefecture (Hyōgo-ken) and one of Japan's chief ports. There are many satellite industrial and residential cities around the two central cities. The total population of Kyōto, Ōsaka, Kōbe, and their contiguous districts was 9,553,000 in 1965 and had increased to 14,886,000 at the 1970 census.

Ōsaka's ancient origins

History. The plain of Ōsaka was settled in prehistoric days and early became a political centre. The national capital was situated at nearby Nara in 710 AD. Among many ancient burial mounds in the vicinity of Ōsaka is that of the Emperor Ōjin (died AD 310?). The palace and town were built in 645, 651, and again in 724.

When Kyōto became the capital of Japan in 794, the road and water routes between Ōsaka and Kyōto were improved. The reclamation of the delta of the Yodo-gawa (Yodo River) allowed the building of new settlements, including Watanabe, which became a provincial capital and port during the Middle Ages. South of Ōsaka, on the eastern shore of the bay, is the port of Sakai founded in the 13th century. Like some of the medieval European towns, it was run by merchant guilds, and the accounts of Christian missionaries tell of the great houses of its wealthy men. In the 15th century, when Kyōto was destroyed in a war, Sakai became a centre of the arts.

In 1496, Buddhist priests built a temple in the uplands near Ōsaka, called Ishiyama Jinaimachi ("temple town with fortifications"). In 1580 it fell to the conqueror Oda Nobunaga after a siege of many years. His successor, Toyotomi Hideyoshi (1537–98), built a great castle on the site with massive stone walls and broad moats, and until 1596 it was the seat of the government of Japan. Destroyed by fire in 1868, the castle has been partly reconstructed, and it towers above the modern city. During the Tokugawa era (1603–1867) Ōsaka became the country's largest commercial city; feudal lords from all Japan established their warehouses along the canals, where they traded the rice of their villagers. Many other goods were traded in Ōsaka, and the city became an expanding industrial centre. In the 17th century the towns of Nada (now part of Kōbe), Nishinomiya, and Itami became famous for their sake (rice wine).

As its economic prosperity grew, Ōsaka became a centre of culture. It had schools of art, music, Kabuki, classical studies, and modern science. In the mid-19th century, when Japan was still closed to most Westerners, the Dutch language and Western science were studied by Japanese in Ōsaka.

Lawrence Smith—Photo Researchers

Ōsaka Castle.

Ōsaka remained pre-eminent both as a port and as a centre of industry until after World War II, when the Tokyo–Yokohama area underwent a great expansion. The revolution in China deprived Ōsaka of its important China trade, while the increasing economic role of the national government tended to encourage industrial location in the Tokyo–Yokohama area.

The contemporary city. *The city site.* The city of Ōsaka is situated on the delta of the Yodo-gawa. To the east of the central city, Hideyoshi's castle stands on the upland, which is the northern extension of the upland that rises in the southern part of Ōsaka Urban Prefecture and is about 65 feet above sea level. The metropolitan area spreads over the deltas of the Yodo, Yamato, and other rivers, and into their diluvial uplands. The area is bounded by the Ikoma-sanchi (Ikoma Mountains) in the east, the Izumi-sanchi in the south, and the Rokkō-sanchi in the northwest. The southwestern boundary of Ōsaka-wan is formed by Awaji-shima (Awaji Island). On the northwestern shore of the bay is Kōbe, above which rises the granite peak of the Rokkō-zan (Mt. Rokkō; 3,057 feet, or 932 metres). The coastline here has been altered by reclamation for port facilities and industries. Along the coast and in the uplands are the best residential areas of Kōbe and the cities of Ashiya, Nishinomiya, Ikeda, Itami, and Toyonaka. On the delta of the Kanzaki-gawa, just west of Ōsaka, is the city of Amagasaki, a centre of heavy industry. To the north of Ōsaka are the cities of Toyonaka, Suita, and Ibaraki. Above them, on Senriyama (Senri Mountain), are new towns developed after 1967. Northeast of Ōsaka, along the Yodo-gawa, are the industrial and residential cities of Takatsuki, Moriguchi, Neyagawa, and Hirakata. To the east of Ōsaka are the cities of Kadoma, Higashiōsaka, and Yao. To the southeast are Fujiidera, Tondabayashi, Matsubara, and others, most of them old historical towns. To the southwest, on the coastal plain, are Sakai, Izumi-Ōtsu, Kaizuka, Kishiwada, and Izumi-Sano, some of them industrial and others residential. Urbanization extends to Nara, 25 miles east of Ōsaka, and to Kyōto 25 miles northeast. A dense network of railways winds throughout the area.

River deltas and uplands

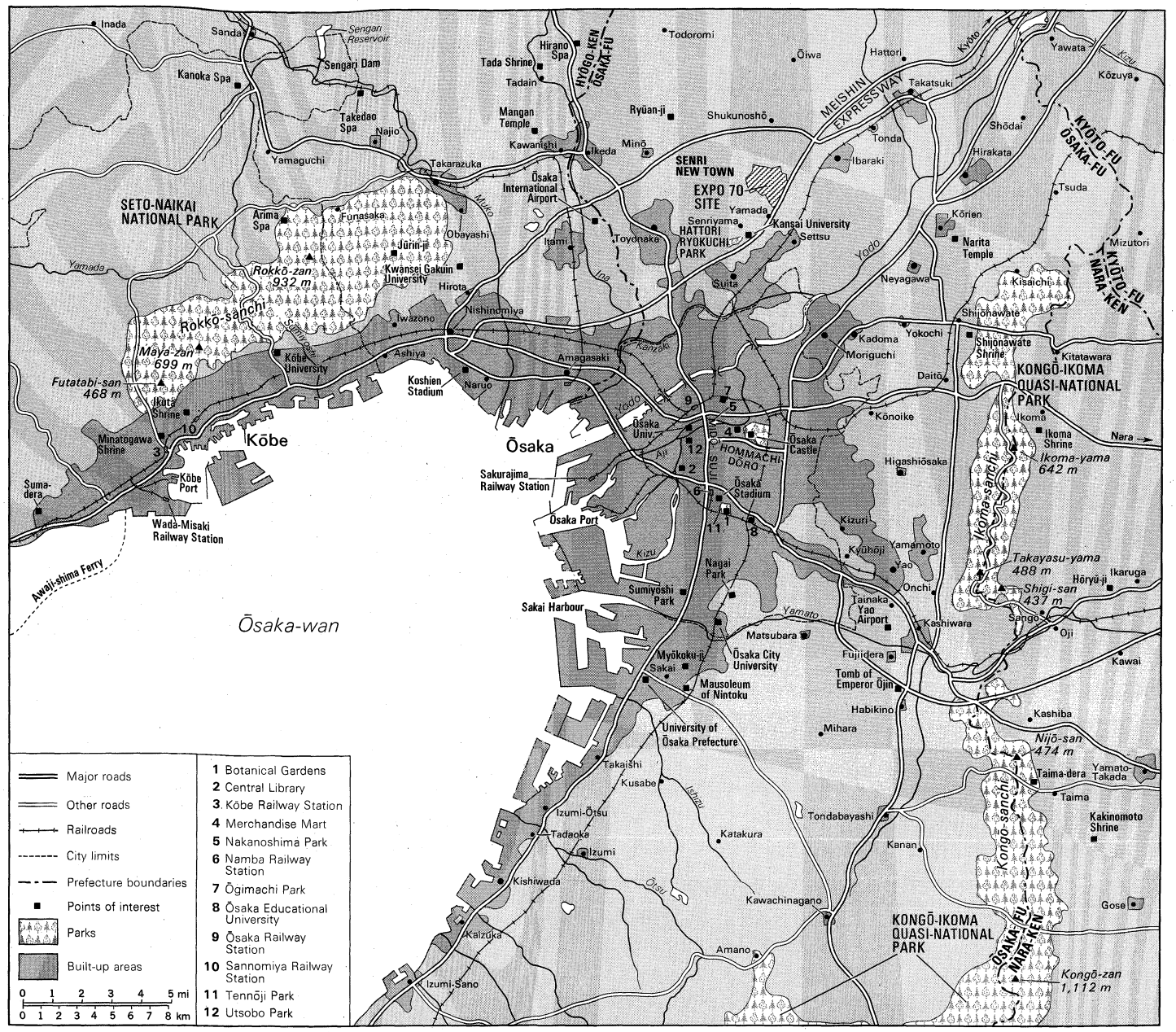
The physical environment. Ōsaka has a temperate climate. The annual mean temperature is 59.9° F (15.5° C), and the annual rainfall averages 53.5 inches (1,359 millimetres). The temperature in August is often 86° F or more (above 30° C), with no breeze from the sea at night. The January mean is 40.1° F (4.5° C), and snow falls several times in winter. The rainy seasons are in June–July and September–October. In September the region will usually be struck by one or two typhoons. The greatest typhoon disaster in the Ōsaka–Kōbe area's history occurred in 1934, when 3,000 persons were killed and 476,000 houses damaged. During the rainy season of June–July 1938, huge landslides from the Rokkō-sanchi buried wide areas of Kōbe.

The citizens of Ōsaka once took pride in its smoky atmosphere as a mark of industrial progress, but by the mid-1970s its smog and air pollution were seen as harmful. Other environmental problems were water pollution and subsidence (sinking) of the earth in the Amagasaki region caused by overuse of water beneath the ground.

Street patterns

City plan. Ōsaka's streets are laid out in checkerboard style. The north–south axis is Midō-suji (Mido Street), connecting Ōsaka railroad station in the north and Namba station in the south. The east–west axis is Hommachi-dōri (Hommachi Street), running east to the castle. Parallel to Midō-suji is the narrow Shinsaibashi-suji, the southern part of which is the central shopping district. Dotombori, at the south end of Shinsaibashi-suji, is a crowded theatre and cabaret area.

The central business district is the northern part of the downtown area. Nakanoshima, situated on an island formed by arms of the Yodo-gawa, contains the city offices, Ōsaka University, the Sumitomo Bank, and the headquarters of a number of large businesses. The traditional commercial centres were Semba and Shimanouchi



The Ōsaka–Kōbe metropolitan area.

streets, where the old-style white-walled shops with family quarters behind them continued for centuries, until World War II. Ōsaka's industrial areas are on the lower Yodo delta and in the eastern and northeastern parts of the city.

The street pattern of Kōbe is governed by its location between the mountains and the shore. Main streets run east and west, crossed by short north–south streets and occasional longer streets going up into the hills. The central shopping street, Motomachi, runs between the Sannomiya and Kōbe railroad stations. The central business district is near the harbour.

Transportation. Ōsaka is a junction point for the Japanese National Railways, feeding traffic from the north and east to the west and south. The JNR also runs local and interurban rapid transit lines, but the best service for suburban commuting and for regional passenger traffic is provided by the private electric lines, of which there are two between Ōsaka and Kōbe and two between Ōsaka and Kyōto. Nonstop electric trains run between Ōsaka and the city of Nagoya. Rail lines run from Kōbe to the Arima Spa on the north side of Rokkō-zan, to other northern mountain areas, and westward to Himeji, the third largest city of Hyōgo Prefecture.

There were 40 miles of subway lines in Ōsaka in 1970, and the system was being expanded. Because of heavy automobile traffic, the city's main streets were one-way. Elevated expressways ran through the central parts of Ōsaka and Kōbe and out to Ōsaka International Airport, near Itami. An express highway extends from Nagoya to Kōbe, and ordinary highways span the whole region. Kōbe and Ōsaka are both international and domestic ports; passenger ships, freighters, and car ferries sail to the islands of Shikoku and Kyūshū and to various ports of the Inland Sea.

Population. The population of Ōsaka declined from 3,156,000 in 1965 to 2,980,000 in 1970 because of migration to the suburbs. The density of the population in 1970 was 5,636 persons per square mile. The highest density is not in the centre of the city but in the peripheral wards, because the population decrease has been greatest in the central wards, while the outer wards were still growing. The population of Ōsaka Urban Prefecture in 1970 was 7,620,000 as compared with 6,637,000 in 1965. The increase was partly natural and partly the result of migration; on the average, the prefecture receives about 390,000 persons from other prefectures each year and sends out 310,000.

The city of Kōbe had a population of 1,289,000 in 1970 and of 1,217,000 in 1965; there was a slight decrease in the central ward and high increases in peripheral wards. Hyōgo Prefecture, of which Kōbe is the centre, had 4,668,000 in 1970 and 4,310,000 in 1965.

There were 27 cities of more than 100,000 population in the Ōsaka–Kōbe metropolitan area in 1970.

Burial mounds and housing developments

Housing and buildings. The Ōsaka area has been settled and built upon since prehistoric times. There are many ancient burial mounds and relics of Korean settlers, who introduced pottery and weaving to the Japanese. The ancient villages of the rice farmers were on the marshy plains, while the palaces, shrines, and temples were located on higher ground. Medieval settlements were in the uplands. Some of the present residential areas have been built on the sites of several former settlements. Thus, Tezukayama, a middle class residential development in Ōsaka south of the old castle, formerly an upper class area, is built over a number of ancient mounds.

The central part of Ōsaka is now primarily commercial; since 1920 there has been a migration from the city to the suburbs, helped along by the suburban railways that have made building land available along their rights-of-way. The Hankyu Electric Railway was particularly active in developing the city of Toyonaka (population 369,000) northwest of Ōsaka. Two of the large postwar housing developments are Koori Danchi, accommodating 22,000, and Senri New Town built in 1967–70, with 150,000.

Ōsaka Urban Prefecture had 1,830,000 dwelling units in 1968, and Hyōgo Prefecture had 1,124,000. About two-thirds of Ōsaka Urban Prefecture's dwellings were apartment houses. Inner-city areas were occupied mostly by Western-style multistoried buildings. Nineteenth-century architectural styles can be seen in the Public Hall of Ōsaka and in many residential houses in Kōbe. The redevelopment of Ōsaka's central business district in the 1960s produced some large new buildings: the Merchandise Mart, rising 256 feet above ground level and descending 66 feet below, and the Senba building under an elevated expressway—only four floors in height but about six-tenths of a mile in length.

The economy. Ōsaka was once known as the Lancashire of the Orient because of its great textile industry. At the beginning of the present decade its leading industry was the manufacture of machinery. The relative importance of its various industries was as follows (in percentages of the total production of Ōsaka Urban Prefecture in 1968): electric machinery, 12.6%; ordinary machinery, 12%; iron and steel, 11.2%; fabricated metals, 8.3%; textiles, 8.3%; chemicals, 8.2%; food, 6.3%; nonferrous metals, 5.7%; transportation machinery, 5.7%; printing and publishing, 3.9%; pulp and paper, 3.3%; other, 14.4%. Nearly 2,000,000 workers were employed in manufacturing in Ōsaka Urban Prefecture in 1968, 535,000 in the city itself.

Major industries

Between Ōsaka and Kōbe are several other industrial cities. The largest, Amagasaki, employed 100,000 workers; it is a centre of machinery, metallurgy, chemicals, cement, and paper production. Kōbe is pre-eminent in shipbuilding and steel production. Heavy industry and chemical plants are situated along the shore of Ōsaka-wan, while light industry and assembly plants are inland.

The merchants of Ōsaka greet one another in the mornings with the query, "Are you making money?" The city had 27,000 wholesale establishments in 1969, or about one-fifth of the country's total. The wholesale area is in the central part of Ōsaka, where the streets are named after commodities: Medicine Street, Textile Street, and so on.

Ōsaka is Japan's second largest financial centre. Together with Kōbe it is the leading port for foreign trade, handling about 30 percent of all exports.

Administration and social conditions. Ōsaka is the capital of Ōsaka Urban Prefecture, consisting of 31 cities and 15 towns and villages. It is also the centre of the Kinki-chiho (Kinki District), which consists of the seven prefectures of Ōsaka, Kyōto, Hyōgo, Nara, Wakayama, Shiga, and Mie. Various prefectural and regional institutions have their main offices in Ōsaka. Kōbe is the capital of Hyōgo Prefecture.

Public utilities. Ōsaka's main source of water is the Yodo-gawa. Kōbe has several reservoirs on the slope of the Rokkō-sanchi. Sewage services in the metropolitan area are generally inadequate, but most central city sections have flush toilets. Electricity is available everywhere, and gas is available in most city areas.

Health and education. Medical care in Ōsaka hinges on the hospitals of Ōsaka University and of the Ōsaka City University. Other hospitals and health centres are distributed throughout the metropolitan area. Ōsaka Urban Prefecture had 113 persons per hospital bed in 1969. The number of physicians per 100,000 persons was 132.6, which was higher than the national average. In general, the medical services compared favourably with those of other industrially developed countries.

The educational system

Ōsaka has the Japanese six–three–three–four educational system—six years of elementary schooling, three of junior high, three of high school, and four years of college. In 1969 the prefecture had 696 primary schools, 321 junior high schools, and 191 high schools, most of them run by public authorities. In Ōsaka and Hyōgo prefectures there were 55 universities with 177,000 students and 63 junior colleges with 38,000 students. There are four national universities: Ōsaka University, Ōsaka University of Foreign Studies, Ōsaka Educational University, and Kōbe University. Some of the public universities are Ōsaka City University and the University of Ōsaka Prefecture. Kansai University and Kwansei Gakuin University were the oldest and largest private universities located in the area.

Cultural life. Ōsaka and Kyōto have long been leading centres of culture—Ōsaka, famous for its restaurants, has a more bustling, democratic tone than Kyōto, which is one of the great centres of Japanese culture, and the difference is expressed in the popular sayings: "The Kyōto people spend too much money on clothes," and "The Ōsaka people eat too much rich food" (*Kyō no kidaore, Ōsaka no kuidaore*).

Traditional and modern Japanese drama and music are performed at theatres and halls in the metropolitan area, as are Western music, operas, and plays. There are numerous science and art museums, art galleries, and libraries.

Ōsaka is the place of publication of the national newspapers *Ōsaka Asahi* and *Ōsaka Mainichi*. There are also several local newspapers and many specialized publications. The national radio and television system has stations in Ōsaka, as do four commercial radio stations and four commercial television stations.

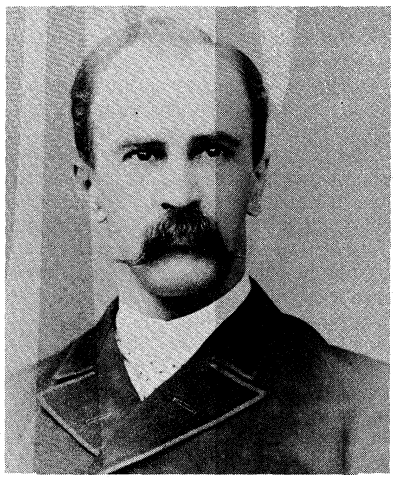
Green space in the city of Ōsaka is scarce. Two important parks are Nakanoshima and Tenno-ji, the latter with a zoo and botanical gardens. The suburbs have many historical sites and large recreation areas. Besides the large man-made Hattori Ryokuchi Park, there are the Kii-hantō (Kii Peninsula) on the Pacific, the beaches of the Inland Sea, the historical towns of Nara and Kyōto, and beautiful Biwa-ko (Lake Biwa)—a little larger than Switzerland's Lake Geneva—near Kyōto. At Kōbe, Rokkō-zan can be ascended by motor road or by cable car; there is a golf course at the top and ponds for swimming. There are four professional baseball teams in the metropolitan area, and the national high school baseball championships are played in the summer at Koshien Stadium. The town of Takarazuka in the northwest has been developed as an amusement centre; it houses the renowned Girls Opera and Dancing Theatre. In 1970 the Japan World Exposition (Expo 70) was held near Senriyama; the site is now used as a university campus and recreation area. (S.K.)

Osler, Sir William

When William Osler died in 1919, he was probably the most famous and beloved physician in the English-speaking and perhaps the whole world. He remains so more than 50 years later. His renown was not due to his contributions to science, which were small, or his contributions

to medicine, though these were sizable. He adored and fascinated the young, and he transformed medical education in the United States and elsewhere. Before Osler, the relations between teacher and student, teacher and teacher, and teacher and patient had been cold and formal. After him, the relations were warm and friendly. He personified the learned, scholarly, skillful physician who was also a warm human being. Osler thought his epitaph should be that he took the teaching of medicine into the wards.

By courtesy of the National Library of Medicine, Bethesda, Md.



Osler, 1886.

Early life

William Osler was born at Bond Head in western Ontario on July 12, 1849. He was the youngest of the nine children of the Rev. Featherstone Osler, who had gone to Canada as an Anglican missionary, and his wife, Ellen. William, like his father, was intended for the church. But while at school he read *Religio Medici* and became fascinated by natural history. He began to study arts at Trinity College, Toronto, but decided that the church was not for him and entered the Toronto Medical School in 1868. He subsequently transferred to McGill, where he took his medical degree in 1872. During the following two years he visited medical centres in Europe, spending the longest period at University College, London, in the physiology laboratory of John Burdon-Sanderson, who was making experimental physiology pre-eminent in medical education.

In 1873 Osler demonstrated that hitherto unidentified bodies in the blood were in fact the third kind of blood corpuscles, which were later named the blood platelets. These corpuscles had been observed before, but no one before Osler had studied them so thoroughly. Thus began what he called his periods of "brain dusting"—travel and studies that made him almost as much a part of Europe as of America.

Osler returned to Canada and began general practice in Dundas but was soon appointed lecturer in the institutes of medicine at McGill. He became professor there in 1875. A year later he became pathologist to the Montreal General Hospital and in 1878 physician to that hospital. At McGill he taught physiology, pathology, and medicine. His research was conducted largely in the postmortem room. In 1884, while in Leipzig, he was invited to occupy the chair of clinical medicine in Philadelphia. He decided to do so on the toss of a coin. While in Philadelphia he became a founder member of the Association of American Physicians.

Work at Johns Hopkins

In 1888 Osler accepted an invitation to be the first professor of medicine in the new Johns Hopkins University Medical School in Baltimore. There he joined William H. Welch, chief of pathology, Howard A. Kelley, chief of gynecology and obstetrics, and William S. Halsted, chief of surgery. Together, the four transformed the organization and curriculum of clinical teaching and made Johns Hopkins the most famous medical school in the world. Students studied their patients in the wards and presented the results to the "Chief." They were also encouraged to take their problems to the laboratory. Finally, the experts pooled their knowledge for the benefit of the patient and the student in public teaching sessions. Thus was born the pattern of clinical teaching that spread throughout the United States. Osler was not only professor of medicine but physician in chief to the hospital, an office first devised by the president of the university on the basis of his experience of running a large department store, and later to spread to most of the medical centres of the United States. For the first four years there were no students at Hopkins, and Osler used the time to write *The Principles and Practice of Medicine*, first published in 1892. In the same year he married Mrs. Grace Gross, widow of a surgical colleague at Philadelphia and great-granddaughter of Paul Revere.

Osler's textbook was lucid, comprehensive, interesting, and scholarly. It quickly became the most popular textbook of its day and has continued to be published since under a succession of editors, though never regaining the quality with which Osler endowed it. The textbook had an unexpected sequel. In the summer of 1897 it was read word by word by F.T. Gates, who had been engaged by John D. Rockefeller to advise him in his philanthropic endeavours. As a result of his reading, Gates inspired Rockefeller to direct his foundation toward medical research and to establish the Rockefeller Institute of Medical Research in New York.

Years at Oxford

In 1904, while visiting in England, Osler was invited to succeed Sir John Burdon-Sanderson in the Regius Chair of Medicine in Oxford. Osler's practice and teaching had for many years imposed enormous demands on his time and energy. His forceful wife telegraphed him from America: "Do not procrastinate. Accept at once." Osler did. The Regius Chair at Oxford is a crown appointment for which only citizens of the crown are eligible, but Osler had kept his Canadian nationality. He took up his chair in the autumn of 1905, becoming a student, roughly a lifetime fellow, of Christ Church, one of the Oxford colleges, a member of its Hebdomadal (weekly) Council, and curator of the Bodleian Library. In Oxford he taught only once a week, did a small amount of practice, and spent most of his time on his books. His library became one of the best of its kind, and after his death it passed intact to McGill, where it is specially housed. His scholarship was recognized by his election as president of the Classical Association. He was also active in medical affairs and inspired the formation of the Association of Physicians of Great Britain and Ireland and the establishment of the *Quarterly Journal of Medicine*. He was elected a fellow of the Royal College of Physicians of London in 1884 and a fellow of the Royal Society of London in 1898; he became a baronet in 1911. He and his wife were immensely hospitable, especially to visiting Americans, among whom their house was known as the "Open Arms."

Osler gave many lectures, mostly on the subjects of medical history and literature, some of which were collected and published. *Aequanimitas*, which he regarded as the most desirable quality for doctors, was the title of the most famous of these. Osler had a puckish wit and wrote some admirable medical nonsense under the pseudonym of Egerton Yorrick Davis, a retired surgeon captain of the U.S. army.

In medical terminology, Osler is immortalized in Osler's nodes (red, tender swellings in the palm or palmar surface of the fingers and hands in subacute bacterial endocarditis), Osler-Vaquez disease (polycythaemia vera or rubra), and Rendu-Osler-Weber disease (a familial form of recurring nose bleeds with multiple teliangectases of the skin and mucous membranes).

The Oslers had one son, Revere, named after his great-great-grandfather, Paul Revere. His death in action during World War I took the spirit out of his father, who died of pneumonia on December 29, 1919. His funeral was in the cathedral and chapel of Christ Church at Oxford on January 1, 1920. Lady Osler died in 1928. Her ashes, and those of Sir William, are now in the Bibliotheca Osleriana at McGill.

BIBLIOGRAPHY. H.W. CUSHING, *The Life of Sir William Osler*, 2 vol. (1925), is the standard life. See also the special commemorative issues devoted to Osler: *Arch. Intern. Med.*, vol. 84, no. 1 (1949); and *J.A.M.A.*, vol. 210, no. 12 (1969).

(G.Pi.)

Oslo

Occupying a site of rare natural beauty, the city of Oslo has grown into the capital of the kingdom of Norway and the centre of national economic, political, and cultural life. It is also the most important Norwegian port and the leading road, rail, and air junction; car ferries connect with Germany and Denmark, and there are regular passenger connections with Britain and the United States. Oslo is situated in southeastern Norway, at latitude 60° N, longitude 10° E, at the northern end of the 60-mile-long Oslo Fjord. The city covers over 175 square miles (453 square kilometres) and is, in area, one of the largest cities in the world. The built-up area, however, comprises only 60 square miles (155 square kilometres). Outside this historic waterside core, the greater part of the city area has been kept in a natural state. This region, which includes the areas known as Østmarka, Lillomarka, Nordmarka, Krokskogen, and Vestmarka, consists of extensive pine- or spruce-clad hills, interspersed with lakes (totalling 9 square miles), and marshes. It has been preserved primarily for recreational purposes, and a network of marked paths and trails threads the whole region, which is a popular destination for hiking and skiing trips on the part of Oslo citizens and their families.

The people of Oslo are in fact often prouder and fonder of the outlying regions than of central Oslo, even though the city centre has much to commend it: attractive parks; handsome public buildings; and the renowned thoroughfare of Karl Johansgate, which links the parliament building, the old university, the National Theatre, and the royal palace, which is surrounded by a park.

History. The *Snorres kongesagaer* ("Snorr's King's Sagas"), a historic Norse chronicle, records that

> King Harald established a town in the east, in Oslo, and spent a great deal of his time there, as supplies were easily obtained and the countryside thriving. Here he was in a good position to protect the country against the Danes, and also to make raids on Denmark.

The reference is to King Harald Hardraade, who lived around the year 1050, and illustrates the strategic advantages of Oslo's original site.

This lay close to the Oslo Fjord, to the east of the Aker River at the foot of the Ekeberg Heights. The old city district, as it is now known, contains only a few archaeological traces of this early settlement, which once contained a bishop's residence and several church and monastery buildings, some of them built on Hovedøya, an island in the fjord.

Original foundation

To the west of the old city, a rocky cliff juts out into the fjord, and it was on this commanding site that the Akershus Fortress was established (*c.* 1300). Rebuilt and extended, it dominates the centre of Oslo Harbour to this day, is still used for state occasions, and has lent its name to the county (*fylke*) encircling the city. It was under the walls of this castle that, following a disastrous fire in 1624, one of several which ravaged the young city, King Christian IV of Denmark–Norway built a new town laid out with a regular rectangular street system. This replaced the Ekeberg site. A famous statue, standing in the marketplace in front of the Cathedral, together with the names—Christiania 1624–1877, Kristiania 1877–1925—by which Oslo was formerly known, commemorate the modern city's founder.

During the 19th century, the growing city gradually grew beyond its old limits; its population grew from a mere 11,000 in 1814 to 225,000 in 1910; and it replaced its rival, the west coast port of Bergen, as Norway's largest and most influential city. A university was founded in 1811, while at midcentury, to the west of the then built-up area, an impressive royal palace was constructed. This was erected at the behest of Karl Johan (Charles XIV) the joint king of Norway and Sweden, who was to give his name to the city's main thoroughfare, Karl Johansgate. The Norwegian monarch still maintains his residence at the palace.

The 19-century expansion of the city initially took place along roads leading out from the old city centre, notably Grønlandsleiret and Strømsvein. The latter connected the city and its harbour with the hinterland, primarily the sawmills at Strømmen. Traffic in lumber was heavy, and the road was crowded with horse-drawn vehicles. In 1854, the nation's first railway line, connecting Oslo and Eidsvoll, was opened: this took over much of the lumber and freight traffic and, by stimulating the urban economy, made an important contribution to the growth of the city. Within the city itself, the turbulent Aker River became a focus for sawmills and other industrial development; and outlying settlements, consisting of the wooden dwellings of the new working class, soon sprang up, while the western sections of the city developed more as residential areas for the prosperous elements of society. By the 1970s much of the physical legacy of this period of growth was in a state of deterioration and extensive programs of urban renewal were under way. The Vika area around the city hall, has been the scene of an ambitious program of new construction since World War II.

Nineteenth-century growth

In the 20th century the most rapid period of growth for the city also occurred after World War II. In 1948 Oslo incorporated the nearby township of Aker, and in the following decades a number of satellite towns and residential areas—notably Grorud, Veitvet, Bøler, and Lambertseter—grew up to the east of the city.

The contemporary city. The geological conditions underlying Oslo are varied and interesting. At the end of the Ice Age, about 10,000–12,000 years ago, the sea level in the Oslo Fjord was nearly 700 feet higher than it is today. As a result the ancient rocks underlying much of the Oslo region have been covered by a thick layer of marine clay. Consequently, variations in the subsoil of the city have created problems during the construction of modern buildings, roads, and railways. Parts of the centre of Oslo had to be built on large rafts laid on top of the clay deposits, while subway construction has been a complicated and costly affair.

The hills surrounding the city offer magnificent views, with the highest points—Grefsenåsen, Vettakollen, Holmenkollen, and Tryvasshøgda—lying to the north of the city. A television tower 393 feet high rises from the top of Tryvasshøgda (1,760 feet [536 metres] above sea level). The Maridalen and Sørkedalen valleys, also on the outskirts of the city, have been maintained as agricultural areas: no residential building is allowed, except on the part of the local farmers. The nearby Maridalsvatn Lake is Oslo's largest lake and forms the most important reservoir for the city's supply of drinking water.

The perimeter of the city is some 73 miles long, with the Lysaker River forming the boundary line of the municipality of Bærum, the most highly populated of the suburban municipalities in Akershus County. In addition to Akershus, the city of Oslo also adjoins the counties of Buskerud and Oppland.

Demography. In 1801 Oslo proper had 9,500 inhabitants: in 1875, 76,800; in 1948, 429,500; and in 1971, 481,000, together with a steadily increasing suburban population. Newcomers from other parts of the country made up over half of Oslo's population in the early 1970s, and Oslo also has many more foreigners than other Norwegian cities, although it is not as cosmopolitan in character as other western European capitals. In the central parts of the city the population is decreasing, and the average age of the inhabitants there is fairly high. In the developing suburban districts, the population is younger, and there is an increasing need for more schools; in the inner city, on the other hand, schools are being closed down and used for other purposes.

The expansion of industry in Oslo and a housing shortage within the city have caused a steadily increasing flow of commuters. New housing areas are being opened constantly in the outlying municipalities to assist families who are affected by this trend. As a result the population of greater Oslo reached 700,000 by 1970, while the Oslo region as a whole had in excess of 850,000 inhabitants.

Development of outlying areas

Economic life. Oslo is the focus of Norwegian trade, banking, industry, and shipping. The value of the city's imports in 1971 was 12,600,000,000 kroner, while exports amounted to 2,860,000,000 kroner (7.12 kroner = U.S. $1; 17.09 kroner = £1 sterling). Oslo Harbour is the largest, as well as the busiest, in the country, with a waterfront extending for eight miles. The city has over 130 shipping firms, which command a merchant fleet of around 7,000,000 gross tons, of which 4,300,000 gross tons are dry cargo tonnage and 2,800,000 gross tons are tank tonnage.

Industry in Oslo is primarily concerned with the domestic market, with industrial production constituting about a quarter of the total value of output for the entire country. In addition, Oslo firms pay out nearly 30 percent of the national industrial wages. The leading industries are the production of consumer goods, shipbuilding, and the electrotechnical and graphic industries. The most important national daily newspapers—including the *Aftenposten, Dagbladet,* and *Arbeiderbladet*—are also published in the city. Industrial districts have been sited along the Loelva and Akerselva from an early stage in the city's growth, while smaller industrial districts are to be found near the outlying communities of Bislett, Majorstna, and Skøyen.

Around 50 percent of Norway's wholesale trade, together with 25 percent of all retail business, is transacted in Oslo. The Norwegian Trade Fair exhibition hall is located in Skløyen, to the west of the city. The important Norwegian fur auctions are held at Økern, in the northeast of Oslo.

Administration. Oslo, as the capital city of Norway, contains numerous government offices and other public institutions. These include the parliament, the supreme court, the Bank of Norway, and the Norwegian broadcasting corporation.

The city council is made up of 84 representatives and a mayor, who is elected for two years. The city hall, which contains a number of municipal offices, is an impressive building with two massive towers. Situated near the harbour in the Vika district, it dominates the seaward approach to the city. A modern building, it has been richly decorated by 28 Norwegian artists.

Services. Oslo has two main railway centres. From the east station, main lines radiate to Bergen, Fagernes, Trondheim, Stockholm, and Gøteborg, while from the west station, the Sørland line leads to Stavanger. Freight is transported by rail along the harbour partly through tunnels alongside the Akershus Fortress. Electric suburban trains connect the city centre with the east, north, and west suburbs, with some of the suburban lines running in tunnels under the central section of the city.

Transportation improvements

The network of roads around Oslo is being extended rapidly, with major new stretches of modern highways opened every year, although, as in all parts of Norway, the difficult terrain makes road construction costly and arduous.

Automobile traffic in Oslo itself had also become a problem by the 1970s, especially in rush-hour periods. Some of the traffic is channelled along roads encircling the centre, but there is a growing need for a new motor road through the city. Such a road, in the planning stage in the early 1970s, would alter radically the face of the city.

Oslo's electrical energy requirements are set by deliveries from water-powered stations, the most important of which are located in Numedal, Hallingdal, Gudbrandsdalen, and at Glomma (or Glåma).

Cultural life and recreation. The leading Norwegian cultural institutions are located in Oslo. The city centre contains the National Theatre, the Norwegian Theatre, the Oslo Nye Teater, and Den Norske Opera. Near the oldest part of the university are found the Historial Museum (with the Ethnographic Museum) and the National Gallery, the latter containing an excellent collection of paintings. At Tøyen, in the east of the city, are the botanical gardens and several museums. At Bygdøy, the Norwegian Folk Museum, a Viking ship hall, the Framhuset (containing a famous polar exploration vessel), the Kon-Tiki Museum (commemorating the Pacific expedition of Thor Heyerdahl), and the Norwegian shipping museum serve as reminders of Oslo's maritime connections. Frogner Park, in the west section of the city, is justly famous for its impressive display of works by the modern sculptor Gustav Vigeland.

A number of scientific institutions are attached to the university, which had around 16,000 students by the 1970s; the university library is the main library in the country. The majority of university departments are located at Blindern, in the northwest of the city, while apartment buildings for students have been built just outside the university centre, at Sogn and Kringsjå. These students' apartments are used by tourists during the summer months. The city's most distinguished auditorium is the university's Aula (Hall), decorated by the painter Edvard Munch, and it is here that the Philharmonic Society's concerts are held. Oslo also has a number of other national centres of higher learning.

Winter sports

Oslo has exceptional natural advantages for winter sports, especially cross-country skiing; the annual Holmenkollen competitions bring together participants from all the leading ski nations, and the site also houses an interesting ski museum. The Bislett arena, located in the city centre, is an internationally famous speed-skating rink.

Oslo Fjord is used extensively for swimming, although water pollution has become a serious problem in the innermost portions. Boating and sailing are also popular sports on the fjord in summer. The freshwater lakes of Bogstad and Sognsvatn are also frequently visited by swimming enthusiasts, and popular camping-grounds for tourists are to be found at Ekeberg and Bogstad.

Considerable care has been taken by city planners to safeguard the recreational facilities surrounding Oslo: green belts or park areas have been kept free of development, and it is still possible to ski almost to the city centre from outlying regions.

BIBLIOGRAPHY

Topography: JOAN WRIGHT, *Norway* (1970), a guide, has a useful section on Oslo, while OLIVER WARNER, *A Journey to the Northern Capitals* (1968), is a travel book. See also A. ARSTAL, *Oslo byleksikon,* new ed. (1968). The OSLO TRAVEL ASSOCIATION, *Social Service, Oslo, Norway* (1960), highlights this particular aspect of the city; VILHELM BJORSET (ed.), *The Oslo Book* (1950), is mostly pictures.

History: GERHARD FISCHER, *Oslo Under Eikaberg* (1950), deals with the period 1050–1624, while F.N. STAGG, *East Norway and Its Frontier: A History of Oslo and Its Uplands* (1956), has useful later material. A.J. STENSENG, *Akershus Castle: Official Guide with a Brief Historical Survey* (1950), deals with a part of the city. The OSLO UNIVERSITET, *Facts About the University of Oslo* (1967), supplements C.F. ENGELSTAD (ed.), *Oslo, the Capital of Norway: Art and Intellectual Life at Its 900-Years Jubilee* (1950).

Scientific aspects and economics: The *Statistical Yearbook for the City of Oslo* is a useful source of figures, while the OSLO BOURSE COMMITTEE and the OSLO CHAMBER OF COMMERCE, *Report of the Trade, Industry and Shipping of Oslo* (annual), is a serviceable guide to past patterns of commerce. OLAF HOLTEDAHL and JOHANNES A. DONS, *Geological Guide to Oslo and District* (1957), is instructive. See also T.F. RASMUSSEN, *Storbyutvikling og arbeidsreiser. En undersøkelse av pendling, befolkningsutvikling, neeringsliv og urbanisering i Oslo-området* (1966), on metropolitan growth, commuting, and urbanization in the Oslo area (with an English summary).

(Öi.R.)

Ostariophysi

The fishes of the superorder Ostariophysi include the majority of freshwater fishes throughout the world. Familiar representatives of this group are the minnows, suckers, characins, loaches, gymnotid "eels," and innumerable catfishes. The 31 recognized families of catfishes constitute the order Siluriformes, the remaining 26 families the order Cypriniformes. Man consumes huge quantities of these fishes for food and derives pleasure from the beauty of tropical aquarium fishes. A few harmful species can inflict painful injuries; some others serve as intermediate hosts for parasites of man. Strange and fascinating

behaviour is exhibited by many of these fishes—nest building, oral incubation, egg laying in mollusk shells, walking and flying, air breathing, production of sound and electricity, and communication by chemical secretions.

GENERAL FEATURES

Size range and diversity of structure. Most ostariophysians are small to moderate in size, from two to 30 centimetres (about one to 12 inches) long; others rank among the giants of the freshwater world. The elegant mahseer (Cyprinidae) of Asia grows to two metres (6½ feet) and 90 kilograms (200 pounds); the wels, a Eurasian catfish (Siluridae), attains 4.5 metres (15 feet) and 30 kilograms (660 pounds). The extent of morphological diversity is at least as great as that in any other group of living vertebrates.

Ostariophysians abound in nearly all freshwater habitats, including subterranean caverns and those on all major land masses and continental islands of the world except for Greenland and Antarctica. A few invade brackish waters, and two families consist largely of marine species. Approximately 6,000 species are recognized, about one-fourth of all known species of fish. Their undisputed success may be attributed at least in part to two remarkable features: a sense of hearing more acute than that in any other group of fish and a warning system by chemical communication unique among fishes.

Importance to man. Many moderate to large ostariophysians are utilized for food, and commercial fisheries harvest huge quantities of marketable species. The common carp (*Cyprinus carpio*), originally from China, has been introduced nearly worldwide and is extensively cultured in the warmer regions. Other Chinese carp under cultivation include the grass carp (*Ctenopharyngodon*), silver carp (*Hypothalmichthys*), snail carp (*Mylopharyngodon*), and bighead carp (*Aristichthys*). Culture of the channel catfish (*Ictalurus punctatus*) is an important industry in the southern United States. Numerous ostariophysians provide sport fishermen with recreation and food; several rank among the world's prized game fish; *e.g.*, mahseers (several species of *Barbus*) of Asia and the dorado (*Salminus maxillosus*) of South America.

Ostariophysians as aquarium fish

The tropical-fish industry has multiplied phenomenally since World War II. In 1968 more than 60,000,000 tropical fishes were imported into the United States. Among the most popular are the characins, tetras, rasboras, danios, barbs, loaches, and innumerable catfishes. Adaptability of many ostariophysians to aquarium life has resulted in their widespread use as experimental animals in scientific research. Foremost among these are the goldfish (*Carassius auratus*) and the common carp.

In eastern Asia and parts of Europe, humans frequently become infected with liver flukes acquired by eating raw or imperfectly cooked fish. The carps, especially *Ctenopharyngodon idellus*, are the second intermediate host of the Chinese liver fluke (*Clonorchis sinensis*). Many cyprinids serve as intermediate hosts for the cat liver fluke (*Opisthorchis felineus*). Domestic animals similarly become infected with flukes and tapeworms.

Some ostariophysians are pests or are potentially dangerous to man. The common carp is a nuisance in many localities in the U.S. Introduced species such as the walking catfish (*Clarias batrachus*) pose a serious threat to the native fauna. In South America, on occasion, the piranha (*Serrasalmus*) voraciously attacks man and domestic animals, and the diminutive candiru (*Vandellia cirrhosa*) can penetrate the urogenital openings of human bathers and cause intense pain and hemorrhaging.

NATURAL HISTORY

Behaviour. *Reproductive cycle.* Like most fishes, ostariophysians are bisexual; eggs develop in the ovaries of the female and spermatozoa (milt) in the testes of the male. In temperate zones most species breed in the spring, when water temperatures are rising and day lengths (photoperiods) are increasing. In tropical regions many fishes spawn the year round. All ostariophysians lay eggs; none give birth to living young. Eggs are fertilized externally in the water in all species except a few auchenipteriod catfishes and the South American characins of the subfamily Glandulocaudinae in which fertilization is internal. Development is direct; newly hatched individuals do not pass through morphological changes (metamorphosis) but instead are miniature replicas of their parents. The age at sexual maturity depends on the species and relative body size. Many small species reproduce when only a few months old and rarely live more than one to a few years. Large species attain sexual maturity when several years old, and, in captivity, the common carp has been known to live more than 50 years.

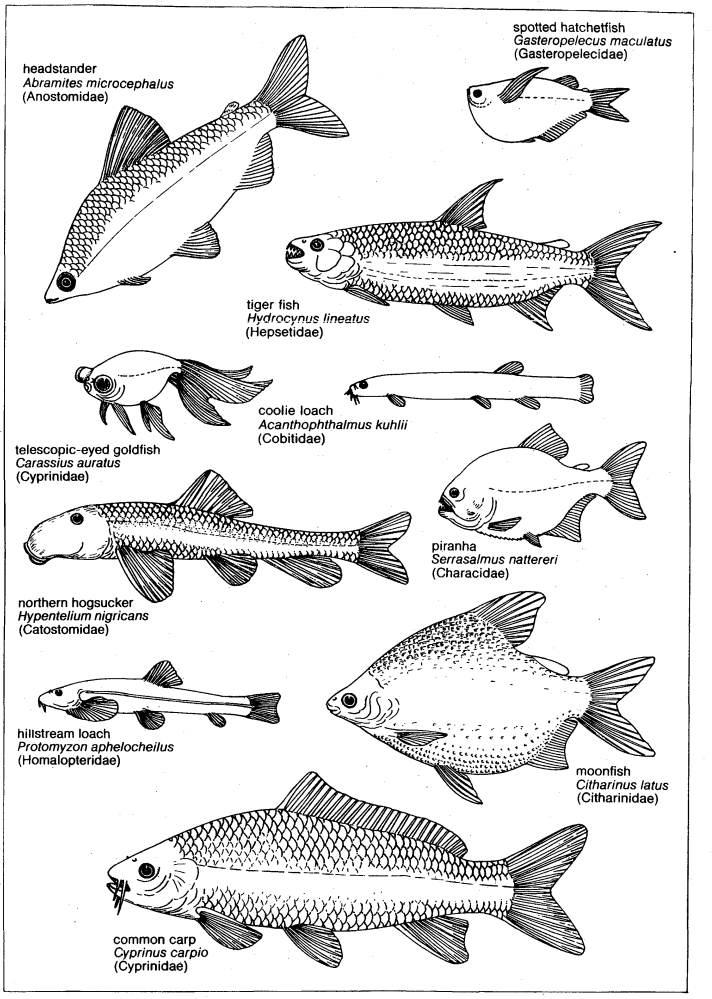


Figure 1: Body plans of representative ostariophysian fishes.
Drawing by D.P. Janson

Breeding. Distinct pairing occurs in most ostariophysians, and courtship behaviour in characoids and cyprinoids often consists of elaborate displays by males in brilliant nuptial coloration. The eggs are heavier than water (demersal) and sink; most are sticky and adhere to the surface or to various objects. Characins and cyprinids generally deposit their eggs among aquatic plants, under stones and logs, or in shallow pits in gravel and sand. Among the many exceptions is the characin *Copeina arnoldi*; the female actually leaps out of the water to lay her eggs on the undersides of overhanging leaves (or, in captivity, of aquarium covers), to which she clings, joined by the male, during egg deposition. The parents then splash water on the eggs during development. The female bitterling (*Rhodeus sericeus*) deposits its eggs in the gill cavity of freshwater mussels by means of an elongated ovipositor, which she inserts into the mussel's incurrent siphon. Catfishes choose breeding sites in streams and ponds, generally in quiet water among plants or on mud, sand, gravel, or debris. The nest may be a simple circular depression (as in bullheads) or a tunnel-like affair in the bank (as in the channel catfish). Migrations comparable

Egg laying on leaves out of water

to those of salmon and eels are unknown among the ostariophysians, but the tendency to migrate occurs among suckers (Catostomidae), which swarm upstream into small tributaries and spawn over gravel or sand bottoms, and in other riverine species, as the mahseer (*Barbus tor*) and the African tiger fish (*Hydrocynus*).

Parental care. Although many species exhibit no parental care, nest building and egg guarding are widespread among this group of fishes. Some cyprinids, such as the chubs (*Nocomis*), build massive pyramidal nests of stones; they desert the nests once spawning is completed. Other species of breeding minnows often swarm over these nests, and hybrids frequently are produced by the mixing of eggs and sperm from different species. The eggs of characids are commonly guarded by the male. Catfishes provide their eggs with considerable protection, either by guarding nests or by carrying the eggs with them. Oral incubation is practiced in sea catfishes (Ariidae); the male carries from ten to 50 marble-sized eggs in the mouth cavity until hatching. The male continues to protect the hatchlings in his mouth even after the young have begun to feed independently. In certain species of banjo catfishes (Aspredinidae), the eggs are anchored to spongy tentacles on the underside of the female's abdomen. Some female callichthyid catfish carry eggs on the abdomen only for fertilization; others deposit their adhesive eggs in froth nests and guard them. The loricariid catfishes employ various methods; some lay adhesive eggs in cavities, others carry them under the lower lip, and a few deposit them on rocks, where they are cleaned, fanned, and guarded by the male.

Defense. Ostariophysians with bright colours and gaudy patterns are popular among tropical-fish fanciers; however, many other small species are somberly coloured, relying on this protective coloration for passive defense from enemies. Large carnivorous forms such as the African tiger fish and the South American piranhas have powerful jaws and strong teeth, extremely effective weapons for defense as well as for offense. Most catfishes and some Old World cyprinids possess spines (hardened fin rays) in the dorsal and pectoral fins. The spines alone afford a considerable degree of protection; in addition, venom glands develop at the base of the spines in some bullheads and madtoms of North America (Ictaluridae), labyrinthic catfishes (Clariidae), and sea catfishes (Ariidae and Plotosidae). Painful but rarely fatal injuries result when the skin of a victim is punctured and venom injected.

Protection by poisonous spines

Although a variety of freshwater fishes can generate an electrical charge, only two develop sufficient voltage to stun other animals, including man—the electric eel (*Electrophorus electricus*) and the electric catfish (*Malapterurus electricus*).

Ecology. *Habitat and distribution.* Ostariophysians are the dominant fishes in virtually all types of freshwater habitats throughout the tropical, temperate, and sub-Arctic regions of the world. Only a few species of the families Cyprinidae and Aspredinidae are known to invade low-saline or brackish waters. The only truly marine members of this superorder are the sea catfishes (Ariidae and Plotosidae), which inhabit tropical coasts. Some plotosids, however, live in freshwater.

The upper regions of small mountain streams are characterized by steep gradients, waterfalls and rapids, and torrential currents. Here occurs a variety of ostariophysians (Homalopteridae, Sisoridae, Akysidae, Loricariidae, Astroblepidae), which exhibit fascinating structural adaptations, such as holdfast organs and specialized respiratory mechanisms. In river systems where the gradients are not steep, currents are slow and quiet pools alternate with riffles, large numbers of characins, cyprinids, and suckers and other types of catfish are conspicuous elements in the fauna. In the sluggish waters of large rivers live large species of suckers, cyprinids (*e.g.*, carp), and many catfishes generally characterized by environmental tolerances and nonrestrictive feeding habits. Ponds and lakes also support large populations of characids, cyprinids, catostomids, and siluroids that prefer and are adapted to standing-water habitats. Although a few are benthic (bottom-dwelling) forms, most of the characins and cyprinids tend to live and feed in the middle and upper layers of the water column. Suckers, loaches, and most catfishes are typically benthic animals and thus are highly adapted to such an existence. Catfishes are generally most active at night or under conditions of reduced light intensities.

Among the most unusual habitats for fishes are those in subterranean waters, wells, and caves. A relatively large number of ostariophysians, belonging to unrelated families, present a striking example of convergent adaptation to life in more or less total darkness. The evolutionary trends have led to a reduction or loss of eyes, loss of pigment, and special development of certain sense organs, especially the lateral-line system, to compensate for the loss of sight. Ostariophysians adapted to such a mode of life include six genera of cyprinids in Africa, the Middle East, and Java; a characin (*Astyanax jordani*) from Mexico; ictalurids from the U.S. (*Trogloglanis* and *Satan*) and Mexico (*Prietella*); six genera of pimelodids and trichomycterids from South America; and two genera of clariids from Africa.

Cave fishes

Feeding habits. The remarkable diversity of feeding habits among ostariophysians is associated with a fantastic variety of adaptations in mouth shapes and tooth types (especially in the suborder Characoidei) probably unsurpassed by any other group. At one extreme are certain cyprinids (*e.g.*, *Notropis atherinoides*) with highly developed gill rakers that strain phytoplankton (minute plants) from the water. Mountain-stream fishes (*e.g.*, Gyrinocheilidae, Homalopteridae, Loricariidae) possess sucker-like lips for scraping algae from the rocks; their teeth are minute or entirely lacking. Because they devour large quantities of plants, herbivores such as the Chinese grass carp are used experimentally to control vegetation in weed-choked waters. Omnivores are especially common among the characins and catfishes. Suckers, long-snouted knife fishes, many catfishes, and some minnows suck up mud and bottom debris, extract the nutriments, and eject the residue. Small carnivorous species consume insect larvae, small crustaceans, worms, mollusks, and other invertebrates. At the top of the food chain are the voracious predators, the most famous of which are the piranhas. Although modest in size, they have short, powerful jaws armed with razor-sharp teeth. These fearsome predators often occur in large schools and can quickly strip the flesh from their victims. Other fishes are their usual prey, but cattle and occasionally humans are also attacked. Probably the largest predatory ostariophysian is the tiger fish, which attains a weight exceeding 45 kilograms; its huge, sharp teeth and large, tuna-like tail endow it with ferocity and speed. Parasitic habits are rarely found among bony fishes, but certain species of trichomycterid catfishes attach themselves to the gills of other fishes and feed on their hosts' blood.

Feeding habits of piranhas

FORM AND FUNCTION

Distinguishing characteristics. *Weberian apparatus and swim bladder.* The single character unique to the superorder Ostariophysi is the presence of the so-called Weberian apparatus, a complex connection between the inner ear and gas bladder (swim bladder). It is formed by the modification of the first four (or five) vertebrae immediately behind the skull, small portions of which have become separated and form a chain of four paired bones, or ossicles, named from front to back the claustrum, scaphium, intercalarium, and tripus. The first is in contact with a membranous window, or extension of the inner ear; the last touches the anterior wall of the swim bladder. The diverse modifications of the Weberian apparatus are diagnostic of orders and certain families; *e.g.*, the claustrum is absent in Gymnotidae. Although much remains to be learned about its functions, it is known to serve as a hearing organ. Changes in volume of the swim bladder due to sound waves in the water cause the ossicles to move and transmit pressure changes to the ear.

The gas bladder varies in shape and size but typically consists of two, sometimes three, chambers. In bottom-dwelling fishes, such as the Homalopteridae, Cobitidae,

and many catfishes, the posterior chamber is greatly reduced and the anterior one often more or less surrounded by a bony capsule. In some catfishes (Sisoridae), only the anterior chamber remains, and it may be encapsulated with bone.

Body covering. The nature of the body covering is variable. Most cypriniforms possess cycloid scales (smooth, overlapping scales more or less circular in shape). Exceptions are found among the Ctenoluciidae, Distichodontidae, Citharinidae, and Ichthyboridae, which have ciliate, or ctenoid, scales (*i.e.*, posterior margins of scales with fine teeth). Most catfishes have lost the scaly covering and are naked, but several families possess bony plates forming an overlapping armour on the sides of the body (Doradidae, Callichthyidae, Loricariidae).

Bony armour

Fin spines and adipose fin. Ostariophysians possess segmented, branched, flexible, soft rays in the fins, unlike the stiff spines of perchlike fishes. In some species, however, soft-ray elements may fuse during development and give rise to a spinous ray (usually called a spine), commonly found in the dorsal and pectoral fins of most catfishes and in the dorsal and anal fins of some Old World cyprinids. The presence or absence of these spines is frequently diagnostic for genera and families.

An adipose fin consists of a small to elongated fleshy or fatty structure without fin-ray supports, located dorsally between the rayed dorsal fin and caudal (tail) fin. It is present in most ostariophysian fishes.

Barbels. Diverse morphological differences in the mouth region are related to the type of diet and to the modes of locating, capturing, and ingesting food. Barbels are short to filamentous, fleshy, fingerlike projections located at the corners of the mouth or on the snout and chin of many suctorial and bottom-feeding fishes (some minnows, loaches, and catfishes). Highly sensitive to touch, they bear numerous taste buds. Taste and touch probably function together in the selection of food before ingestion.

Teeth. Teeth may be present along the jaws, in the roof of the mouth, on the tongue, or in the pharynx, or they may be entirely absent. In the minnows (Cyprinidae) and suckers (Catostomidae), the mouth is toothless, but an array of teeth is borne on a pair of branchial bones, the lower pharyngeals, located in the throat. In the minnows the pharyngeal teeth, arranged in one, two, or three rows, press or bite against a horny pad in the roof of the mouth. They have undergone specialization paralleling the diversity found in jaw teeth of other fishes. Vegetarians such as the carp have grinding, molar-like teeth; carnivores have pointed or hooked teeth. Suckers have numerous pharyngeal teeth aligned in a single row. Oral and pharyngeal teeth are of great value in classifying many families of ostariophysians.

Secondary sexual characteristics. With the onset of the breeding season, many secondary sexual characteristics develop: size differences, nuptial coloration, enlarged and modified fins, breeding tubercles, and contact organs. These features are related chiefly to courtship and mating, but differences in size obviously play a role in guarding nests and care of the young; the sex that exercises parental care is usually the larger. Brilliant red, orange, yellow, green, and blue coloration may develop on various parts of the head, body, and fins, especially in the males. Some characids and cyprinids are among the most beautiful of all fishes. The male usually has larger and more brightly coloured fins than the female. In some characins, the median and pelvic fins of the males may possess small hooks or contact organs, which aid in maintaining contact with the female during spawning. In the six families of cyprinoids, breeding tubercles, or pearl organs (epidermal excrescences), develop on the head, body, and fins of males under the influence of sex hormones. The tubercles function in maintenance of body contact during spawning, in defense of nests and territories, and possibly in the stimulation of females during breeding.

Sexual coloration

Sexual differences among the siluroids are more marked in the highly specialized families. Pelvic fins of female ariid catfishes and, to a lesser extent, of ictalurid catfishes show specialized developments whose functions are not yet fully known. Some male loricariid catfishes develop elaborate dermal, branching growths and spines around the head; in others, the lower lip is enlarged to accommodate the transport of eggs.

Drawing by D.P. Janson

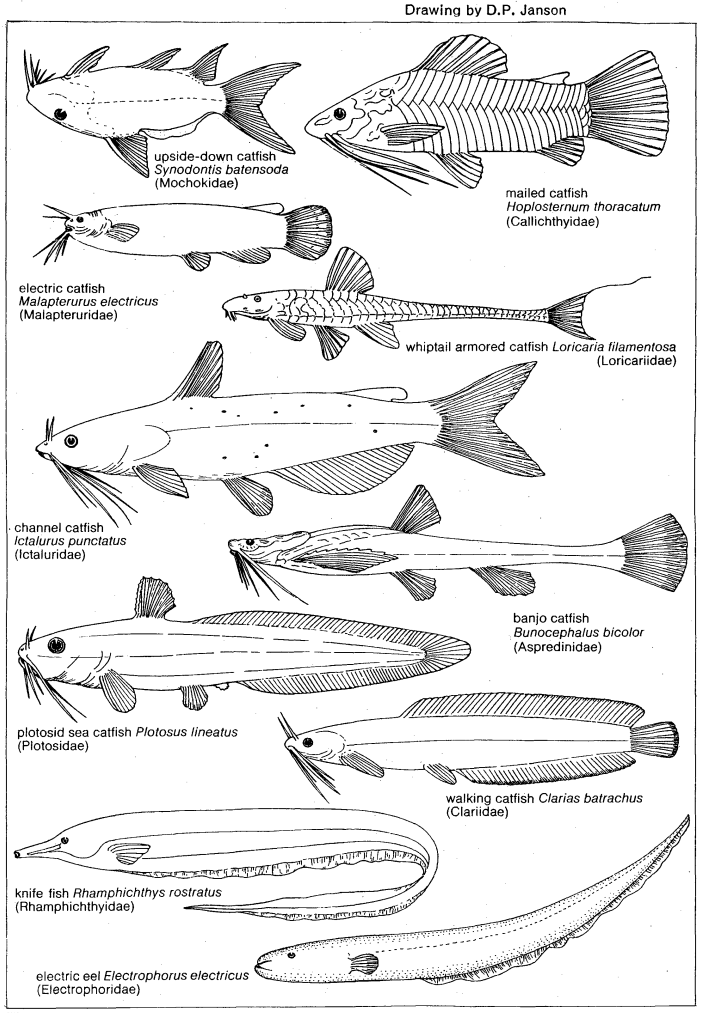


Figure 2: Body plans of representative ostariophysian fishes.

Adaptations for locomotion. *Swimming.* The body of most ostariophysian fishes is more or less streamlined, taking the most efficient form for movement through water. In this highly diversified group, however, a large array of adaptations occurs. Lateral compression (flattened from side to side) is common, especially among characins and cyprinids that inhabit quiet, weedy lakes, ponds, and backwaters. Extreme examples are the flying hatchetfishes (Gasteropelecidae) and the knife fishes (Rhamphichthyidae and Apteronotidae). Depressed body form (flattened from top to bottom), especially in the head region, is widespread among fishes spending much time on or near the bottom or under rocks and similar objects (most catfishes) or among those inhabiting torrential mountain streams (Homalopteridae, some Loricariidae). An elongated, eellike form has evolved in certain loaches (Cobitidae) and electric eels (Electrophoridae), fishes that live on soft, muddy, and sandy bottoms or in rock crevices.

The most common form of locomotion among ostariophysians is swimming by lateral undulations of the body, resulting from the contractions of muscles along the sides of the body and base of the tail. These undulating flexures culminate in a powerful back and forth sweeping of the caudal fin, which produces as much as 85 percent of the total thrust. Some fishes have departed from the normal horizontal swimming posture. The headstanders (Anostomidae) move with the head pointing downward at a slant; some of the pencil fishes (Hemiodontidae) assume a tail-standing position. Most bizarre of all are the upside-down catfishes (Mochokidae) of Africa, which can

Departures from normal posture

swim either in the normal position or inverted, with the belly uppermost. In *Synodontis batensoda,* the coloration of the belly is darker than the back, a reversal of the usual pigmentation pattern. Displacement of the gas bladder toward the underside is a further adaptation to this unusual swimming behaviour.

In fishes with specialized modifications of body form and habits, the fins are frequently modified and used for propulsion. The electric eels and knife fishes (Gymnotoidei) have lost the dorsal fin and, in some cases, the caudal fin. Slow forward and backward movements are made possible by undulations of an extremely long anal fin.

Associated with locomotion is the need for maintaining position in the water, particularly in the rapid torrents of mountain streams. A variety of modifications have evolved that function as holdfasts, anchoring the fish to rocks or similar objects. The hill-stream loaches (Homalopteridae) of southeastern Asia possess a large ventral suction disk formed by the expanded pectoral and pelvic fins. Some of the mountain-stream catfishes (Sisoridae) of Asia have an adhesive organ on the thorax (chest). Mountain-inhabiting catfishes of South America may use a sucker-like mouth (Loricariidae) or employ a combination of a disklike mouth and disklike paired fins (Astroblepidae) for adhesion to the surface.

Walking and flying. A few ostariophysians have evolved the capability to emerge from their aquatic abode and move over land, climb walls, or even fly through the air. The walking catfish (*Clarias batrachus*), recently introduced into southern Florida, uses its pectoral-fin spines as anchors to prevent jackknifing as its body musculature produces snakelike movements and can progress remarkable distances over dry land. Using suction disks and fins, the mountain-stream catfishes (Sisoridae and Astroblepidae) can climb vertical rock walls above the water surface.

Flying by characins

The small hatchetfishes, or flying characins (Gasteropelecidae), of South America normally swim near the surface of the water but are capable of jumping clear and flying short distances. They vibrate enlarged pectoral fins rapidly back and forth by using highly specialized musculature on the shoulder girdle.

Air breathing. Although gills are typical respiratory structures in fishes, many freshwater species occupy habitats where the oxygen may be depleted occasionally or where droughts may force them to live out of water temporarily. These fishes have evolved a variety of air-breathing organs, most of which are outgrowths or pouches from the pharynx, branchial (gill) chamber, or digestive tube. Some catfishes (*Clarias* and *Heterobranchus*) of Asia and Africa have treelike respiratory structures extending above the gill chambers; others (*Heteropneustes*) have elongated, tubular, lunglike sacs extending backward as far as the tail. The electric eel is a mouth breather, gaseous exchange taking place through the wrinkled mucous membrane lining the mouth cavity. Some fishes actually swallow air into the lower part of the digestive tract, which then also serves as a respiratory structure. In the armoured catfishes (*Doras, Plecostomus, Callichthys*) of South America the thin-walled stomach serves this function. The loaches swallow air into a reservoir-like bulge from the intestine and void the remaining gases through the anus.

Communication and sensory perception. *Sound.* Sounds produced by ostariophysians are usually associated with the swim bladder. Minnows produce noises by expelling air through the pneumatic duct, which connects the gas bladder with the digestive tract, and the mouth; loaches do the same by expulsion through the anus. In several catfish families the expanded ends of a spring-like mechanism (derived from modified portions of the fourth vertebra) are attached to the swim bladder. The contraction of muscles extending from the spring mechanism to the skull cause the springs and bladder wall to vibrate rapidly, producing a growling or humming noise. In other catfishes the rubbing or grating movements of the dorsal and pectoral spines produce sounds.

The sense of hearing in Ostariophysi is more highly developed than in any other fishes. The walls of the gas bladder are set in vibration by waves of underwater sound, and the Weberian ossicles then increase the amplitude of these vibrations, transmitting them to the internal ears. This combination is analogous to that of a hydrophone and endows these fishes with a remarkable sensitivity to sound. The normal frequency range detectable by ostariophysians is from 16 to 7,000 hertz (cycles per second); for some characins the maximum is 10,000 hertz. Among other functions, sound production and hearing in fishes may assist in bringing schooling fishes together; even more significant is the role of sound in reproduction. Experiments with North American cyprinids provide evidence that sounds are produced by both sexes and may serve for sexual recognition. A male is able to distinguish the calls of females of his own species from those of closely related species. Consequently, sounds may serve as isolating mechanisms in maintaining the genetic integrity of the species. For fishes living in muddy waters, sounds may be a vital communication link between individuals, especially in the breeding season. It is reasonable to suggest that the combination of sound production and acute hearing may have aided the ostariophysians in attaining their dominant role in freshwaters.

Hearing range

Electric organs. Members of the suborder Gymnotoidei and of the siluriform family Malapteruridae possess the unusual capacity to generate electricity. The best known and most powerful of this group is the electric eel (*Electrophorus electricus*). The electrical organs, three on each side of the body, are derived from modified muscle tissue. The force of the discharge has been measured at 350 to 650 volts and can produce a current strong enough to stun animals as large as a horse or man. The electric catfish (*Malapterurus electricus*) can deliver shocks up to 450 volts, but this power is apparently used only as a defensive measure. The electrical organ of this species, also derived from muscle tissue, consists of a specialized, gelatinous coat of tissue that sheathes most of the body just under the skin.

Production of electric currents

The gymnotid eels and knife fishes (*Gymnotus* and other genera) produce currents of low voltage only, emitting a continuous series of pulses (from 35 to 1,700 hertz), which create an electrical field around the fish. When this field is broken, either by a moving animal or by inanimate objects in the vicinity, the fish can locate the objects, which otherwise would be difficult to see at night or in muddy water. Experiments indicate that electrical cues may also facilitate social interactions. Perception of electrical stimuli occurs in specialized electrical receptors in the skin, and portions of the brain are enlarged to process electrosensory information.

Taste and smell. Catfishes and other fishes living in muddy waters have relatively poor vision but possess chemosensory acuity. Lips, barbels, and most of the body are covered with innumerable taste buds. Experiments have proved that taste plays a leading role in the location of food by these fishes.

Recent studies on the sense of smell have yielded interesting results. Odours emanating from mucus produced in the skin, from secretions of the gonads, and from other body parts serve as chemical signals called pheromones. These odours provide a means of communication between individuals of the same or different species. Certain minnows (Cyprinidae) can discriminate between the odours of at least 15 species of fishes belonging to eight different families. The social behaviour of bullheads (*Ictalurus*) and other ostariophysians is related to a system of communications utilizing chemical signals. An individual not only recognizes individuals of other species but can identify and remember the identification of a particular individual of its own species after a time lapse of three weeks. Territorial and communal behaviour are evidently influenced by different pheromones.

Reactions to wounded companions

Alarm substances. In 1938 an Austrian biologist, Karl von Frisch, introduced an injured minnow (*Phoxinus*) into a school of the same species and observed that the school rapidly retreated and became very frightened. By experimentation he demonstrated that a chemical substance released from the lacerated skin produced a fright reaction when perceived through the nasal organs of

other fishes. This "alarm substance," secreted by specialized cells in the epidermis, is released only when the skin is injured. Alarm substances are present in almost all species of ostariophysians tested (except for a few species of Characidae, Hemiodontidae, Chilodontidae, and Rhamphichthyidae) and are absent in all nonostariophysian fishes examined. Although the fright reaction appears to be important insurance for the individual against predation, the alarm substances are of greatest value among those species exhibiting social behaviour by warning other members of the school. It is possible that alarm substances and the fright reaction have contributed markedly to the biological success of the Ostariophysi.

CLASSIFICATION

Distinguishing taxonomic features. Many characteristics are useful in classifying this large, diverse superorder —the nature of the body covering; presence or absence of barbels, fin spines, and adipose fin; modifications of mouth and fins; types of teeth. Less obvious but especially significant are numerous skull features, specializations of the Weberian apparatus, configuration of the gas bladder, and fusions of vertebral elements.

Annotated classification. This classification, a recent revision by P.H. Greenwood (U.S.) and colleagues, raises to family rank many groups of fish often treated as sub-families. The smallest families are grouped for brevity or are included under a closely related family.

SUPERORDER OSTARIOPHYSI

Gas bladder and internal ear connected by chain of ossicles (Weberian apparatus). All inhabit freshwater unless otherwise noted.

Order Cypriniformes

Body usually covered with cycloid scales; about 3,500 species.

Suborder Characoidei

Cretaceous (about 136,000,000 years ago to present). Mouth not protractile; jaws toothed. Characidae most generalized; other families have specialized skeletal structures, jaws, and teeth.

Family Characidae (characins). Fresh to brackish waters; Africa, South and Central America. Tremendous morphological and ecological diversity. Many brilliantly coloured. Variable food habits. Popular aquarium and food fishes. Size 2.5–150 centimeters (1–60 inches). Examples: tetras, piranhas.

Families Erythrinidae, Ctenoluciidae, and Cyndontidae. South America. Large mouths, canine teeth. Erythrinidae lacks adipose fin; Ctenoluciidae has ciliated scales; Cynodontidae, long anal fin. Carnivorous. Food fishes for man. Size to 120 cm (4 ft).

Family Hepsetidae. Africa. Pikelike; large canine teeth. Carnivorous. Food fishes. Size to 100 cm (40 in.), 55 kilograms (120 pounds).

Family Lebiasinidae. South and Central America. Lateral line and adipose fin usually absent. Small to moderate-sized predators.

Family Gasteropelecidae (hatchetfishes). South and Central America. Deep, strongly compressed body; pectoral fins with well-developed musculature. Capable of true flight. Insectivorous. Aquarium fishes. Size to 10 cm (4 in.).

Family Anostomidae (headstanders). South America. Elongated snout; small mouth with folded or fleshy lips or sucking disk. Head-standing habits. Herbivorous. Aquarium and food fishes. Size to 40 cm (16 in.). The South American families Prochilodontidae (predorsal spine, rough scales), Curimatidae (toothless jaws), and Chilodontidae (specialized pharyngeal teeth) are similar to the Anostomidae.

Family Hemiodontidae (pencil fishes). South and Central America. Lower jaw toothless. Tail-standing posture. Herbivorous. Aquarium fishes. Size to 20 cm (8 in.). Family Parodontidae is similar.

Family Citharinidae (moonfishes). Africa. Deep-bodied, scales often denticulate (toothed), small mouth and teeth. Herbivorous. Aquarium and food fishes. Size to 90 cm (3 ft). The African families Distichodontidae (upper jaw not or scarcely movable) and Ichthyboridae (slender body, upper jaw freely movable, carnivorous) are similar but have ctenoid (ciliate) scales.

Suborder Gymnotoidei

No fossil record. Body elongated; anal fin very long; electric organs present.

Families Gymnotidae (gymnotid eels), *Apteronotidae, and Rhamphichthyidae* (knife fishes). South and Central America. Body greatly compressed, scaled. Weak electrical powers. Rhamphichthyidae with elephant-like snout; herbivorous. Other families carnivorous. Size to 90 cm (3 ft).

Family Electrophoridae (electric eels). South America. Body eellike, scaleless. Powerful electric organs. Size to 275 cm (about 9 ft), weight to 22 kg (48 lb).

Suborder Cyprinoidei

Paleocene to present. Mouth toothless, protractile. Adipose fin rarely present.

Family Cyprinidae (minnows and carps). Most in fresh but some in brackish water; Asia, Europe, Africa, North America. Pharyngeal teeth in 1 to 3 rows. Some with 1 or 2 pairs of small barbels. Food habits variable. Food fishes of sport and commercial value; aquarium fishes. Size 2.5–250 cm (1 in. to more than 8 ft). Examples: minnows, carp, goldfish, barb, bitterling.

Family Catostomidae (suckers). North America, Asia. Protractile, sucking mouth on underside of head. Detritus feeders. Food fishes. Size to 90 cm.

Families Gyrinocheilidae (algae eaters), *Psilorhynchidae, and Homalopteridae* (hill-stream loaches). Mountain streams, Asia. Adaptations to fast currents include fleshy, suctorial mouth and inhalant–exhalant gill openings (Gyrinocheilidae); ventral sucking disk formed by paired fins (Homalopteridae). Algae feeders. Size to 10 cm.

Family Cobitidae (loaches). Asia, Europe, Africa. Wormlike; scales minute or absent; barbels 3–6 pairs. Intestine sometimes modified for aerial respiration. Mostly carnivorous. Aquarium fishes. Size to 30 cm.

Order Siluriformes (catfishes)

Paleocene to present. Body naked or covered with bony plates; adipose fin usually present; pectoral and dorsal fins often with spines. Mostly omnivorous. About 2,500 species.

Family Diplomystidae. South America. One pair of barbels; primitive Weberian apparatus. Size to 24 cm.

Family Ictaluridae (North American freshwater catfishes). Few enter brackish water. North America; widely introduced. Barbels 4 pairs; some with venom glands. Valuable food fishes (sport and commercial). Size to 170 cm (67 in.), 50 kg (110 lb). Examples: bullhead, channel catfish.

Family Bagridae. Asia, Africa. Similar to Ictaluridae but with elongated adipose fin. Food, aquarium fishes. Size to 90 cm.

Family Siluridae. Asia, Europe, Africa. Body compressed; adipose fin lacking, anal fin very long; short dorsal fin (often lacking) without spine. Food aquarium fishes. Size to 400 cm (13 ft), 300 kg (660 lb). Examples: wels, glass catfish.

Family Schilbeidae. Asia and Africa. Similar to Siluridae, but with adipose fin usually present and spine in dorsal fin. Food fishes. Size to 230 cm (91 in.), 110 kg (240 lb).

Families Amblycipitidae and Akysidae. Asia. Similar to Bagridae but with reduced gas bladder. Akysids inhabit mountain streams, have tuberculated skin. Small size.

Family Amphiliidae. Africa. Similar to Bagridae, but paired fins expanding horizontally for adhesion in fast currents. Size to 21 cm (8½ in.).

Family Sisoridae (mountain-stream catfishes). Asia. Ventral surface flat; thorax with longitudinal plates or adhesive organ. Size to 30 cm (12 in.).

Family Clariidae (labyrinthic catfishes). Asia, Africa; widely introduced elsewhere. Long dorsal and anal fins without spines; adipose fin usually lacking. Treelike air-breathing organ. Food fishes. Size to 130 cm (51 in.). Example: walking catfish. The similar family Heteropneustidae has long, hollow air sacs.

Families Cranoglanididae, Pangasiidae, Chacidae, and Olyridae. Small Asian families, each containing 1 to several species.

Family Malapteruridae (electric catfishes). Africa. Rayed dorsal fin lacking; spines lacking. Electric organs. Food fishes. Size to 120 cm (47 in.), 23 kg (50 lb).

Family Mochokidae (upside-down catfishes). Africa. Bony shield on head and nape. Some swim upside down. Food fishes. Size to 60 cm (24 in.).

Families Ariidae and Plotosidae (sea catfishes). Marine, a few entering freshwater. Tropical coasts; Plotosidae restricted to Indo-Pacific. Nasal barbels lacking; oral incubation of eggs in Ariidae. Adipose fin lacking; long anal and caudal fins (confluent in Plotosidae). Food fishes. Size to 115 cm (45 in.).

Family Dordadidae (thorny catfishes). South America. Overlapping plates cover sides of body. Intestinal modifications for aerial respiration. Aquarium fishes. Generally small, to 100 cm (40 in.). The related family, Auchenipteridae, has naked flanks and apparent internal fertilization.

Family Aspredinidae (banjo catfishes). A few enter brackish waters and salt waters. South America. Adipose lacking; broad, flat head; large tubercles on naked body. Aquarium fishes. Size to 30 cm (12 in.).

Family Pimelodidae. South and Central America. Similar to Bagridae, but lack nasal barbels. Food, aquarium fishes. Size to 130 cm (51 in.), 65 kg (145 lb). Families Ageneiosidae (maxillary barbels only), Hypophthalmidae (1 species, toothless), and Helogeneidae (1 species, no dorsal spine) are South American families similar to the Pimelodidae.

Family Trichomycteridae (parasitic catfishes). South America. Operculum (gill cover) usually with spines. Many parasitic. Size to 10 cm (4 in.). Example: candiru. The similar family Cetopsidae lacks opercular spines.

Families Callichthyidae (mailed catfishes), *Loricariidae* (armoured catfishes), *and Astroblepidae.* South and Central America. Two longitudinal series of overlapping bony plates in Callichthyidae. Three or 4 rows of bony scutes in Loricariidae. Skin naked in Astroblepidae. Sucking mouth (Loricariidae), or mouth and fins modified for adhesion to rocks in mountain streams (Astroblepidae). All herbivores, closely related. Aquarium fishes. Size to 75 cm (30 in.).

Critical appraisal. Ostariophysians are relatively primitive bony fishes, singularly distinct from all other fishes except the Gonorynchiformes. Their specialized Weberian apparatus precludes their having given rise to any higher groups. Differences among various classifications of the superorder are not as great as they appear; the same or similar subgroups are widely recognized, but they may be assigned to different levels in the taxonomic hierarchy. L.S. Berg placed all ostariophysians in one order (Cypriniformes) with two divisions (Cyprini and Siluri). The many families of characins listed by Greenwood are recognized as subfamilies under the single family Characidae by many authorities. Present consensus favours the characoids as the most primitive suborder, an ancestral stock giving rise to cyprinoids in southeastern Asia and to gymnotoids in South America.

The siluriform fishes are more highly specialized than the cypriniforms, and the diplomystids undoubtedly are the most primitive of the catfishes. There is little agreement on the relationships of the other families, but recent research on the caudal skeleton indicates that the Ictaluridae, Bagridae, and Schilbeidae (among others) tend to retain primitive characters. Advanced features indicate a relationship between the Clariidae and Heteropneustidae; the Doradidae and Auchenipteridae; the Loricariidae, Astroblepidae, and Callichthyidae; and the Plotosidae and Chacidae. Extensive studies of morphological and other genetic characters of both cypriniform and siluriform fishes are needed before a satisfactory classification and phylogeny can be achieved.

BIBLIOGRAPHY. JAMES W. ATZ, *Dean Bibliography of Fishes 1968* (1971), the first volume of a comprehensive, computerized bibliographic series; GEORGE ALBERT BOULENGER, *Catalogue of the Fresh-Water Fishes of Africa*, 4 vol. (1909–16), an important, illustrated, systematic account of Old World tropical groups; PHILIP J. DARLINGTON, JR., *Zoogeography: The Geographical Distribution of Animals* (1957), an excellent account of the distribution of freshwater fishes; CARL H. EIGENMANN and GEORGE S. MYERS (coauthor of pt. 5), *The American Characidae*, 5 pt. (1917–29), a classic treatise, only one-third completed upon death of the author; M.M. ELLIS, "The Gymnotid Eels of Tropical America," *Mem. Carneg. Mus.*, 6:109–204 (1914), a comprehensive, systematic, and morphological study; WILLIAM K. GREGORY and G. MILES CONRAD, "The Phylogeny of the Characin Fishes," *Zoologica*, 23:319–360 (1938), an old, somewhat equivocal, but important contribution on classification; HARRY GRUNDFEST, "Electric Fishes," *Scient. Am.*, 203:115–124 (1960), a semipopular but authoritative article; WILLIAM T. INNES, *Exotic Aquarium Fishes*, 19th ed. (1956), a well-illustrated, informative handbook of popular aquarium fishes; JOHN G. LUNDBERG and JONATHAN N. BASKIN, "The Caudal Skeleton of the Catfishes, Order Siluriformes," *Am. Mus. Novit. 2398* (1969), a description of anatomy, evolution, and relationships; W. PFEIFFER, "Alarm Substances," *Experientia*, 19:113–123 (1963), an excellent review article; C.T. REGAN, "The Classification of the Teleostean Fishes of the Order Ostariophysi," *Ann. Mag. Nat. Hist.*, Series 8, 8:13–32, 553–577 (1911), an old, but historically useful treatise on morphology and classification; JOHN H. TODD, "The Chemical Languages of Fishes," *Scient. Am.*, 244:99–108 (1971), a description of contemporary experiments on communication; STANLEY H. WEITZMAN, "The Osteology of *Brycon meeki*, a Generalized Characid Fish, with an Osteological Definition of the Family," *Stanford Ichthyol. Bull.*, 8:1–77 (1962), an important review of characid classification and osteology.

(R.W.Y.)

Osteoglossomorpha

The superorder Osteoglossomorpha is a group of morphologically and biologically diverse primitive fishes primarily found in freshwaters; a few species enter slightly brackish water. Their relationship with other teleosts (*i.e.*, advanced bony fishes) is obscure; they probably were an early offshoot from the basal teleost stock. Osteoglossomorpha comprises six extant families and about 150 species. Although the group is of little importance to man, in parts of Africa, Asia, and South America certain osteoglossomorph species are sometimes sought commercially as food fishes.

Except for one North American family (Hiodontidae), the Osteoglossomorpha are tropical fishes. The families Mormyridae (elephant-snout fishes, mormyrs), Gymnarchidae, and Pantodontidae (butterfly fishes) are confined to Africa; the Notopteridae (featherbacks) occur in Africa, Southeast Asia, and India. The distribution of the Osteoglossidae (*e.g.*, pirarucu, arawana) in Africa, South America, and Australasia (believed by many authorities to have once been joined as a single landmass called Gondwana) is of particular zoogeographical interest.

The pirarucu (*Arapaima*) of the Amazon, one of the world's largest freshwater fishes, attains a length of three metres (about ten feet); other osteoglossomorphs—for example, certain mormyrids—are only a few centimetres long.

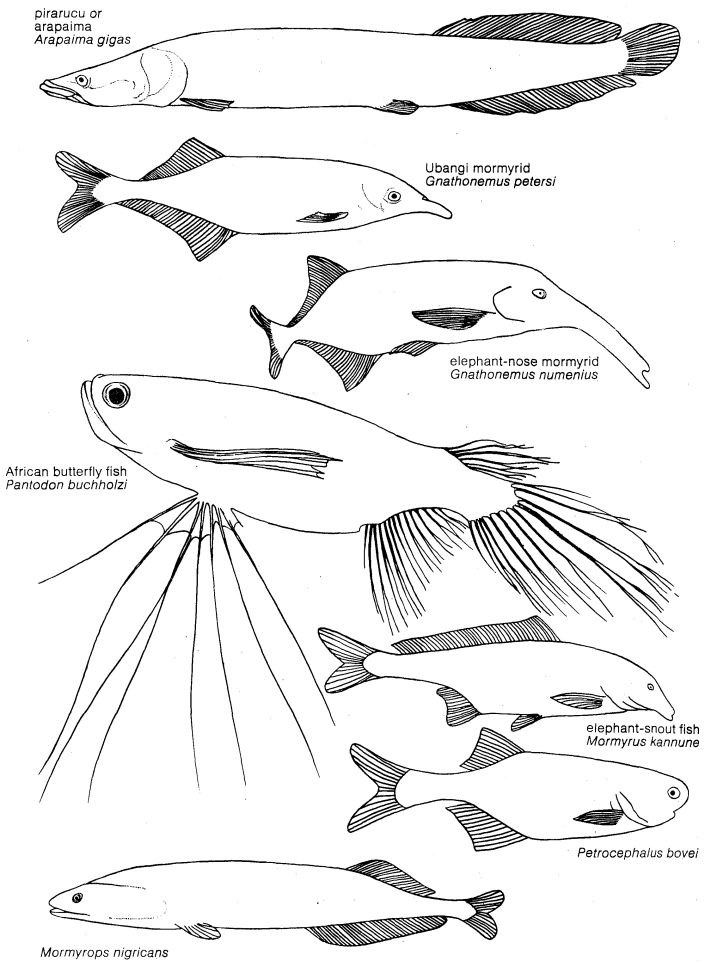
Natural history. *Life cycle and reproduction.* A variety of breeding habits have evolved among the Osteoglossomorpha. In some species there is considerable care of the young by the parents. Although no species is known to undertake extensive breeding migrations, many leave the usual habitat and move into floodplains or streams at breeding time.

The breeding biology of the Mormyridae has been little studied; it does not seem likely, however, that they prepare spawning nests or exercise much parental care. In contrast, *Gymnarchus niloticus* (Gymnarchidae) prepares a large floating nest from the matted stems of swamp grasses, biting off the stems and fashioning them into a trough-shaped structure with an internal length of about 50 centimetres (20 inches). Spawning takes place in the nest, and one or both parents guard the developing young for approximately 18 days.

Nests are not made by notopteroids, but they do establish a breeding territory. Both *Hiodon* species (goldeye and mooneye) spawn in the spring. Eggs are laid on gravel or rocks in shallow, quiet water, which the adults reach after a short migration. The young of the goldeye (*H. alosoides*) remain there until late summer before migrating downstream. In *Notopterus chitala* (Notopteridae), one parent, probably the male, clears an area of the bottom near some submerged object (*e.g.*, rock, plant stem, or piling) on which the eggs are later laid in circular bands; the male guards the developing embryos.

Care of the young

Members of the Osteoglossidae care for their young in a variety of ways. The South American *Osteoglossum bicirrhosum* and its Indo-Australian relatives *Scleropages leichardti* and *S. formosus* carry the eggs and young in the mouth of one parent; little else is known of their breeding habits. The African *Heterotis niloticus* prepares a crude nest from grasses in newly flooded swamp plains. The male guards the young and leads them from the nest on feeding excursions. Both sexes of *Arapaima gigas* of South America dig a spawning pit and guard the developing embryos, which hatch and leave the pit after about seven days; care is provided by the male, around whose



Representative osteoglossomorphic fishes.

Drawing by J. Helmer based on (*Gnathonemus petersi, Mormyrus kannune, Petrocephalus bovei, Mormyrops nigricans*) G. Sterba, *Freshwater Fishes of the World*, (*Gnathonemus numenius*) New York Zoological Society Photo, (*Pantodon buchholzi*) T. Dennett in *International Wildlife Encyclopaedia* (1969)

head the young congregate. The dark colour of the male's head seems to provide the chief stimulus for shoal orientation; *i.e.*, forming groups or schools. Apparently there is also a gustatory (*i.e.*, taste) or olfactory (*i.e.*, smell) stimulus in a secretion from glands on the male's head. Parent–young groups persist for two to three months; if the male dies, the young join other shoals.

Spawnings in aquaria suggest that *Pantodon* (Pantodontidae), the African butterfly fish, produces floating eggs and provides minimal parental care. A complicated courtship in this species has also been observed.

Electrical organs

Behaviour and ecology. Mormyridae and Gymnarchidae are of particular interest because they have electrical organs. Electrical discharges from these organs are in the form of pulses, the frequency and nature of which are different in each species. In nature, electrical discharges ranging from an output of about 120 to 300 pulses per second have been recorded.

The flow pattern of the electric field around the fish is distorted by the different conductivity of objects that pass into it. These variations are detected by modified nerve cells (mormyromasts) in the skin. The greatly varying conductivity differences among animals and inorganic objects make it possible for mormyrids to use their electrical organs to distinguish between prey, predators, and obstacles in the turbid water they often inhabit. Discharges from the organs also serve as signals to other mormyrids.

Some mormyriforms tend to swim with little body movement, using instead the dorsal fins for propulsion. This unusual swimming method is probably associated with the use of electric organs in navigation and detection; *Gymnarchus*, for example, swims with its body held straight, propulsion being provided by undulations of the dorsal, or back, fin. Since electrical organs lie near the tail, side-to-side movements of the tail end (as in normal swimming movement) would constantly change their position relative to that of the receptor organs, which are in the head area.

Rigid-bodied swimming like that of the mormyriforms also occurs in the featherbacks (Notopteridae), which use the long anal fin for propulsion. There is no electrical organ in the notopterids, however; the rigid body of these fishes may be correlated with the long, gas-filled swim bladder that extends into the tail end of the body.

All other osteoglossomorphs swim by using the body musculature and caudal (tail) fin in the usual manner. The little African butterfly fish (*Pantodon*) has greatly expanded winglike pectoral fins (behind the gills), which are used for short flights in the air, either to escape predators or to catch insects. *Pantodon* habitually swims or drifts just below the water surface. It leaps from the water by means of a powerful thrust of the pectorals, sending the fish 30 centimetres (about a foot) or more vertically out of the water. Short horizontal flights of about one metre (40 inches) are also executed.

Some species of Osteoglossidae (*Heterotis, Arapaima*), the related Pantodontidae, certain Notopteridae (*Notopterus, Papyrocranus, Xenomystus*), and the Gymnarchidae are able to breathe air at the surface; thus they can live in areas where the water is deoxygenated.

The swim bladder, noise-making, and hearing

The notopterid *Xenomystus* produces sounds that are used as warnings and in courtship. Swim-bladder structure in other Notopteridae suggests that they are also capable of emitting sounds. In all Notopteroidei (including the Hiodontidae) the swim bladder is closely connected with the inner ear, a condition that may be an aid to hearing. Except for the osteoglossid *Heterotis*, all the osteoglossomorphs are carnivorous, the smaller species (and young of all species) feeding on insects and other invertebrates, the larger species on fish. *Heterotis* feeds on microscopic plants and animals filtered from the water.

Osteoglossomorph fishes occupy a diversity of habitats in rivers and lakes, often in turbid waters or in regions with dense aquatic vegetation. A few species seem to require open waters, and some Notopteridae can tolerate slightly brackish water.

Form and function. All living Osteoglossomorpha have strongly toothed jaws. The jaws are not protrusible, but in piscivorous (*i.e.*, fish-eating) species the gape is sometimes considerable. The Mormyridae show a great variety of head shape and mouth form. Many insectivorous (*i.e.*, insect-eating) species have a small mouth at the tip of a long, curved, tubular snout; they feed by probing among rocks or in soft mud. Piscivores have larger mouths and short snouts.

The mouth is large in all Osteoglossidae, which, like the Notopteridae and Hiodontidae, have a more typical head shape than do the Mormyridae. *Heterotis* has complex spiralled structures lying above the gill arches on each side of the head and opening into the pharynx. Food particles, drawn into the pharynx with the respiratory current, are filtered out and concentrated in these organs.

In all Mormyridae and in the Gymnarchidae, a short length of body musculature toward the tail is modified to form the electric organ. These muscles have lost the ability to contract and have undergone considerable cellular reorganization, from slender fibres of ordinary muscle into thin, flat, electroplates, the structures in which electricity is produced. There are about 300 to 400 electroplates in mormyrid electric organs, and 600 to 800 in those of *Gymnarchus*. The brain, particularly the cerebellum, of mormyriforms is the largest known among fish; the cerebellum is associated with the electroreceptor organs (mormyromasts) in the skin.

A well-developed swim bladder is present in all Osteoglossomorpha. In the air-breathing osteoglossids, pantodontids, and gymnarchids, it is highly vascularized (*i.e.*, has many blood vessels), and its inner surface is honeycombed with tiny pits.

Paired extensions in the Notopteridae and Hiodontidae connect the swim bladder with the auditory region of the

skull. In the African genus *Papyrocranus,* diverticula (outpocketings) of the swim bladder actually penetrate the skull, a condition that probably improves hearing. Posteriorly the swim bladder in notopterids extends beyond the abdominal cavity and runs back on either side of the vertebral column. In the early embryos of Mormyridae and Gymnarchidae, a pair of thin tubes extends forward from the swim bladder into the skull; later the tubes atrophy, leaving an isolated vesicle—a blister, or balloon-like structure—within the skull surrounded by the semicircular canals of each ear.

The inner ear shows various modifications; in notopterids, gymnarchids, and mormyrids, the upper portion (for balance) is completely separated from the lower part (for hearing).

In the Mormyridae and Pantodontidae, the anal fin shows considerable sexual dimorphism in shape—*i.e.,* the anal fin of males differs from that of females. These differences may be related to spawning activity.

Classification. *Distinguishing taxonomic features.* Classification in the superorder Osteoglossomorpha is based largely on skeletal characters, in particular the caudal-fin skeleton; the bones around the eye; and the gill arches and their associated dentition. Details of the inner ear and swim-bladder anatomy are also of importance.

Annotated classification. The classification used here is based on that of P.H. Greenwood, D.E. Rosen, S.H. Weitzman, and G.S. Myers (1966), with subsequent research by Greenwood (1970) and G.J. Nelson (1968, 1969). Groups indicated by a dagger (†) are extinct and known only from fossils.

SUPERORDER OSTEOGLOSSOMORPHA

Primitive; well-developed teeth on tongue, skull base, and bones of the mouth cavity; caudal fin skeleton of characteristic form. Lower Cretaceous to Recent.

Order Osteoglossiformes

Osteoglossomorph fishes without electric organs.

Suborder Osteoglossoidei

Swim bladder not connected with skull; semicircular canals and lower part of inner ear connected.

Family Osteoglossidae. Fishes of diverse body form; pectoral fins not greatly enlarged, pelvic fins abdominal in position. Genera: *Arapaima* (1 species) and *Osteoglossum* (2 species), South America; *Scleropages* (2 species), Australia, New Guinea, Borneo, Sumatra, Malaysia, Thailand; *Heterotis* (1 species), Africa. Fossils from Eocene of North America and Tertiary of Australia (*Phareodus*); Eocene of Sumatra and Tertiary of India (*Musperia*).

†*Family Singididae.* Extinct. Apparently toothless; monotypic genus (*Singida*) from Paleocene of Tanzania (East Africa).

Family Pantodontidae. Greatly expanded winglike pectoral fins; pelvic fins thoracic. A monotypic genus, *Pantodon buchholzi,* from Africa. No fossil record.

Suborder Notopteroidei

Swim-bladder connected with the skull; semicircular canals separate from lower part of ear, or, if connected, utriculus greatly enlarged.

†*Family Lycopteridae.* Extinct. Lower Cretaceous of northeast Asia; small freshwater fishes resembling the Hiodontidae. 4 genera (6 species).

Family Notopteridae. Long anal fin confluent with reduced caudal; dorsal fin small or absent. Genera: *Papyrocranus* (1 species) and *Xenomystus* (1 species) in Africa, *Notopterus* (4 species) in Asia and Indonesia. Fossil *Notopterus* from Eocene of Sumatra.

Family Hiodontidae (goldeye and mooneye). Probably the most primitive living osteoglossomorphs. One genus, *Hiodon,* confined to North America. Fossil from Eocene of British Columbia (*Eohiodon rosei*).

Order Mormyriformes

With electrical organs; very diverse head shape and mouth form. Confined to Africa; fossils from Pliocene of Egypt; 2 families, about 130 species.

Family Mormyridae (mormyrs and elephant-snout fishes). Anal, caudal, and dorsal fins present; several genera; about 130 species.

Family Gymnarchidae. No anal fin; long dorsal confluent with reduced caudal fin. One genus and species, *Gymnarchus niloticus.*

Critical appraisal. Taxonomic problems of the osteoglossomorphs concern intragroup relationships, particularly whether the Mormyriformes are more closely related to the Osteoglossoidei or to the Notopteroidei. Some authorities relate the Mormyriformes with the Notopteroidei; others suggest closer affinity with the osteoglossoids and propose that one order, rather than two, should be recognized. The superorder as a whole presents problems of relationship with the other teleostean lineages. Certain fossil groups—*e.g.,* the Eocene genus *Brychaetus* and the Mesozoic families Plethodontidae and Ichthyodectidae—may be osteoglossomorphs, but their relationship with the living groups is still obscure. *Brychaetus* can probably be classified in the Osteoglossiformes, but it may represent a distinct suborder.

BIBLIOGRAPHY. P.H. GREENWOOD, "On the Genus *Lycoptera* and Its Relationship with the Family Hiodontidae (Pisces, Osteoglossomorpha)," *Bull. Br. Mus. Nat. Hist. (Zool.),* 19:259–285 (1970); and *et al.,* "Phyletic Studies of Teleostean Fishes, with a Provisional Classification of Living Forms," *Bull. Am. Mus. Nat. Hist.,* 131:339–455 (1966), includes the most recent classification of the Osteoglossomorpha and a discussion of the reasons for that arrangement; E.S. HERALD, *Living Fishes of the World* (1961), a popular, well-illustrated account of the various families, and some species, particularly their biology; K.H. LULING, "*Arapaima,* Giant Fish of Amazonas," *Animals,* 11:222–225 (1968), a popular account; G.J. NELSON, "Infraorbital Bones and Their Bearing on the Phylogeny and Geography of Osteoglossomorph Fishes," *Am. Mus. Novit.,* no. 2394 (1969); "Gill Arches of Teleostean Fishes of the Division Osteoglossomorpha," *J. Linn. Soc. (Zool.),* 47:261–277 (1968); J.R. NORMAN, *A History of Fishes,* 2nd ed. rev. by P.H. GREENWOOD (1963), includes a general account of osteoglossomorph biology, distribution, and anatomy (high school and college level); G. STERBA, *Süsswasserfische aus aller Welt* (1959; Eng. trans., *Freshwater Fishes of the World,* 1962), particularly for the aquarist.

(P.H.G.)

Osteopathic Medicine

Osteopathic medicine is a complete school of medical practice. It parallels the school of medicine, whose members hold the degree of bachelor or doctor of medicine, in utilizing all scientifically accepted methods of diagnosis and treatment, including the use of pharmaceuticals, surgery, and X-rays, but differs from them in its greater emphasis on the relationship between the musculoskeletal structure and organ function. Osteopathic physicians develop skill in recognizing and correcting structural problems through manipulative therapy and other forms of treatment. This use of manipulative therapy in diagnosis and treatment is an integral part of osteopathy.

Origin and development of osteopathy. Osteopathic medicine began in the United States in the 19th century as a reform movement against the then rather primitive armamentarium of drugs and surgical techniques. The founder, Andrew Taylor Still, developed a system of osteopathic medicine in reaction to the conditions and types of treatment that he observed while serving as an army physician during the American Civil War.

After failing to persuade various medical schools to incorporate his ideas into their teachings, Still established a new medical school in 1892 in Kirksville, Missouri, where he began conferring the doctor of osteopathy degree (D.O.) upon those physicians who adhered to his concepts of medicine. Still's emphasis on treating the whole man has remained an ideal of the profession.

Osteopathic medicine still has its main base in the United States. Canadian osteopathic physicians have received their training in the United States, and osteopathic medicine in the British Isles is in part a form of postgraduate specialization by holders of the M.D. degree. Most osteopathic physicians who practice elsewhere in the world have also been trained in the U.S. For these reasons the present article is focussed primarily on osteopathy as it is practiced in the United States.

Despite difficulties during its three quarters of a century

Present status

of existence, the osteopathic profession has achieved status as a recognized school of medicine. There are now 13,000 doctors of osteopathy in the United States, and others who received their education there are now located in other countries. The profession maintains nearly 300 hospitals, with a total of more than 20,000 beds.

General osteopathic hospitals provide health care of all kinds, and specialty-oriented hospitals include maternity centres, proctology clinics, arthritis centres, emergency clinics, and alcoholism and drug abuse centres. In the United States, osteopathic institutions are accredited by the American Osteopathic Association (AOA), and most are members of the American Osteopathic Hospital Association.

Training. Professional education leading to the degree doctor of osteopathy is similar to that for doctors of medicine. Although some students enter osteopathic colleges with a minimum of three years of professional training, the majority have completed degree programs.

Admission to colleges is dependent on scholastic achievement, aptitude as measured by the Medical College Admissions Test, and personal interviews. The four years of osteopathic medical education cover the basic sciences and clinical work. Throughout the curriculum special emphasis is placed on the importance of body mechanics and the relationship of body structure to function.

Upon graduation and completion of an internship in one of the AOA approved hospitals, osteopathic physicians enter either general practice—as do three out of four—or various specialties. Although the association particularly encourages general practice, doctors of osteopathy are also certified by AOA specialty boards in anesthesiology, dermatology, internal medicine, neurology, psychiatry, obstetrics and gynecology, ophthalmology, pathology, pediatrics, physical medicine and rehabilitation, proctology, radiology, and many surgical specialties. These men take two- to five-year residencies in osteopathic hospitals equipped to provide the additional training.

Doctors of osteopathy practicing in Canada receive their education at one of the colleges of osteopathic medicine in the United States. The Canadian Osteopathic Association, however, became a recognized national osteopathic association in 1970.

In Great Britain there are two schools of osteopathic medicine. Entrants to the British School of Osteopathy require only a secondary school education, and their four-year course is limited to manipulative therapy and diagnosis. Upon graduation, their practice is limited to these areas. The London College of Osteopathy limits its entrants to those who possess the M.D. degree, and the 14-month course is meant to be a postgraduate training that stresses the musculoskeletal system and manipulative therapy. Graduates are accepted into the British system of socialized medicine and have unlimited practice rights. Both groups may belong to the British Osteopathic Association. Neither British school is inspected, regulated, or approved by the AOA.

Legal status of osteopathy. Osteopathic physicians are licensed to practice medicine in all states of the United States and have the same professional rights and responsibilities as do holders of the M.D. degree under federal law and under the statutes of 48 states and the District of Columbia. In most cases, physicians of both schools of medical practice take their examinations before a medical licensing board composed of both doctors of medicine and doctors of osteopathy. In a very small number of instances doctors of osteopathy are examined by an all-medical board. Sixteen states have separate medical and osteopathic examining boards.

Licensing of osteopathic physicians

In two states—Louisiana and Mississippi—old laws limiting the scope of osteopathic practice continue to exist. Though there are fewer than 100 osteopathic physicians practicing in these states, legislation to guarantee full practice rights is under way in one of them.

In Canada, osteopaths' rights to practice are determined by the individual provinces. British regulations have been noted above.

In other countries, regulation of the practice of the healing arts varies greatly. While the practice of osteopathic medicine as such is not provided for by law outside the countries already discussed, osteopathic physicians practice in many other places.

Research and special programs. Osteopathic research, while encompassing a fairly wide range, considering its relatively limited financial resources, concentrates on topics that ordinarily would not be studied outside osteopathic auspices. Various phases of interrelationship between structure and function are under scrutiny. In the basic science category, studies include anatomy and function of nerve-muscle junctions, transmission of nervous impulses, somatic reflex functions, renal (kidney) growth and function, and blood flow dynamics. In the category of clinical sciences, there are studies of structural findings in hospitalized patients, the effects of manipulation under anesthesia for specific orthopedic problems, the effect of osteopathic manipulation on hypertension, and management of chronic obstructive lung disease by regular medical means, with and without osteopathic manipulation.

At the College of Osteopathic Medicine and Surgery in Des Moines, Iowa, a rehabilitation and treatment centre has been established on an in-patient basis for the alcoholic. At the Kirksville (Missouri) College of Osteopathy and Surgery, a program of rural clinics provides medical care to 40,000 patients by operating 11 clinics staffed by senior students. The clinics are in areas without resident physicians.

BIBLIOGRAPHY. AMERICAN OSTEOPATHIC ASSOCIATION, *Fact Sheet* (1970), statistical information on osteopathic physicians, education, research, and hospitals; J.M. HOAG *et al.* (eds.), *Osteopathic Medicine* (1969), a complete, modern textbook dealing with theoretical and applied concepts of osteopathic medicine; L.W. MILLS, *Opportunities in Osteopathic Medicine* (1964), a short book intended for prospective students of osteopathic medicine, examining the advantages and disadvantages of a career as a doctor of osteopathy, *The Osteopathic Profession and Its Colleges* (1970), a rather complete treatise on the osteopathic profession, and *Educational Supplement* (1970), a frequently revised appraisal of osteopathic colleges and their students; G.W. NORTHUP, *Osteopathic Medicine: An American Reformation* (1966), a historic overview of the development of osteopathic medicine from its formation to the present; A. STODDARD, *Manual of Osteopathic Practice* (1969), a modern textbook covering osteopathic concepts, treatment, and technique.

(E.P.C.)

Ottawa

Ottawa reflects in many ways the character of the Canadian nation, of which it has been the capital city since being so designated by Queen Victoria in 1857. Its metropolitan area straddles the provincial border of English-speaking Ontario and French-speaking Quebec, and in native language the more than 600,000 inhabitants are divided almost equally. The city itself is entirely in Ontario. In 1884 Sir Wilfrid Laurier, the first French-Canadian premier of Canada, referred justifiably to Ottawa as a city about which it was difficult to say anything good and about which little in the way of visual delight could be predicted. His forebodings, however, held true only until 1937, when the French architect Jacques Gréber began the refashioning of Ottawa into one of the most beautiful of Canadian cities.

The Ottawa area, a harmonious blend of forest, farmlands, and water, is located in a lowland that is hot and humid in summer, nearly polar in winter. It is almost completely walled in by rocks of the so-called Canadian Shield dating from the Precambrian Era over 570,000,000 years ago. Among the deposits left by the great ocean arm known as the Champlain Sea are the deep sand beds along the Rideau River, which flows through the city from the south, and whale skeletons. Across the Ottawa River to the north, in Quebec, lie the heavily wooded Laurentian Hills and the valley of the Gatineau River, which, like the Rideau, joins the Ottawa by the city (for details on related subjects, see CANADA; ONTARIO).

History. The first descriptions of Ottawa's future site were written by the French explorer and founder of New France, Samuel de Champlain, in 1613. The three rivers

Centre block of the Parliament buildings on Parliament Hill, Ottawa. The Peace Tower is at centre, with the Commons wing at left and the Senate wing at right. The spire of the Library of Parliament is in the background.
By courtesy of Information Canada Phototheque

served as passageways for explorers and traders over the following two centuries. Early in the 19th century the Napoleonic Wars increased Britain's need for shipbuilding timber, and the Ottawa Valley offered just such resources. In 1800 an American, Philemon Wright, had begun timbering across the Ottawa River in what became the city of Hull. With the War of 1812 between Britain and the U.S., the Rideau provided the route for a safe canal way from the Ottawa River to Kingston, on Lake Ontario. The settlement of Ottawa was under way. It was hastened by the arrival in 1826 of Lt. Col. John By of the Royal Engineers to work on the canal, and the town became Bytown. The first white child was born the same year, and in 1827 a Methodist meetinghouse became the first church, a school opened, and a team of doctors opened a practice. A primitive Rideau Street stirred with life, later to become the main commercial artery of the Lower Town, the downtown area.

Canada's compromise capital

Ottawa might still be a modest city had not political quarrels between Quebec city and Toronto and between Montreal and Kingston induced leaders to call upon Queen Victoria to designate a capital for United Canada. From a population of only a little more than 7,700 in 1851, Bytown doubled in the following decade and found itself, in 1855, incorporated and rechristened Ottawa after an Indian tribe, the Outaouacs or Outaouais, that had moved down from the Lake Huron region. It became the fastest growing metropolis in eastern Canada, a development due largely to the presence of the national government. In 1937 Prime Minister William L. Mackenzie King brought Gréber from France to begin the redevelopment of the national capital district.

The contemporary city. The more than 1,800 square miles of the Ottawa metropolitan area, on both sides of the Ottawa River and the provinicial border, comprises about 70 municipalities. Nearly three-quarters of the population lives on the Ontario side. Ottawa itself completely encircles the separate village of Rockcliffe Park, known as the wealthiest in Canada, and the city of Vanier, formerly Eastview. With the linguistic split in the region, bilingualism flourishes in Ottawa more than in any other Canadian city.

The federal buildings in Ottawa are dominated by the 291-foot-high Peace Tower, with its 53-bell carillon, atop the Parliament building, which was rebuilt between 1916 and 1927 after a disastrous fire in the former year. There are few apartment buildings in the city, the residents having a strong preference for single-family units. The many private and public lawns and flowerbeds throughout its environs have made Ottawa something of a garden city from spring until late autumn.

Demography. It is often said, as it is of many cities whose growth has been around government, that no one is born in Ottawa. Though the statement is obviously an exaggeration, Ottawa is populated by Canadiáns of many ethnic origins from across the nation. There is really no melting pot, for the predominant French- and English-speaking populations tend to retain their traditions and religion as well as mother tongue. The harmony of their coexistence has been something of a trend-setting example for Canada since the adoption of an official national policy of bilingualism in 1969.

Linguistic and cultural coexistence

Forecasters have projected a population of more than 1,200,000 for the region by the year 2001, all but about 60,000 living in urban areas. To prevent a monstrous, unhealthy city from developing, a policy known as "open-space therapy" has brought about the creation of a greenbelt around the city. The 41,000 acres of the belt extend in a semicircle for more than 24 miles.

Economic life. The fur trade and lumbering of its early years have faded from Ottawa's life, and the city's industry employs only a small fraction of the labour force. The largest industry is a fine-paper company with about 700 workers. The federal government employs more than 75,000 workers in nearly 150 buildings. As the national capital, Ottawa also attracts many commercial and financial associations from around the country as well as embassies, trade associations, and the like.

Political institutions. The city proper is self-governing and forms part of a vast regional community, the Ottawa-Carleton communities, but the city's attentions are focussed on the federal government. Along with the governmental institutions and crown corporations, the diplomatic missions also contribute to its way of life and give Ottawa a multilingual, cultural, and cosmopolitan atmosphere.

Services. Ottawa is served by both of Canada's major railroads and several Canadian and foreign airlines. The Ottawa Transportation Company supplies bus service throughout the city, but navigation on the Ottawa and Rideau rivers, except for pleasure craft, is a thing of the past. Inexpensive hydroelectric power is purchased in bulk by Ottawa-Hydro from Ontario-Hydro.

Two English-language newspapers, the *Citizen* and the *Journal*, plus *Le Droit* in French, serve the metropolitan area. Two of the six radio stations and one of the three television stations are French. There are seven hospitals in the city.

Cultural life and recreation. The open-space therapy has focussed recreation in Ottawa around parks, playgrounds, pools, and beaches. Pollution has made the beaches hazardous, but it is believed that purification plants along the Ottawa and Rideau rivers will cure this by the mid-1970s.

The presence of large numbers of people from other provinces and nations gives Ottawa a cosmopolitan atmosphere. The inauguration in 1969 of the National Arts Centre has brought cultural life, long hampered by lack of facilities, to a level with that of other North American cities. The centre includes a large opera house and two smaller theatres as well as a salon and foyer for huge festivities or balls. The major cultural centres remain the city's three universities. The University of Ottawa and St. Paul University are bilingual institutions, whereas Carleton University is entirely English. A large community college, Algonquin, provides technical training.

New art centres, museums, and galleries

The National Library and Public Archives Building and the National Museum of Science and Technology, both inaugurated during the 1960s, are bright new features of Ottawa's life. Use of the Parliamentary Library is restricted to legislators and governmental officials. The Public Library moved into a high-rise office complex in 1973. The superb collections of the National Gallery of Canada date back to the 13th century, although its emphasis is on the Canadian arts.

BIBLIOGRAPHY. WILFRID EGGLESTON, *The Queen's Choice* (1961), is a history of Canada's capital in which the author describes the development of the city and relates the history of the National Capital Commission. More detailed information pertaining to Ottawa may be found in LUCIEN

BRAULT, *Ottawa, Old and New* (1946), although now somewhat dated. Ottawa's history is so closely linked with the history of the Ottawa Valley that the only way to really know it is to follow the historical guide provided by COURTNEY C.J. BOND in *The Ottawa Country* (1968), the companion volume of *City on the Ottawa*, rev. ed. (1967). An interesting feature of the latter book is that it introduces the reader, briefly, to the Canadian authors, poets, and artists who lived in Ottawa.

(Ma.G.)

Otto I the Great, Emperor

Otto I, German king and Holy Roman emperor, consolidated the German *Reich*—the old kingdom of the East Franks—by his suppression of the rebellious nobles and his decisive defeat of the invading Magyars, leaving a solid and durable government to two successors.

By courtesy of the Metropolitan Museum of Art, New York, gift of George Blumenthal, 1941

Otto I (left) offering a model of Magdeburg cathedral to Christ in majesty, ivory altarpiece, *c.* 970. In the Metropolitan Museum of Art, New York.

Otto was born on November 23, 912, the son of the future king Henry I, of the Liudolfing, or Saxon, dynasty, and his second wife, Matilda. Little is known of his early years, but he probably shared in some of his father's campaigns. He married Edith, daughter of the English king Edward the Elder, in 930; she obtained as her dowry the flourishing town of Magdeburg. Nominated by Henry as his successor, Otto was elected king by the German dukes at Aachen on August 7, 936, a month after Henry's death, and crowned by the archbishops of Mainz and Cologne.

While Henry I had controlled his vassal dukes only with difficulty, the new king firmly asserted his suzerainty over them. This led immediately to war, especially with Eberhard of Franconia and his namesake, Eberhard of Bavaria, who were joined by discontented Saxon nobles under the leadership of Otto's half brother Thankmar. Thankmar was defeated and killed, the Franconian Eberhard submitted to the King, and Eberhard of Bavaria was deposed and outlawed. In 939, however, Otto's younger brother Henry revolted; he was joined by Eberhard of Franconia and by Giselbert of Lotharingia and supported by the French king Louis IV. Otto was again victorious: Eberhard fell in battle, Giselbert was drowned in flight, and Henry submitted to his brother. Nevertheless, in 941 Henry joined a conspiracy to murder the King. This was discovered in time; and, whereas the other conspirators were punished, Henry was again forgiven. Thenceforward he remained faithful to his brother and, in 947, was given the dukedom of Bavaria. The other German dukedoms were likewise bestowed on relatives of Otto.

Suppression of the nobles

Despite these internal difficulties, Otto found time to strengthen and to extend the frontiers of the kingdom. In the east the margraves Gero and Hermann Billung were successful against the Slavs, and their gains were consolidated by the founding of the Monastery of St. Maurice in Magdeburg, in 937, and of two bishoprics, in 948. In the north, three bishoprics (followed in 968 by a fourth) were founded to extend the Christian mission in Denmark. Otto's first campaign in Bohemia was, however, a failure, and it was not until 950 that the Bohemian prince Boleslav I was forced to submit and to pay tribute.

Having thus strengthened his own position, Otto could not only resist France's claims to Lorraine (Lotharingia) but also act as mediator in France's internal troubles. Similarly, he extended his influence into Burgundy. Moreover, when the Burgundian princess Adelaide, the widowed queen of Italy whom the margrave Berengar of Ivrea had taken prisoner, appealed to him for help, Otto marched into Italy in 951, assumed the title of king of the Lombards, and married Adelaide himself, his first wife having died in 946. In 952 Berengar did homage to him as his vassal for the kingdom of Italy.

Otto had to break off his first Italian campaign because of a revolt in Germany, where Liudolf, his son by Edith, had risen against him with the aid of several magnates. Otto found himself compelled to withdraw to Saxony; but the position of the rebels began to deteriorate when the Magyars invaded Germany in 954, for the rebels could now be accused of complicity with the enemies of the *Reich*. After prolonged fighting, Liudolf had to submit in 955. This made it possible for Otto to defeat the Magyars decisively in the Battle of the Lechfeld, near Augsburg, in August 955; they never invaded Germany again. In the same year Otto and the margrave Gero also won a victory over the Slavs. A further series of campaigns led, by 960, to the subjection of the Slavs between the middle Elbe and the middle Oder. The archbishopric of Magdeburg was founded in 968 with three suffragan bishoprics. Even Mieszko of Poland paid tribute to the German king.

Foreign conquests

In May 961 Otto procured the election and coronation of the six-year-old Otto II, his elder son by Adelaide, as German king. Then he went for a second time to Italy on the appeal of Pope John XII, who was hard pressed by Berengar of Ivrea. Arriving in Rome on February 2, 962, Otto was crowned emperor, and 11 days later a treaty, known as the *Privilegium Ottonianum,* was concluded, to regulate relations between emperor and pope. This confirmed and extended the temporal power of the papacy; but it is a matter of controversy whether the proviso enabling the emperor to ratify papal elections was included in the original version of the treaty or added in December 963, when Otto deposed John XII for treating with Berengar and set up Leo VIII as pope. Berengar was captured and taken to Germany, and in 964 a revolt of the Romans against Leo VIII was suppressed.

When Leo VIII died in 965, the Emperor chose John XIII for pope, but John was expelled by the Romans. Otto, therefore, marched for a third time to Italy, where he stayed from 966 to 972. He subdued Rome and even advanced into the Byzantine south of Italy. Prolonged negotiations with Byzantium resulted in the marriage of Otto II to the Byzantine princess Theophano, in 972. Having returned to Germany, the Emperor held a great assembly of his court at Quedlinburg on March 23, 973. He died in Memleben on May 7, 973, and was buried in Magdeburg at the side of his first wife.

Otto I's achievement rests mainly on his consolidation of the *Reich*. He deliberately made use of the bishops to strengthen his rule and thus created that "Ottonian church system of the *Reich*" that was to provide a stable and long-lasting framework for Germany. By his victorious campaigns, he gave Germany peace and security from foreign attack; and the pre-eminent position that he won as ruler gave him a sort of hegemony in Europe. His Italian policy and the acquisition of the imperial crown constituted a link with the old Carolingian tradition and was to prove a great responsibility for the German people in the future. All areas under Otto's rule prospered, and the resultant flowering of culture has been called the Ottonian renaissance.

Assessment

BIBLIOGRAPHY. RUDOLF KOEPKE and ERNST DUEMMLER, *Jahrbücher der deutschen Geschichte: Kaiser Otto der Grosse* (1876; 2nd ed., 1962), the yearbook of the German government that offers a description of the actual events in chronological order and the people involved, using all critically han-

dled sources; GERD TELLENBACH, "Otto der Grosse, 912-973," in *Die grossen Deutschen*, new ed., vol. 1, pp. 35–51 (1956), a short character sketch in essay form; HERMANN AUBIN, *Otto der Grosse und die Erneuerung des abendländischen Kaisertums im Jahre 962* (1962); and HELMUT BEUMANN and HEINRICH BUETTNER, *Das Kaisertum Ottos des Grossen* (1963)—all the above-mentioned works assess the historical significance of the renewal of the empire in reaction to the idea of the Millennium; HAGEN KELLER, "Das Kaisertum Ottos des Grossen im Verständnis seiner Zeit," *Deutsches Archiv für Erforschung des Mittelalters*, 20:325–388 (1964), a discussion of the thesis that Otto was crowned emperor by the Pope versus the idea that he was already emperor, although only with the title of king. Very little has been published on Otto I in English. JOHN J. GALLAGHER, *Church and State in Germany Under Otto the Great* (1938); and chapters by C.W. PREVITE-ORTON and A.L. POOLE in *The Cambridge Medieval History*, vol. 3 (1964), are of interest.

(Ku.R.)

Otto, Rudolf

German theologian, philosopher, and historian of religions, Rudolf Otto exerted worldwide influence upon religious thought during the early decades of the 20th century following the publication of his pioneering work, *Das Heilige* (1917; Eng. trans., *The Idea of the Holy*, 1923). In this book Otto undertook to characterize the experience of apprehending the Holy—the sense of a sacred, ultimate power and reality—as disclosed in all religions, and thus to identify what is common to all of them as a distinctly religious phenomenon (see also SACRED OR HOLY). Within a decade after its publication, the book became world-famous and Otto stood forth among Protestant theologians as potentially the most fruitful religious thinker of the postwar period. With the more recent upsurge of interest in the phenomenological study of religion (the systematic study of religious experience and its expressions), his work is receiving attention, especially among younger historians of religion.

Foto-Jannasch, Marburg/L.

Rudolf Otto, 1925.

Academic and public career. Otto was born on September 25, 1869, in Peine (Hanover, now Niedersachsen), Germany, the son of William Otto, a manufacturer, and Karoline Reupke Otto. Little is known of Otto's early life, except that he was educated at the gymnasium in Hildesheim before becoming a student of theology and philosophy at the University of Erlangen and, later, at the University of Göttingen, where he was made a *Privatdozent* ("lecturer") in 1897, teaching theology, history of religions, and history of philosophy. In 1904 he was appointed professor of systematic theology at Göttingen, a post he held until 1914, when he became professor of theology at the University of Breslau. In 1917 he became professor of systematic theology at the University of Marburg and for one year (1926–27) served as rector of the university. He retired from his university post in 1929, though he continued to live in Marburg the rest of his life.

Otto took time from his scholarly pursuits, more out of a sense of duty than of preference, to participate in community and public affairs. He was a member of the Prussian Parliament from 1913 to 1918 and a member of the Constituent Chamber in 1918, where he asserted a liberal and progressive influence. And he was later to concern himself with the political questions of the Weimar Republic. Otto also participated widely in Christian ecumenical activities, both as they related to divisions within the Christian community and as they concerned relations between Christianity and other religions of the world.

Scholarly pursuits. *Early development of Otto's inquiry.* What initially prompted Otto's inquiry into man's experience of the Holy was a specifically Christian, even Protestant, concern that had awakened in him while studying the life and thought of Martin Luther. This concern—to elucidate the distinctive character of the religious interpretation of the world—is reflected in his first book, *Die Anschauung vom heiligen Geiste bei Luther* (1898), literally, "The Perception of the Holy Spirit by Luther." He was to expand his inquiry in his book, *Naturalistische und religiöse Weltansicht* (1904; Eng. trans., *Naturalism and Religion*, 1907), in which he contrasted the naturalistic and the religious ways of interpreting the world, indicating first their antitheses and raising the question of whether the contradictions can be or should be reconciled. Otto resisted an easy reconciliation between the world view offered by the sciences and the religious interpretation but opposed equally the religionist's hostility toward science and the scientist's disregard of religion. The two perspectives, he insisted, are to be embraced and heeded for what they purport to disclose concerning the world in which men live; but it was clear that Otto's principal concern was to justify and to clarify what it is that the religious interpretation of the world, even within its rational aspect, conveys to man as a distinctive dimension of understanding beyond the discoveries of the sciences and the generalized knowledge following from them. Five years later came his work, *Kantische-Fries'sche Religionsphilosophie* (1909; Eng. trans., *The Philosophy of Religion Based on Kant and Fries*, 1931), a discussion of the religious thought of the German philosophers Immanuel Kant and Jacob Friedrich Fries, in which he sought to specify the kind of rationality that is appropriate to religious inquiry.

Distinctive character of religious understanding

During 1911–12 Otto undertook an extended journey, visiting many countries of the world, beginning with North Africa, Egypt, and Palestine, continuing to India, China, and Japan, and returning by way of the United States. These experiences were to set his problem in a worldwide context, turning him to an extended and searching exploration of the diverse ways in which the religious response had manifested itself among various religions of the world. He proved to be remarkably well equipped for such an exploration, both in his mastery of languages and his knowledge of the history of world religions. In addition to being at home with the languages of Near Eastern religions, he had mastered Sanskrit sufficiently to translate many ancient Hindu texts into German as well as to write several volumes comparing Indian and Christian religious thought.

From Schleiermacher to "The Idea of the Holy." Otto's initial mentor guiding his inquiry into the specific character of the religious response was the eminent German philosopher and theologian Friedrich Schleiermacher. It was Schleiermacher's early work, specifically his book *Über die Religion. Reden an die Gebilden unter ihren Verächtern* (1799; Eng. trans., *On Religion: Speeches to its Cultured Despisers*, 1893), to which Otto gave particular attention. What appealed to him in this work was Schleiermacher's fresh way of perceiving religion as a unique feeling or awareness, distinct from ethical and rational modes of perception, though not exclusive of them. Schleiermacher was later to speak of this unique feeling

"Absolute dependence" and "creature-feeling"

as man's "feeling of absolute dependence." Otto was deeply impressed by this formulation and credited Schleiermacher with having rediscovered the sense of the Holy in the post-Enlightenment age. Yet he later criticized the formulation on the grounds that what Schleiermacher had pointed up here was no more than a close analogy with ordinary, or "natural," feelings of dependence. For "absolute dependence" Otto substituted "creature-feeling." Creature-feeling, he said,

> is itself a first subjective concomitant and effect of another feeling element, which casts it like a shadow, but which in itself indubitably has immediate and primary reference to an object outside of the self.

Otto called this object "the numinous" or "Wholly Other"—*i.e.*, that which utterly transcends the mundane sphere, roughly equivalent to "supernatural" and "transcendent" in traditional usage.

Various influences had played upon Otto's reflections through the years, aiding him in reformulating the religious category that was to carry him beyond Schleiermacher. His early teacher at Göttingen, Albrecht Ritschl, had located religion in the realm of value judgments, whereas, more significantly, his theological colleague at Göttingen, Ernst Troeltsch, sought for a religious a priori as the ground of religious interpretation and judgment. Otto was impressed by William James's shrewd insights in *The Varieties of Religious Experience* (1902), yet he found James's empirical method inadequate for interpreting such phenomena. Otto was particularly attracted to the thought of J.F. Fries, already mentioned, whose notion of *Ahndung* (obsolete form of *Ahnung*; literally, "presentiment," or "intuition"), a yearning that yields the feeling of truth, opened up to him a way of dealing with religious phenomena, sensitively and appropriately. These "feelings of truth" Otto sought to schematize in his *Idea of the Holy*.

The "numinous" and *mysterium tremendum*

In that work, however, Otto was conscious of moving beyond his previous efforts, exploring more specifically the nonrational aspect of the religious dimension, for which he coined the term "numinous," from the Latin *numen* ("god," "spirit," or "divine"), on the analogy of "ominous" from "omen." The numinous, the awe-inspiring element of religious experience, Otto contended,

> evades precise formulation in words. Like the beauty of a musical composition, it is non-rational and eludes complete conceptual analysis; hence it must be discussed in symbolic terms.

Thus, *The Idea of the Holy*, while benefitting from earlier studies, represented for Otto a new venture and a radical shift in the nature and ground of his inquiry. The concern here was to attend to that elemental experience of apprehending the numinous itself. In such moments of apprehension, said Otto,

> we are dealing with something for which there is only one appropriate expression, *mysterium tremendum*. . . . The feeling of it may at times come sweeping like a gentle tide pervading the mind with a tranquil mood of deepest worship. It may pass over into a more set and lasting attitude of the soul, continuing, as it were, thrillingly vibrant and resonant, until at last it dies away and the soul resumes its "profane," non-religious mood of everyday experience. . . . It has its crude, barbaric antecedents and early manifestations, and again it may be developed into something beautiful and pure and glorious. It may become the hushed, trembling, and speechless humility of the creature in the presence of—whom or what? In the presence of that which is a *Mystery* inexpressible and above all creatures.

Although the *mysterium*, which Otto represents as the form of the numinous experience, is beyond conception, what is meant by the term, he insists, is something intensely positive. This can be experienced in feelings that convey the qualitative content of the numinous experience. This content presents itself under two aspects: (1) that of "daunting awfulness and majesty," and (2) "as something uniquely attractive and *fascinating*." From the former comes the sense of the uncanny, of divine wrath and judgment; from the latter, the reassuring and heightening experiences of grace and divine love. This dual impact of awesome mystery and fascination was Otto's characteristic way of expressing man's encounter with the Holy.

After "The Idea of the Holy." Otto employed the method he had developed in *The Idea of the Holy* in three major publications that followed: *West-Östliche Mystik* (1926; Eng. trans., *Mysticism East and West*, 1932); *Die Gnadenreligion Indiens und das Christentum* (1930; Eng. trans., *India's Religion of Grace and Christianity*, 1930); and *Reich Gottes und Menschensohn* (1934; Eng. trans., *The Kingdom of God and Son of Man*, 1938). Of the three books, the latter is especially important for glimpses of new insight that seem to point beyond the earlier, more widely acclaimed volume; it renders the hint of ultimacy that appears in present history.

Otto's concern with experiencing the numinous also gave rise to experimenting with new forms of liturgy designed to give urgency and vividness to such experiences in Protestant services of worship under critically controlled conditions. Here he employed a "Sacrament of Silence" as a culminating phase, a time of waiting comparable to the Quaker moment of silence, which he acknowledged to have been the stimulus to his own innovation.

Interest in non-Christian religions

Otto took all religions seriously as occasions to experience the Holy and thus pressed beyond involvement in his own historic faith as a Christian to engage in frequent encounter with people of other religious traditions. He had much respect for the distinctive characteristics of the various religions and thus resisted universalizing religion in the sense of reducing all to the lowest common denominator. Yet he strongly argued for a lively exchange between representatives of the various religions. It was this concern that led him to create in Marburg the Religious Collection of religious symbols, rituals, and apparatus on a worldwide basis for purposes of inspection and study and to advocate establishing an Inter-Religious League as "a cultural exchange in which the noblest . . . of our art and science and of our whole spiritual heritage would be mutually interpreted and shared."

Personal life and characteristics. Never having married, Otto shared his home with his widowed sister and her daughter. A tall, erect man of subtle mien and movement, Otto created an imposing presence. His students at Marburg called him "Der Heilige" ("the Holy One" or "the Saint"). Someone said of him after his death, "After all, he was something of a king; yet a king who did not lack humility."

Rudolf Otto was first and foremost a scholar, absorbed in his special field of inquiry and rigorous in the pursuit of it. This in itself presented him on initial contact as being austere in manner and remote in his reflections. People who encountered him did not readily intrude upon the man or engage him in light conversation. To interrupt his reflections, they felt, one must have something of significance to talk about. That this was in part a facade of their own making, based upon the impressiveness of the man himself, was discovered again and again by those who came to know him well; for, despite this outward austerity, Otto was a gentle person, reticent in his response to people. Yet he could be gracious and outgoing once the barrier of formality had relaxed. Nevertheless, the style of the man was that of a serious and dedicated scholar, and no amount of congeniality or friendliness dispelled that tone of an encounter with him.

Otto died at Marburg March 6, 1937.

BIBLIOGRAPHY. ROBERT F. DAVIDSON, *Rudolf Otto's Interpretation of Religion* (1947), the only full-length biographical study of Otto in English—principally an analysis of his thought; S.P. DUBEY, *Rudolf Otto and Hinduism* (1969), the only book-length study of Otto's philosophy in English, other than Davidson's book; BERNARD E. MELAND, "Rudolf Otto," in DEAN G. PEERMAN and MARTIN E. MARTY (eds.), *A Handbook of Christian Theologians*, pp. 169–191 (1965), a characterization of the man and his thought; JOACHIM WACH, "Rudolf Otto and the Idea of the Holy," in *Types of Religious Experience: Christian and Non-Christian*, pp. 209–227 (1951), a delightful essay by one who knew Otto well (highly informative and revealing); WILLIAM J. WAINRIGHT, "Rudolf Otto," in *The Encyclopedia of Philosophy*, vol. 6, pp. 13–15 (1967), especially helpful as an interpretation and assessment of the philosophical aspects of Otto's thought.

(B.E.M.)

Ottoman Empire and Turkey, History of the

A term without ethnic significance, Ottoman is a dynastic application derived from the Arabic "'Uthmān," for Osman (died 1324), who is regarded as the founder of the Turkish empire that spanned six centuries and came to an end only in 1922, when Turkey was proclaimed a republic. This empire, centred in Anatolia, varied greatly in extent during its checkered history. At various times it included the Balkan States, Greece, Crete, and Cyprus; parts of Hungary, Austria, and southern Russia; Iraq, Palestine, and Egypt; North Africa as far west as Algeria; and parts of Arabia. For the previous history of the Turkish or Turkmen tribes that Osman led, see SELJUQS; and for the history of the area of modern Turkey before the Ottomans, see ANATOLIA, ANCIENT.

This article is divided into the following sections:

I. The period of growth and world power

THE OTTOMAN STATE TO 1481: THE AGE OF EXPANSION

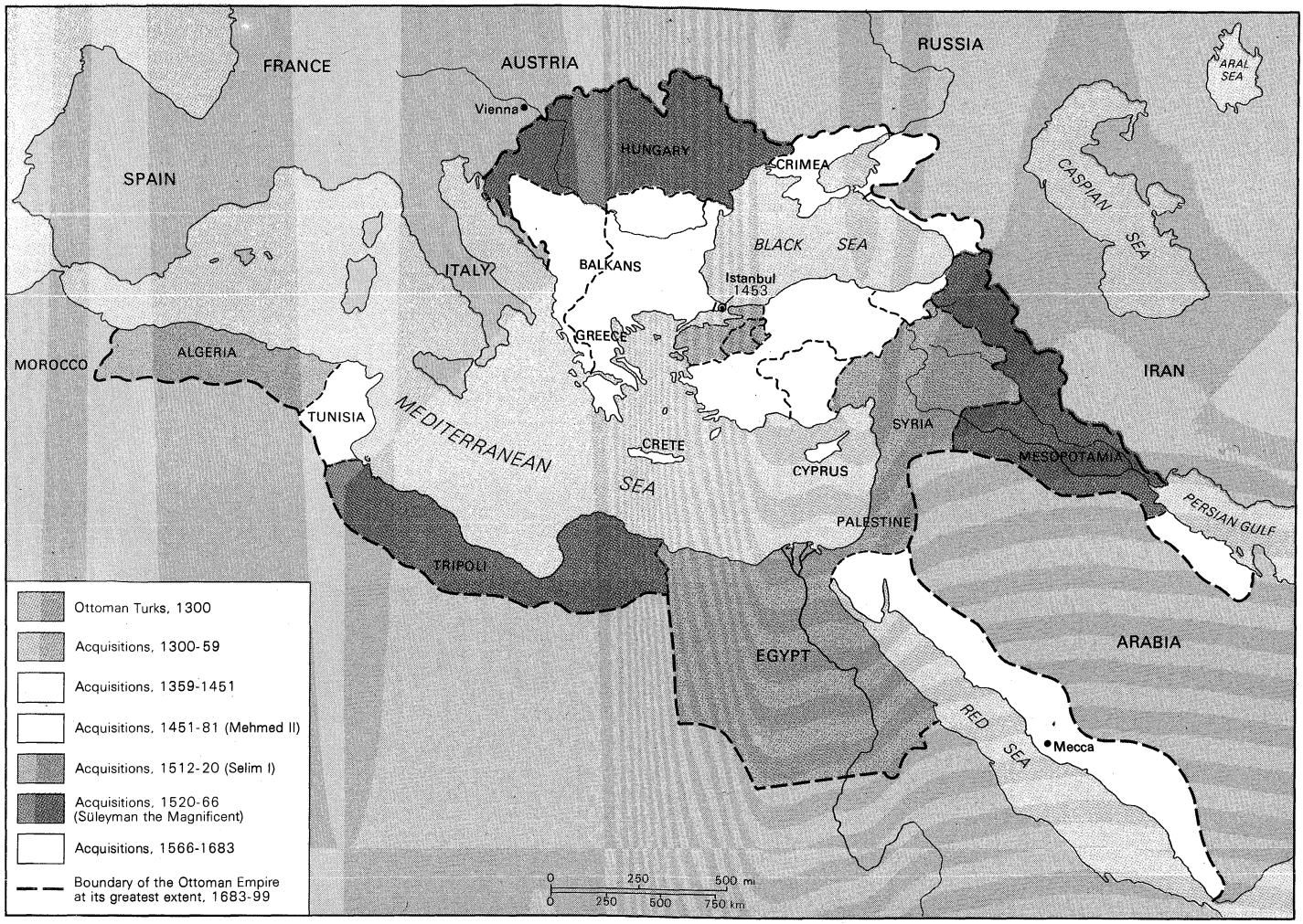
The first period of Ottoman history was characterized by almost continuous territorial expansion, during which the Ottoman dominion spread out from a small northwestern Anatolian principality to cover an empire encompassing southeastern Europe, Anatolia, and the Arab world. At the same time, the political, economic, and social institutions of the Middle East were amalgamated with those inherited from Byzantium and the great Turkish empires of Central Asia and re-established in new forms that were to be characteristic of the area into modern times.

Origins and expansion of the Ottoman state, c. 1300–1402. In their initial stages of expansion, the Ottomans were leaders of the Turkish *gazis*, or fighters for the faith of Islām, against the shrinking Christian Byzantine state. The ancestors of Osman, the founder of the dynasty, were members of the Kayı tribe who had entered Anatolia along with a mass of Turkmen Oğuz nomads who overwhelmed Byzantium after the Battle of Manzikert in 1071 and came to occupy eastern and central Anatolia during the 12th century. The *gazis* fought against the Byzantines and then the Mongols, who invaded Anatolia following the establishment of the Īl-Khānid (Ilhanid) Empire in Iran and Mesopotamia starting in the last half of the 13th century. Following the Mongol defeat of the Seljuq army in 1293, Osman I emerged as prince (*amīr*) of the border principality of Bithynia, in northwestern Anatolia, and was in command of the *gazis* leading the fight against the Byzantines in that area. With the disintegration of Seljuq power and its replacement by Mongol suzerainty, enforced by direct military occupation of much of eastern Anatolia, independent Turkmen principalities emerged in the remainder of Anatolia, among which was that led by Osman. Hemmed in on the east by the more powerful Turkmen principality of Germiyan, Osman and his immediate successors concentrated their attacks on Byzantine territories bordering the Bosporus and the Sea of Marmara to the west. Left as the most important Muslim rivals of Byzantium, the Ottomans became the main focus for the mass of nomads and urban unemployed then searching for means to gain their livelihoods and to fulfill their religious desire to expand the territory of Islām. The Ottomans were able to take advantage of the opportunity offered by the decay of the Byzantine frontier defense system and the rise of economic, religious, and social discontent in the Byzantine Empire (*q.v.*) and, starting under Osman I and continuing under his successors Orhan (Orkhān; ruled 1324–60) and Murad I (ruled 1360–89), took over Byzantine territories, first in western Anatolia and then in southeastern Europe. It was only under Bayezid I (ruled 1389–1402) that the wealth and power gained by this initial expansion was used to assimilate the Anatolian Turkish principalities to the east.

By 1300 Osman ruled an area stretching from Eskişehir (Dorylaium) to the plains of İznik (Nicaea), having defeated several organized Byzantine efforts to curb his expansion. Byzantine attempts to secure Īl-Khānid support against the Ottomans from the east were unsuccessful, and the emperor's use of mercenary troops from western Europe caused more damage to his own territory than to that of the Turks. At the time, however, the Ottomans lacked effective siege equipment and were unable to take the major cities of Bithynia. Nor could they move against their increasingly powerful Turkmen neighbours at Aydın and Karası, in southwestern Anatolia, who had arisen as the result of Byzantine weakness in the area. Orhan's capture of Bursa in 1324 provided the first means for developing the administrative, financial, and military power necessary to make the principality into a real state and to create an army. It was Orhan who began the military policy (successfully developed by his successors) of employing Christian mercenary troops, thus lessening his dependence on the nomads as well as providing fighting forces better able to meet the military needs of the time. Orhan soon was able to capture the remaining Byzantine towns in northwestern Anatolia: İznik (1331), İzmit (1337), and Üsküdar (1338). He then moved against his major Turkmen neighbours to the south. Taking advantage of internal conflicts he annexed Karası in 1345 and gained control of the area between the Gulf of Edremit and Kapı Dağı (Cyzicus), reaching the Sea of Marmara. He thus put himself in a position to end Aydin's lucrative monopoly in providing mercenary troops to competing Byzantine factions in Thrace and Constantinople. In 1346 Orhan replaced Aydın as the principal ally of the Byzantine emperor John VI Cantacuzenus. The consequent entry of Ottoman troops into Europe gave them a direct opportunity to see the possibilities for conquest offered by the decline of Byzantium. The collapse of Aydın following the death of its ruler, Umur Bey, left the Ottomans alone as the leaders of the *gazis* against the Byzantines. Orhan helped Cantacuzenus take the throne of Byzantium from John V Palaeologus, and as reward secured the right to ravage Thrace and to marry the emperor's daughter Theodora. Ottoman raiding parties began to move regularly through Gallipoli into Thrace. Huge quantities of captured booty strengthened Ottoman power and attracted into Ottoman service thousands from the uprooted masses of Anatolia. Starting in 1354, Orhan's son Süleyman transformed Gallipoli, on the European side of the Dardanelles, into a permanent base for expansion into Europe, and refused to leave, despite the protests of Cantacuzenus and others. From this base, his bands moved into the Balkans, up the Maritsa River, raiding as far as Adrianople (Edirne). Cantacuzenus soon fell from power, at least partially because of his

Capture of Bursa

Gallipoli used as base for expansion



Expansion of the Ottoman Empire.
From S. Fischer, *The Middle East: A History* (1960); Alfred A. Knopf

cooperation with the Turks, and Europe began to be aware of the extent of the Turkish danger.

It was only under Orhan's son Murad I that Gallipoli was used for permanent conquests. Constantinople itself was bypassed because its thick walls and well-organized defenses, despite the weakness and disorganization of its defenders, remained too strong for the rudimentary Ottoman army. Murad's initial conquests were northward into Thrace, culminating with the capture in 1361 of Adrianople—the second city of the Byzantine Empire; renamed Edirne, the city became the new Ottoman capital, providing the Ottomans with a centre for the administrative and military control of Thrace. As the main fortress between Constantinople and the Danube, it controlled the principal invasion road through the Balkan mountains, assured Ottoman retention of their European conquests, and gave them the means for further expansion to the north. Murad then moved through the Maritsa Valley and captured Philippopolis (Filibe) in 1363. Control of the main sources of Constantinople's grain and tax revenues enabled Murad to force the emperor to accept his suzerainty. The death of the Serbian emperor Stefan Dušan in 1355 left his successors too divided and weak to defeat the Ottomans, despite an alliance with King Louis the Great of Hungary and Tsar Shishman of Bulgaria in the first European crusade against the Ottomans. The Byzantine emperor John V tried to mobilize European assistance by uniting the churches of Constantinople and Rome, but only further divided Byzantium without assuring any concrete help from the West. Murad was thus able to rout the allies at Chirmen, on the Maritsa in 1371, increasing his own confidence and demoralizing his smaller enemies, who rapidly accepted his suzerainty without further resistance. Murad next inaugurated an Ottoman empire of vassals in Europe. He retained local native rulers, who in return accepted his suzerainty, paid annual tributes, and provided contingents for his army when required to do so. This policy enabled the Ottomans to avoid a great deal of local resistance to conquest by assuring rulers and subjects alike that their lives, properties, traditions, and positions would be preserved if they peacefully accepted Ottoman rule. It also enabled the Ottomans to govern the newly conquered areas without building up a vast administrative system of their own or maintaining occupation garrisons.

Creation of European vassals

Moving rapidly to consolidate his empire south of the Danube, Murad captured Macedonia (1371–87), central Bulgaria (including Monastir (1382), Sofia (1385), and Niš (Nish) (1386), and Serbia, all culminating in the climactic defeat of the Balkan allies at the Battle of Kosovo (Kosovo Polje) in 1389. South of the Danube, only Walachia, Bosnia, Albania, Greece, and the Serbian fort of Belgrade remained outside Ottoman rule, and to the north, Hungary alone was in a position to resist further Muslim advances.

Murad was killed during the Battle of Kosovo. His son and successor Bayezid I was unable to take advantage of his father's victory to achieve further European conquest, and was, in fact, compelled to restore the defeated vassals and return to Anatolia to face the rising threat of the Turkmen principality of Karaman, created on the ruins of the Seljuq Empire of Anatolia, with its capital at Konya. Bayezid's predecessors had avoided forceful annexation of Turkmen territory in order to concentrate on Europe. They had, however, expanded by such peaceful means as marriage alliances and the purchase of territories. The acquisition of territory in central Anatolia from the amirates of Hamid and Germiyan had brought the Otto-

mans into direct contact with Karaman for the first time. Murad had been compelled to take some military action to prevent Karaman from occupying his newly acquired territories, but once this was accomplished he had turned back to Europe leaving the unsolved problem to his successor son. Karaman willingly cooperated with Serbia in stirring opposition to Ottoman rule among Murad's vassals in both Europe and Anatolia. It had strengthened the Balkan Union that was routed by the Ottomans at Kosovo, and had stimulated a general revolt in Anatolia that Bayezid was forced to meet by an open attack as soon as he was able to do so. By the end of 1390, Bayezid had overwhelmed and annexed all the remaining Turkmen principalities in western Anatolia. He attacked and defeated Karaman in 1391, annexed several Turkmen states in eastern Anatolia, and was preparing to complete his conquest in the area when he was forced to turn back to Europe to deal with a revolt of some of his Balkan vassals, encouraged and helped by Hungary and Byzantium. Bayezid quickly smashed the rebels (1390–93), occupied Bulgaria and put it under direct Ottoman administration for the first time, and put Constantinople under siege. In response, Hungary organized a major European crusade against the Ottomans. The effort was beaten back by Bayezid at Nicopolis (Niğbolu) on the Danube in 1396. Europe was terrorized and Ottoman rule south of the Danube was so assured, and Bayezid's prestige in the Islāmic world was so extended that he was given the title of *sulṭān* by the shadow ʿAbbāsid caliph of Cairo—this despite the opposition of the caliph's Mamlūk masters, the rulers of Egypt, Syria, and the Holy Cities, who wanted to retain the title only for themselves.

War with Timur

Turning back to Anatolia to complete the conquests aborted by his move against the crusaders, Bayezid thoroughly overran Karaman, the last Turkmen principality, in 1397. His advances, however, now attracted the attention of Timur (Tamerlane), who had been building a powerful Tatar empire in Central Asia, Iran, Afghanistan, and Mesopotamia, and whose invasion of India in 1398 had been halted by his fear of the rising Ottoman power on his western flank. Encouraged by several Turkmen princes who had fled to his court when their territories were taken by Bayezid, Timur decided to destroy Bayezid's empire before resuming his campaigns in India and thus invaded Anatolia. As Bayezid and Timur moved toward battle, the former's Turkmen vassals and Muslim followers deserted him because he had abandoned the old Ottoman *gazi* tradition of advancing against the infidel. Left only with forces provided by his Christian vassals, Bayezid was decisively overwhelmed by Timur at the Battle of Ankara in 1402. Taken captive by the victor, he died within a year.

Restoration of the Ottoman Empire, 1402–81. Timur's objective in Anatolia had not been conquest, and he followed his victory by retiring from Anatolia after restoring to power the Turkmen princes who had joined him. Even Bayezid's sons were able to assume control over the family's former possessions in western Anatolia, and the Ottoman Empire in Europe was left largely untouched. At this time a strong European crusade might have pushed the Ottomans out of Europe altogether, but weakness and division south of the Danube and diversion in other matters to the north left an opportunity for the Ottomans to restore what had been torn asunder without significant loss.

Disputed succession

But internal divisions were to hinder Ottoman efforts to restore their power. Bayezid's four sons fought for the right of succession. His eldest son Süleyman assumed control in Europe, with his capital at Edirne, and gained the support of the Christian vassals and those who had stimulated Bayezid to turn toward conquest in the East. The descendants of the Turkmen notables who had assisted the early Ottoman conquests in Europe supported the claims of Mehmed, who with the additional support of the Anatolian Muslim religious orders and artisan guilds was able to defeat and kill his brothers Mûsa Bey, who had established his capital at Bursa, and İsa Bey of Balıkesir, in southwestern Anatolia, as well as Süleyman, and so assume undisputed possession of the entire empire as Sultan Mehmed (Muḥammad) I. He was to reign from 1413 to 1420.

Under Mehmed and Murad II (ruled 1421–51), there was a new period of expansion in which Bayezid's empire was restored and new additions made. Mehmed restored the vassal system in Bulgaria and Serbia and promised the vassal princes that he would not undertake new European adventures and would restore his position within the state. Murad II was also compelled to devote most of the early years of his reign to internal problems, and particularly to the efforts of the *gazi* commanders and Balkan vassal princes in Europe, as well as the Turkmen vassals and princes in Anatolia, to retain the autonomy and even independence gained during the decade following the Battle of Ankara. In 1422–23, Murad suppressed the Balkan resistance and put Constantinople under a new siege that ended only after the Byzantines had provided him with huge amounts of tribute. He then restored Ottoman rule in Anatolia and eliminated all of the Turkmen principalities left by Timur with the exception of Karaman and Candar (Jandar), which he left autonomous though tributory so as not to excite the renewed fears of Timur's successors in the east. Murad then inaugurated the first Ottoman war with Venice (1423–30), which had maintained friendly relations with the sultans in order to develop a strong trade position in the Ottoman dominions and the Black Sea area. Venice, however, had accepted Salonica from Byzantium in order to prevent Ottoman expansion across Macedonia to the Adriatic, which it considered to be a Venetian lake. The war was indecisive for some time; Venice was diverted by conflicts in Italy, and the Ottomans needed time to build a naval force sufficient to compete with that of the Venetians. In addition, Murad was diverted by an effort of Hungary to establish its rule in Walachia, between the Danube and the Transylvanian Alps, a move which inaugurated a series of Ottoman-Hungarian conflicts that were to occupy much of the remainder of his reign. Murad finally built a fleet strong enough to blockade Salonica and enable his army to take it (1430), after which Ottoman naval raids against Venetian ports in the Adriatic and the Aegean compelled Venice in 1432 to make a peace in which it abandoned its efforts to prevent the Ottoman advance to the Adriatic, and was allowed to assume a role as the leading commercial power in the sultan's dominions.

Outbreak of Venetian War

Murad, who had been put on the throne by Turkish notables who had joined the Ottoman state during the first century of its existence, soon began to resent the power they had gained in return as well as in consequence of the great new estates they had built up in the conquered areas of Europe and Anatolia. To counteract their power, he began to build up the power of various non-Turkish groups in his service, particularly those composed of Christian slaves and converts to Islām, whose military arm was organized into a new infantry organization called the Janissary corps. To strengthen this group, Murad began to distribute most of his new conquests to its members, and to add new supporters of this sort he developed the famous *devşirme* system by which Christian youths were drafted from the Balkan provinces for conversion to Islām and life service to the sultan. With their revenues and numbers increasing, the *devşirme* men and their supporters achieved considerable political power. Because the new European conquests were being used by the sultan to build up the *devşirme* relative to the Turkish notables, the former wanted them to continue and expand, while the latter opposed them. Murad, wanting to return to aggressive policies of European expansion in order to help the *devşirme* reduce the power of the Turkish notables, renewed the struggle with Hungary in Serbia and Wallachia starting in 1434. He took advantage of the death in 1437 of the Hungarian king Sigismund to reoccupy Serbia, excepting Belgrade, and to ravage much of Hungary, and he then annexed Serbia in 1439, beginning a policy of replacing the vassals with direct Ottoman rule throughout the empire. Hungarian control of Belgrade was now the major bar to large-scale advances north of the Danube. Ottoman attacks on it and raids on Transylvania failed to move the Hungarians, largely because

Organization of the Janissary corps

of the leadership of János Hunyadi (*q.v.*), originally a leader of the Walachian border resistance to the *gazis* in 1440–42. Although Murad finally defeated Hunyadi at the Battle of Zlatica (İzladi) in 1443, the increased influence of the Turkish notables at Murad's court led the sultan to agree to the Peace of Adrianople in 1444. By its terms Serbia regained its autonomy, Hungary kept Wallachia and Belgrade, and the Ottomans promised to end their raids north of the Danube. In 1444 Murad also made peace with his main Anatolian enemy, Karaman, and retired to a life of religious contemplation, voluntarily passing the throne to his son Mehmed II who, still very young, already showed the qualities of leadership that were later to distinguish his long reign. The Byzantines and the pope sought to use the opportunity created by the existence of a youthful sultan to push the Ottomans out of Europe, organizing a new crusade joined by Hungary and Venice when the pope assured them that they were not bound to honour the peace treaty they had signed with Muslim infidels. A crusader army moved through Serbia across the Balkan Mountains to the Black Sea at Varna, where it was to be supplied and transported to Constantinople by a Venetian fleet, which was to sail through the Straits, at the same time using its power to prevent Murad from returning from Anatolia with the bulk of the Ottoman army. Though the crusaders reached Varna, a Serbian decision to remain loyal to the sultan, combined with Venetian reluctance to fulfill its part of the agreement for fear of losing its trade position should the Ottomans win, left the crusaders stranded. Further quarrels among the crusade leaders gave Murad time to return from Anatolia and organize a new army. The Turkish victory at Varna on November 10, 1444, ended the last important European crusading effort against the Ottomans. Murad now reassumed the throne and brought back to power with him the *devşirme* party, whose insistent demands for conquest led him to spend the remainder of his reign eliminating the vassals and establishing direct rule in much of Thrace, Macedonia, Bulgaria, and Greece. In the process, he divided the newly acquired lands into estates the revenues of which further increased the power of the *devşirme* at the expense of the Turkish notables. Only Albania was able to resist because of the leadership of its national hero, Skanderbeg (George Kastrioti). He was routed by the sultan at the second Battle of Kosovo (1448). By the time of Murad's death in February 1451, the Danube frontier was secure, and it seemed that the Ottoman Empire was in Europe to stay.

New European crusade

Whereas the victory at Varna brought new power to the *devşirme* party, the grand vizier Candarlı Halil Paşa was able to retain a dominant position for the Turkish notables whom he led by keeping the confidence of the sultan and by successfully dividing his opponents. Prince Mehmed therefore became the candidate of the *devşirme*, and it was only with his accession that they were able to achieve the political and military power made possible by the financial base built up during the previous two decades. Under Sultan Mehmed II (1451–81), the *devşirme* increasingly came to dominate, and to press their desire for new conquests in order to take fuller advantage of the situation created at Varna. Constantinople became their first objective. To Mehmed and his supporters, the Ottoman dominions in Europe could never reach their full extent or be molded into a real empire so long as their natural administrative and cultural centre remained outside their hands. The grand vizier (the sultan's chief adviser) and other Turkish notables bitterly opposed the attack, ostensibly because it might draw a new crusade, but in fact because of their fear that the capture of the Byzantine capital might bring about a final triumph of the *devşirme*. The siege (April 6–May 29, 1453) and conquest of Constantinople and its transformation into the Ottoman capital of Istanbul marked an important new state in Ottoman history. Internally, it saw the end of power and influence for the old Turkish nobility, whose leaders soon were executed or exiled and whose properties were confiscated, and the triumph of the *devşirme* and their supporters. Externally, the conquest made Mehmed II the most famous ruler in the Muslim world, even though the lands of the old caliphate still remained in the hands of the Mamlūks of Egypt and Timur's successors in Iran. Moreover, it was not long before possession of Constantinople stimulated Mehmed to nourish an ambition to place under his dominion not merely the Islāmic and Turkic worlds but also a re-created Byzantine Empire and, perhaps, the entire world of Christendom.

Final attack on Constantinople

To pursue these objectives, Mehmed II developed various bases of power. Domestically, his primary objective was to restore Istanbul as the political, economic, and social centre of the area that it formerly had dominated. To this end, he worked to repopulate the city, not only with its former inhabitants but also with elements of all the conquered peoples of the empire, whose residence and intermingling there would provide a microcosm for a similar process that Mehmed hoped would weld the entire empire into a powerful and integrated whole. Special tax concessions were established to encourage the most active and skilled of his subjects to settle in the capital. The major religious groups were allowed to establish their own self-governing communities, called *millets*, under the leadership of their religious chiefs, each retaining its own civil laws, traditions, and language under the general protection of the sultan. Mehmed also worked to restore the physical aspects of the city. Old buildings were repaired; streets, aqueducts, and bridges were constructed; sanitary facilities were modernized; and a vast supply system was set up to provide for the city's inhabitants. Great attention was paid to restoring Istanbul's industry and trade, with special concessions to attract merchants and artisans from all parts of the empire.

Restoration of Istanbul

Mehmed also devoted much time in expanding his dominions in Europe and Asia in order to establish his claim to world leadership. To this end, he eliminated the last princes who might dispute his claims to be legitimate successor to the Byzantine and Seljuq dynasties, and replaced the remaining vassal princes with direct, Ottoman administration. In addition, he extended Ottoman rule far beyond the territories inherited from Murad II; from 1454 to 1463, he concentrated mainly on southeastern Europe, annexing Serbia (1454–55), conquering the Morea (1458–60), and eliminating in the process the last major claimants to the Byzantine throne. When Venice refused to surrender its important forts on the Aegean coast of the Morea, Mehmed inaugurated the second Venetian-Ottoman war (1463–79). At the same time he annexed Trebizond (1461) and the Genoese commercial colonies that had survived along the Black Sea coast of Anatolia, including Sinop and Kafa, and began the process by which the Crimean Tatar *khāns* were compelled to accept Ottoman suzerainty. In 1463 he occupied and annexed Bosnia, helped by the native Bogomils, an evangelical Christian sect that had been badly treated by the Catholic Hungarians. When Albania continued to hold out, helped by supplies sent by sea from Venice, Mehmed sent in large numbers of Turkmen irregulars, who in the process of conquering Albania settled there and formed a Muslim community. Whereas the papacy and Venice were unable to raise a new crusade, they were able to divert Mehmed by encouraging attacks by his Eastern enemies, the Turkmen principality of Karaman and the Tatar Ak Koyunlu ("White Sheep") dynasty, which under the leadership of Uzun Ḥasan, had replaced Timur's descendants in Iran. Mehmed, however, skillfully used dynastic divisions to conquer Karaman in 1468, extending direct Ottoman rule in Anatolia to the Euphrates. When Uzun Ḥasan responded by invading Anatolia in the company of many Turkmen princes, Venice intensified its attacks in the Morea, Hungary moved into Serbia, and Skanderbeg attacked Bosnia. Mehmed, however, was able to defeat these enemies, one after the other. In 1473 he routed Uzun Ḥasan, who acknowledged Ottoman rule in all of Anatolia and returned to Iran. This brought the Ottomans into conflict with the Mamlūk empire in Syria, which Mehmed neutralized, though he could not defeat it. He then turned to Venice, and several naval raids along the Adriatic coast finally led to a peace in 1479 whereby Venice surrendered its bases in Albania and the

Second Venetian war

Morea and agreed to pay a regular annual tribute in return for restoration of its commercial privileges. Mehmed then used his new naval power to send a large force that landed at Otranto in southern Italy in August 1480, and to attack Rhodes. Success seemed within sight when his premature death in 1481 brought his effort to an end. But he had, indeed, laid the foundations for Ottoman rule in Anatolia and southeastern Europe that was to survive for the next four centuries.

In addition to conquering a large empire, Mehmed worked to consolidate it and to codify the political, administrative, religious, and legal institutions developed during the previous century in a series of law codes called *kanun-names*. The immensity of the task, however, and his diversion in numerous campaigns delayed the process to such an extent that it was completed only during the mid-16th century. Nor was Mehmed overly successful in building the economic and social bases of his empire. His most important problem was securing sufficient money to finance his military expeditions and the new apparatus of government and society. The tax systems inherited from his predecessors did not provide the required money, particularly because most of the conquered lands were turned into estates whose taxes went entirely to their holders in return for military and administrative services. Mehmed therefore turned to a number of expedients that achieved their immediate objectives, but at the cost of grave economic and social difficulties. He regularly withdrew all coins from circulation and issued new ones with a larger proportion of base metal alloys. To enforce acceptance of the new issues, he sent armed bands around the empire with the right to confiscate without compensation all of the older and more valuable coins that had not been voluntarily exchanged for the new. The debasement of the coinage soon caused a rapid inflation, and this in turn greatly disturbed the industry and trade that the sultan had hoped to assist. In addition, in his search for revenues, Mehmed created monopolies over the production and use of essential goods, distributing them among the highest bidders, who in turn charged excessive prices and created artificial scarcities to secure their profits. Finally, Mehmed established the principle that all revenue-producing property belonged to the sultan. In pursuance of this, he confiscated much private property and religious foundation lands, creating tremendous resentment and opposition among those who lost their revenues, including members of the religious *'ulamā'* class, the Turkish notables, and even some *devşirme* men, whose discontent threatened to undermine both state and sultan. It was only by playing these groups off against each other that Mehmed was able to maintain his own position and power and to continue his conquests.

Financial difficulties

Ottoman institutions in the 14th and 15th centuries. Ottoman dynasts were transformed from simple tribal leaders to border princes (*uc bey*) and *gazi* leaders under Seljuq and then Īl-Khānid suzerainty in the 13th and early 14th centuries. With the capture of Bursa, Orhan had been able to declare himself independent of his suzerains and assume the title of *bey*, which was retained by his successors until Bayezid I was named sultan by the shadow 'Abbāsid caliph of Cairo following his victory over the Christian crusaders at the Battle of Nicopolis (1396). These title changes were not immaterial in the position of the Ottoman ruler within the state and in the organization of the state itself. As *uc bey* and even *bey*, the Ottoman leader remained little more than a tribal chief, sharing administrative and military leadership with the Turkmen chiefs surrounding him. Like the tribal chiefs, he had the right to the loyalty and obedience of his followers only so long as he led them to victory, and only in relation to his military functions. Beyond this, he was only one among equals in the councils that decided general internal policies; the tribes and clans remained autonomous in their internal affairs. The *bey* was accessible to the tribe and clan leaders as well as to their followers. He could intervene in disputes among the clans, but jurisdiction was limited indeed. Muslim law and jurists had little influence, whereas Turkish tribal law and custom prevailed. In such a situation, the idea of rule was very limited; administration was conceived mainly in financial terms, with each clan or family or tribe accepting Ottoman military leadership largely for the financial rewards it could bring. Ottoman chiefs collected the booty in conquered lands, and following the conquest had the right to collect taxes from lands left in their possession. The only advantage that the *bey* had over the chiefs surrounding him was, as tribal war leader, the right called *pençik*, to collect an extra fifth of the booty taken by his followers. Because the *bey* was dependent for his power and revenues on the assent of his followers, his authority was limited in scope and in time.

Powers of Ottoman rulers

But as the territory of the Ottoman principality expanded, and the Ottomans fell heirs to the administrative apparatus left by the Byzantines, this simple tribal organization was replaced by a more complex form of government, so that by the time the Ottoman rulers became sultans, they already had far more extensive power and authority than had been the case a half-century earlier. The simple tribal organization of the Ottoman *bey* could suffice only for so long as the state was small enough for the individual tribal leaders to remain on their lands and fight the nearby enemy at the same time. But as the principality expanded and the frontiers and enemies became further removed from previously conquered territory, the financial and administrative functions at home had to be separated from the military. Taxes had to be collected to exploit the conquered territories and support the officers and soldiers while they were away. The treasury of the sultan had to be separated from that of the state so that each would have an independent income and organization. Throughout the 14th and 15th centuries, the Ottoman state gradually evolved its institutions of government and the army to meet the needs of administering and defending an expanding empire. As this was done, it was natural that it should be influenced by those states that had preceded it, not only in the areas it came to rule, but also in the lands of its ancestors. So it was that the developing Ottoman state was influenced by the traditions of the nomadic Turkic empires of Central Asia, particularly in military organization and tactics. It was also influenced heavily by the classical High Islāmic civilization of the 'Abbāsids, as passed through the hands of the Seljuqs, particularly in the development of orthodox Islām as the basis of its institutions of administration, religion, law, and education, and in the development of the *mukataa* as the basic unit of its administrative and financial systems. In the court hierarchy, the central financial structure, and the tax and administrative organizations developed in the European provinces, the Ottomans were influenced by the Byzantines and, to a lesser extent, by the Serbian and Bulgarian empires. Whereas conversion to Islām was not demanded of those conquered at this time, many Christians and Jews voluntarily converted to secure full status in the new empire. Most, however, continued to practice their old religions without restriction. A particularly important source of Christian influence during the 14th century came from the close marriage ties between the Ottoman and Christian courts. Sultan Orhan married Theodora, daughter of Byzantine emperor John VI Cantacuzenus. Theodora was the mother of Murad I, who in turn married Byzantine and Bulgarian princesses, whereas Bayezid I married Despina, daughter of the Serbian prince Lazar. Each of these marriages brought Christian followers and advisers into the Ottoman court, and it was under their influence that Bayezid I in particular abandoned the simple nomadic courts and practices of his predecessors and isolated himself behind elaborate court hierarchies and ceremonials borrowed primarily from the Byzantines. At the same time, the Greek and Serbian languages tended to dominate Ottoman court life, and to a lesser extent its administration.

Influences of early states

The triumph of Sultan Mehmed I in 1413 was due at least in part to the support of the Turkish notables and Muslim religious orders of Anatolia who strongly resented the Christian predominance in Bayezid's court and

attributed his abandonment of the *gazi* tradition and attacks in Turkish Muslim Anatolia to their influence. As a result, Turkish and Muslim influences dominated the Ottoman court during the 15th century, although the hierarchies, institutions, and ceremonies introduced in the previous century remained largely without change.

Growth of administration

The same process that isolated the sultans from their subjects also removed them from the daily administration of government. Formal institutions of administration therefore had to be evolved to take their place, with the rulers delegating more and more of their duties to executive ministers, to whom the Seljuq title *vezīr* (vizier) was given. The continued close connections of the Ottoman family with the urban guilds and orders of Anatolia, many of the members of which were descendants of officials of the Great Seljuq and Īl-Khānid empires, as well as the empire of the Seljuqs of Konya, provided continuity with the Islāmic Turkish traditions of government. With them came the basic unit of Islāmic administrative and financial organization, the *mukataa*, which intimately associated each office with a source of revenues and made each official the collector of his own salary, at the same time that it circumscribed his administrative powers within those tasks directly involved with the financial function. It was relatively simple for the Ottomans to preserve previous methods of taxing on the local levels in different parts of the empire while weaving them into a united whole through the veneer provided by the *mukataa* units in which the resulting tax revenues were assigned to Ottoman officials.

As the central administration was divided into functional departments, a vizier was appointed to direct each. Most of the early viziers were former Turkmen princes who had entered Ottoman service, though some, particularly under Bayezid I, were Christians and Christian converts. State policy was discussed and decided in a council (*divan*) of these viziers, who were joined by religious, judicial, and military leaders under the direction and chairmanship of the sultan. As the duties of the state became more extensive and complex, the individual viziers gained increased financial and political power, and as the Byzantine influence caused the sultan to isolate himself, it was inevitable that the viziers would come to dominate. As if to emphasize his removal from the daily affairs of state, the sultan began to appoint one of his viziers as his chief minister, or grand vizier. From 1360 to the conquest of Constantinople, this powerful position was reserved for members of the Çandarlı family, which came to lead and represent the powerful and assertative Turkmen notable families, who thus benefitted most from the 14th-century expansion of the empire.

Office of grand vizier

The first Ottoman army had been composed entirely of Turkmen nomads, who had remained largely under the command of their own tribe and clan leaders, and under the influence of the *şeyh*s of the heterodox and mystic religious orders that had converted most of them to Islām. Armed with bows and arrows and spears, these nomadic cavalrymen had lived mostly on booty, but those assigned as *gazi*s to border areas or sent to conquer and raid Christian lands had also been given more permanent revenues in the form of taxes levied on the lands they garrisoned. These revenue holdings were formalized as *mukataʿa*s, with the tribal leaders and *gazi* commanders holding them and collecting their revenues to feed, supply, and arm their followers. It was this type of *mukataa* that developed into the Ottoman form of fief called *timar*, that was the basis of Ottoman military and administrative organization as the European portions of the empire were conquered from the vassals in the 15th century and placed under direct Ottoman administration.

These troops had predominated through Orhan's reign until he had seen that such mounted and undisciplined men were of limited use in besieging and taking large cities. In addition, once he had established his state, he had found it difficult to maintain order with such an army because the nomads still preferred to maintain themselves by the traditional forms of looting, in the lands of their commander as well as in those of the enemy. To replace the nomads, Orhan had organized a separate standing army of hired mercenaries paid by salary rather than booty or by *timar* estates. Those mercenaries organized as infantry were called *yaya*s, those organized as cavalry, *müsellem*s. Although the new force included some Turkmen who were content to accept salaries in place of booty, most of its men were Christian soldiers from the Balkans who were not required to convert to Islām as long as they obeyed their Ottoman commanders. As Murad I had conquered more and more of southeastern Europe these forces had become mainly Christian, and as they came to dominate the Ottoman army, the older Turkmen cavalrymen were maintained mainly as irregular shock troops, called *akıncı*s, who were compensated only by booty. As the *yaya*s and *müsellem*s expanded in numbers, their salaries became too burdensome for the Ottoman treasury, so in most cases the newly conquered lands were assigned to their commanders in the form of *timar*s. It was this new regular army that developed the techniques of battle and siege used to achieve most of the 14th-century Ottoman conquests, but because it was commanded by members of the Turkish notable class, it became the major vehicle for their rise to predominance over the sultans, whose direct military supporters were limited to the vassal contingents.

Growth of a regular army

Only late in the 14th century did Murad I and Bayezid I attempt to build up their personal power by building a military slave force for the sultan under the name *kapıkulu* (pl. *kapikullari*), or slaves of the Porte. Murad based the new force on his right to a fifth of the war booty, which he interpreted to include captives taken in battle. As these men entered his service, they were converted to Islām and trained as Ottomans, gaining the knowledge and experience required for service in the government, as well as the army, while remaining in the sultan's personal service. During the late 14th century, this force became the most important element of the Ottoman army, particularly its infantry branch, called *yeniçeri* ("new soldiers"), or the Janissary corps. The provincial forces provided by the timar holders comprised the Ottoman cavalry and were called *sipahi*s, while the irregular *akıncı*s and salaried *yaya*s and *müsellem*s were relegated to rear-line duties and lost their military and political importance. But when Bayezid I abandoned the *gazi* tradition and moved into Anatolia, he lost the support of the Turkish notables and their *sipahi*s before his new *kapıkulu* army was fully established. He thus had to rely only on the Christian vassal forces in the Battle of Ankara, and whereas they demonstrated considerable valour and fighting ability, they were not alone sufficient to resist Timur's powerful army.

When the Ottoman Empire was restored under sultan Mehmed I, the Turkish notables, in order to deprive the sultan of the only military force he could use to resist their control, required him to abandon the *kapıkulu* as contrary to the Islāmic tradition that Muslims could not be kept in slavery. The European and Anatolian revolts that arose early in the reign of Murad II were at least partly stimulated and supported by members of the *kapıkulu* as well as the Christian slaves and vassals who had been losing their power to the Turkish notables. But as soon as Murad II was established, he resumed earlier efforts to make the sultanate more independent, building up the strength of the Janissaries and their associates and playing them off against the notables. He distributed most of his conquests to members of the *kapıkulu* force, occasionally as *timar*s, but more often as tax farms (*iltizam*s), so that the treasury could obtain the money it needed to maintain the Janissary army entirely on a salaried basis. In addition, in order to man the new force, Murad developed the *devşirme* system of recruiting the best Christian youths from southeastern Europe.

The *devşirme* system

Whereas Mehmed II used the conquest of Constantinople to destroy the major Turkish notable families, and build up the power of the *devşirme*, he sought only to establish a balance of power and function between the two groups so that he could use and control both for the benefit of the empire. So it was that he enlarged the concept of *kapıkulu* to include members of the Turkish nobility and their Turkmen followers as well as the products of the

devşirme. Now only persons accepting the status of slaves of the sultan could hold positions in the Ottoman government and army. Persons of Muslim and non-Muslim origin could achieve this status, so long as they accepted the limitations involved, namely absolute obedience to their master and the devotion of their lives, properties, and families to his service. After this time, all important ministers, military officers, judges, governors, *timar* holders, tax farmers, Janissaries, *sipahis*, and the like were made members of this class and attached to the will and service of the sultan. The salaried Janissary corps remained the primary source of strength of the *devşirme* class whereas the *sipahis* and the *timar* system remained the bases of power of the Turkish notables.

Mehmed II thus avoided the fate of the great Middle Eastern empires that had preceded that of the Ottomans, in which rule had been shared among members of the ruling dynasty and with others, and rapid disintegration had resulted. The Ottomans established the principle of indivisibility of rule, with all members of the ruling class subjected to the absolute will of the sultan. To assure this, Mehmed II began the process of developing a firm law of succession, beginning the practice of executing all the brothers of the reigning sultan so that succession would be limited to one of his sons, preferably the ablest one.

THE PEAK OF OTTOMAN POWER, 1481–1566

Domination of southeastern Europe and the Near East. During much of the century that followed the reign of Mehmed II, the Ottoman Empire achieved the peak of its power and wealth. New conquests extended its domain well into central Europe and throughout the Arab portion of the old Islāmic caliphate; and a new amalgam of political, religious, social, and economic organizations and traditions was institutionalized and developed into a living, working whole.

Bayezid II (ruled 1481–1512). The reign of Mehmed II's immediate successor, Sultan Bayezid II, was largely a period of rest. The previous conquests were consolidated, and many of the political, economic, and social problems caused by Mehmed's internal policies were dealt with, leaving a firm foundation for the conquests of the 16th-century sultans.

The economic stringencies imposed to finance Mehmed II's campaigns had led during the last year of his reign to a virtual civil war encouraged and participated in by the major factions in Istanbul—the *devşirme* party and the Turkish aristocracy. Bayezid was put on the throne by the Janissaries because of their military domination of the capital, whereas his more militant brother Cem fled to Anatolia, where he led a revolt initially supported by the Turkish notables. Bayezid managed to conciliate the latter, however, by exposing to them his essentially pacific plans, which he concealed from the Janissaries. Left without major support, Cem fled into exile to Mamlūk Syria in the summer of 1481. He returned the next year with the help of the Mamlūks and the last Turkmen ruler of Karaman, but his effort to secure the support of the Turkmen nomads failed because of Bayezid's heterodox religious policies. Cem remained in exile, first at Rhodes, then with the pope in Rome, until his death in 1495. European efforts to use him as the spearhead of a new crusade effort against Istanbul were unsuccessful.

Conciliation of Turkish notables

In the meantime, however, the threat that Cem might lead a foreign attack compelled Bayezid to concentrate on internal consolidation. Most of the property expropriated by his father was restored to its original owners. Equal taxes were established around the empire so that all subjects could fulfill their obligations to the government without the kind of disruption and dissatisfaction that had characterized the previous regime. Particularly important was the establishment of the *avârız-i divaniye* (war chest) tax, which provided for the special expenditures of war without special confiscations or heavy levies. The value of the coinage was restored, and Mehmed II's plans for economic expansion were at long last brought to fruition. The vassal system was replaced by direct Ottoman administration throughout the empire. For the first time the central government was given a budget system in which expenditures were balanced against revenues on a regular basis.

Culturally, Bayezid stimulated a strong reaction against the Europeanizing trends of the previous half century. The Turkish language and Muslim traditions were emphasized. He worked to develop and establish the institutions of orthodox Islām in order to save the empire from the increasing menace of heterodox Shīʿism among the tribes of eastern Anatolia, until, late in his life, his own tendency toward mystic Ṣūfism led him largely to abandon the effort.

Though Bayezid preferred to maintain peace so as to have the time and resources to concentrate on internal development, he was forced into a number of campaigns by the exigencies of the time and the demands of his more militant *devşirme* followers. In Europe, he rounded off the empire south of the Danube and Sava by taking Hercegovina (1483), leaving only Belgrade outside Ottoman control. The Hungarian king Matthias Corvinus (ruled 1458–90) was interested mainly in establishing his rule over Bohemia and agreed to peace with the Ottomans (1484), and after his death, struggles for succession left this front relatively quiet for the remainder of Bayezid's reign. To the northeast, the sultan pushed Ottoman territory north of the Danube, along the shores of the Black Sea, capturing (1484) the ports of Kilia and Akkerman, which controlled the mouths of the Danube and Dniester, thus giving the Ottomans control over the major entrepôts of northern Europe's trade with the Black Sea and Mediterranean. Bayezid was there able to compel Moldavia to accept his suzerainty—an important step toward the incorporation of the Romanian principalities in his empire. Because these advances conflicted with the ambitions of Poland, war followed (1483–89) until the diversion of Poland by the threat of Muscovy under Ivan the Great (ruled 1462–1505) left this front quiet also after 1484.

Extension of the empire

Bayezid then turned to the East, where previous conquests as far as the Euphrates had for the first time brought the Ottomans up to the Mamlūk Empire. Conflict over control of the small Turkmen principality of Dulkadir (Dhū al-Qadr), which controlled much of Cilicia and the mountains south of Lake Van, and an Ottoman desire to share in the control of the Muslim holy cities of Mecca and Medina, led to an intermittent war (1485–91); but there were no concrete results, and Bayezid's disinclination to commit major forces to the endeavour led to dissention and criticism on the part of his militant followers. To counter this, Bayezid tried to use Hungarian internal dissention to take Belgrade; but he failed, and raiding forces sent into Transylvania, Croatia, and Carinthia were turned back.

In the same year that Cem died (1495) a new peace with Hungary left Bayezid's objectives unfulfilled, so he turned toward his other major European enemy, Venice, to rebuild his reputation. Venice had been encouraging revolts against the sultan in the Morea, Dalmatia, and Albania, which it had ceded to the Ottomans in 1479. It also gained control of Cyprus (1489), where it built a major naval base, which it refused to allow Bayezid to use against the Mamlūks, thus pointing up the strategic importance of Cyprus to the sultan. Bayezid also hoped to conquer the last Venetian ports in the Morea to establish the bases for complete Ottoman naval control of the eastern Mediterranean. All these objectives except control of Cyprus were achieved in the war that followed (1499–1503). The Ottoman fleet emerged for the first time as a major Mediterranean naval power, and the Ottomans became an integral part of European diplomatic relations.

Bayezid never was able to use this situation to make new conquests in Europe because the rise of revolts in eastern Anatolia occupied much of his attention during the last years of his reign. There the old conflict resumed between the autonomous, uncivilized nomads and the stable, settled Middle Eastern civilization—the Turkmen nomads resisted the efforts of the Ottomans to expand their administrative control to all parts of the empire. In reaction to the orthodox Muslim establishment, the nomads

showed a fanatical attachment to the leaders of the Ṣūfī and Shīʿah mystic orders, of whom the most successful were the Ṣafavids of Ardabīl, who at this time used a religious-military appeal to conquer most of Iran. Under Shāh Esmāʿīl (ruled 1502–24), the Ṣafavids sent missionaries throughout Anatolia, spreading a message of religious heresy and political revolt, not only among the tribesmen but also to cultivators and some urban elements, who began to see in this movement the answers to their own problems. A series of revolts resulted, with Bayezid, because of his involvements in Europe, able to do little to suppress them. Finally, at the start of the 16th century, a general Anatolian uprising forced Bayezid into a major expedition (1502–03), which pushed the Ṣafavids and many of their Turkmen followers back into Iran; there the Ottomans turned from Ṣūfism to Shīʿism as a means of gaining the loyalty of the Persians to a Turkish dynasty. Shāh Esmāʿīl continued, however, to spread his message as Ṣūfī leader in Anatolia, leading to a second major revolt of his followers against the Ottomans (1511). All the grievances of the time coalesced into what was basically a religious uprising against the central government, and only a major expedition led by the grand vizier Ali Paşa could suppress it. But the conditions that had caused the uprising remained a major problem for Bayezid's successor. In the end, Bayezid's increasingly mystical and pacific nature led the Janissaries to dethrone him in favour of his militant and active son, Sultan Selim I.

The Ṣafavids

Selim I (ruled 1512–20). Whereas Bayezid, despite his pacific nature, had been put on the throne by the Janissaries, Selim I was their candidate because he shared their desire to return to an aggressive policy of conquest. But Selim did not wish to be dependent on, or under the control of, those who had brought him to power; he killed not only his brothers but also all seven of their sons and four of his own five sons, leaving only the ablest, Süleyman, as the sole possible heir to the throne. This deprived potential opponents of alternative leaders around whom they could coalesce. Selim was then able to leave the *devşirme* in control of the government; but with a single heir, and with the sultan establishing his personal control over the Janissaries, it was he who dominated, rather than the *devşirme*.

Selim's ambitions encompassed Europe as well as Asia; but Bayezid had left the European fronts relatively quiet, so the new sultan turned first to the East, and chose the Ṣafavids of Iran as his initial victims. Selim first launched a vigorous campaign against the Ṣafavid supporters in eastern Anatolia, massacring thousands of tribesmen and missionaries and espousing a strict defense of Islāmic orthodoxy as a means of regaining political control. In the summer of 1514 he launched a major expedition against the Ṣafavids, hoping to add Iran to his empire and finally to eliminate the threat of heterodoxy. Esmāʿīl employed a scorched-earth policy, retiring into central Iran and hoping that winter would force the Ottomans to retire without a battle. But the militant Kızılbâş (Kizilbash) followers of the Ṣafavids forced the *shāh* to accept battle by intercepting the Ottomans before they entered Azerbaijan. The Ottoman and Ṣafavid armies clashed at Chāldirān, on the eastern side of the Euphrates (August 23, 1514), and the Ṣafavids were routed by Ottoman superiority in weapons and tactics (particularly because of Selim's use of cannons and gunpowder, in contrast to Ṣafavid reliance on spears and arrows). Though Azerbaijan was occupied, the Ottoman victory led to neither the conquest of Iran nor to the collapse of the Ṣafavid empire. The Ottoman army became increasingly discontented under the impact of Ṣafavid propaganda among the already-heterodox Janissaries and because of a relative lack of booty and supplies compared to campaigns in Europe. Selim was compelled to retire, and the Ṣafavids regained their lost province without resistance. The major result of the Chāldirān battle was to convince Shāh Esmāʿīl and his successors to avoid open conflict with the Ottomans at all costs—a policy followed for the next century. This preserved the Ṣafavid army, but it enabled Selim to overcome the last independent Turkmen dynasties in eastern Anatolia (1515–17) and to establish a strong strategic position relative to the Mamlūk Empire, which was falling into internal decay and was ripe for conquest by either the Ottomans or the Ṣafavids.

Defeat of the Ṣafavids

With Shāh Esmāʿīl still busy restoring his army, Selim I was able to overwhelm the Mamlūks in a single year's campaign during the summer and winter of 1516–17. The Mamlūk army fell easily to the well-organized and disciplined Ottoman infantry and cavalry, supported by artillery. The conquest was aided by the support of many Mamlūk officials, who betrayed their masters in return for important positions and revenues promised by the conquerors. In addition, most of the major populated centres of Syria and Egypt turned out their Mamlūk garrisons, preferring the security and order offered by the Ottomans to the anarchy and terror of the prior century of Mamlūk dominion.

Thus in a single sweep, Selim doubled the size of his empire, adding to it all of the lands of the old Islāmic caliphate with the exception of Iran, which remained under the Ṣafavids, and Mesopotamia, which was later taken by his successor. These acquisitions were of immense importance to the Ottomans. Under an efficient administration, the Arab world provided Istanbul with new revenues that solved the financial problems left from the 15th century and made the empire into one of the most powerful and wealthy states in the 16th century. Acquisition of the holy places of Islām cemented the position of the sultan as the most important ruler of Islām. The Ottomans had gained direct access to the intellectual, artistic, and administrative heritage of high Islāmic civilization, previously transmitted to them only indirectly. Now from the Arab world there came to Istanbul the leading Muslim intellectuals, artisans, administrators, and artists of the time. They penetrated every facet of Ottoman life and made the empire much more of a traditional Islāmic state than it ever had been before. Finally, the Ottomans replaced the Mamlūks in control of the Middle Eastern trade routes—part of the old international routes between Europe and the Far East.

Control of the Arab world

One of the major reasons for the Mamlūk decline had been Portuguese discoveries in India and the establishment of an all-water route around southern Africa in place of that through the Middle East. It now remained for the Ottomans to restore the full prosperity of their Arab dominions by countering Portuguese naval activities in the Eastern seas. The Ottoman conquests in the East, combined with the Ṣafavid survival in Iran, ended the long period of political vacuum and anarchy that had followed the collapse of the universal ʿAbbāsid Empire, starting in the 11th century. Order and security finally were re-established, and the stability of Middle Eastern society was restored under the guidance and protection of powerful imperial orders. But the Islāmic world was left permanently divided, with Iran and Transoxania, once centres of the Islāmic caliphates, separated from the Arab world, to which Anatolia and southeastern Europe were for the first time added as integral parts of the Middle East.

Süleyman I (ruled 1520–66). Selim I's last years were spent in Istanbul solidifying the supremacy of the sultan, exploiting the prestige and revenues resulting from his Eastern victories. It was therefore only during the long reign of his son and successor Süleyman I (ruled 1520–66)—called the Magnificent in Europe and the Lawgiver among the Ottomans—that the foundations left by Selim were fully used to establish the classical Ottoman state and society and to make important new conquests in East and West. Süleyman assumed the throne with a position unequalled by any sultan before or after him. He was without opposition and with a great deal of control over the *devşirme* class as well as over the remnants of the Turkish notables. The conquest of the Arab world had doubled the revenues of the treasury without imposing important additional financial obligations, leaving Süleyman with wealth and power unparalleled in Ottoman history. Although Süleyman never took full advantage of the opportunities left him and, in fact, began a process of Ottoman decline, his reign still marked

the peak of Ottoman grandeur, and it has always been regarded as the Golden Age of Ottoman history.

Expansion in Europe

The chief battlefields of Ottoman expansion in Europe now were Hungary and the Mediterranean. The weak southeastern European enemies of Süleyman's predecessors had been replaced by the powerful Habsburg Empire, bolstered by the appeals of the pope against the menace of Islām. Süleyman's main European ally was France, which sought to use Ottoman pressure in the East to lessen the pressure of the Habsburgs.

The land war with the Habsburgs centred in Hungary and was fought in three main stages. From 1520 to 1526 the independent Hungarian kingdom bore the direct brunt of the Ottoman attack and acted as a buffer between the two great empires; but a weak king, Louis II (ruled 1516–26) and feudal anarchy and misrule made a united defense impossible. A split among Hungarian nobles over the question of accepting Habsburg rule, combined with the social and national divisions stimulated by the Reformation, further weakened the opposition to Ottoman attack, and, as a result, Süleyman was able to take Belgrade in August 1521, opening the way for a large-scale advance north of the Danube. The only real army the Hungarian notables could muster was routed at the Battle of Mohács (August 29, 1526), and the death of Louis II ended the last hope for Hungarian unity and independence.

The second period of Ottoman-Habsburg relations (1526–41) was characterized by Hungarian autonomy under the anti-Habsburg prince of Transylvania, John Zápolya (ruled 1528–40), who accepted the suzerainty of the sultan in return for the right to continue native administration and military defense. The Habsburg prince Ferdinand, brother of the Emperor Charles V (ruled 1519–58), occupied the northern areas of Hungary with the support of the Hungarian nobles who desired Habsburg aid against the Turks, and for all practical purposes he annexed them to Austria before, in 1527–28, undertaking an effort to conquer the remainder of Hungary. In response Süleyman returned from Anatolia; drove the Habsburgs from all of Hungary; and besieged Vienna (1529), an effort that failed due to the difficulty of supplying a large force so far from the major centres of Ottoman power. Vienna thus was the chief remaining European bulwark against further Muslim advance. Under the existing conditions of supply, transport, and military organization, the Ottomans had reached the limit of their possible expansion in the West from a winter base that had to be maintained in Istanbul because of the constant threat of possible military needs in the East.

Siege of Vienna (1529)

The siege of Vienna had important benefits for the Ottomans—it secured Süleyman's rule of Hungary, and it prevented Ferdinand from launching a new attack against Zápolya for some time to come. Although the siege frightened the other states of Europe sufficiently for them to agree to a temporary Catholic-Protestant truce (1532), the result was only temporary, and Ferdinand never was certain of the support of the independent German princes and of other European rulers who promised him help. Even Charles V was too preoccupied with the problems of the Reformation and with France to devote much attention to the Ottomans. Thus, when Süleyman went on a second Austrian campaign (1532), he was unable to draw the imperial army into conflict and had to content himself with devastating large parts of the Habsburg realm.

By the peace of 1533 Ferdinand abandoned his claims to central Hungary and recognized Zápolya's rule there as Ottoman vassal, whereas Süleyman agreed to accept Ferdinand as ruler of northern Hungary in return for the payment of an annual tribute. This arrangement lasted until 1540, when Zápolya died and left his dominions to Ferdinand in defiance of his agreement with the sultan. When Ferdinand tried to assume his heritage by force, Süleyman occupied and annexed Hungary (August 1541) under the guise of championing the cause of Zápolya's infant son, John Sigismund Zápolya. Thus began the third and final period of Ottoman-Habsburg relations, with the two great powers in direct contact and almost continuous border conflict; diversions on both sides, however, prevented long periods of open warfare.

Many historians have accused Francis I of France (ruled 1515–47) of encouraging Ottoman expansion into central Europe to relieve Habsburg pressure on him. But the Ottoman advances were due less to any French overtures than to Süleyman's own ambitions, together with his fears of a possible alliance among the Habsburgs, the Hungarians, and the Ṣafavids and of Habsburg rule in Hungary. The sultan regarded the French king largely as a supplicant for commercial favours, which were granted in the Capitulations of 1536—an agreement by which French subjects were given the freedom to travel and trade in the sultan's dominions, and subjects of other states wishing to do the same were required to secure French protection as a condition of the necessary permission. French and other merchants and travellers in the Ottoman Empire were allowed to remain under French laws and courts in cases concerning themselves and to have special privileges when involved with Ottoman law. Thus was established the foundation of the French predominance in the Levant, which remained to modern times. The Capitulations served as a model for later trade agreements between the Ottomans and the other European powers, who subsequently used them, during the centuries of Ottoman weakness, as means to dominate commerce within the Ottoman dominions and thus prevent the rise of a native commercial class.

The stalemate between the Ottomans and Habsburgs in Hungary led their conflict to shift periodically from the land to the sea, with the Ottomans emerging as a major naval power for the first time. The decline of the Venetian navy led Charles V to try for complete control of the Mediterranean, enlisting as his naval commander a great Genoese seaman, Andrea Doria, and thus securing the support of the powerful Genoese fleet. Süleyman responded by driving the Knights Hospitallers—a Christian religious and military order—from Rhodes (1522), but Charles then established them on Malta (1530) and captured Tunis (1535). While Süleyman was busy in Anatolia, Doria captured a number of ports in the Morea (Peloponnese) and began to raid the Ottoman coasts, largely severing the sea lines of communication between Istanbul and Alexandria. To counter this, Süleyman in 1533 enrolled in his service as grand admiral Barbarossa (Khayr ad-Dīn), a Turkish captain who had built a major pirate fleet in the western Mediterranean and had used it to capture Algiers (1529) and other North African ports. The Ottomans annexed Algiers to the empire, but as a special province permanently assigned to the grand admiral to support the fleet. Ottoman land troops were sent to defend Algiers against Habsburg attacks—probably the main reason Barbarossa agreed to join the sultan.

The Mediterranean front

Barbarossa built a powerful Ottoman fleet, able to meet the Habsburgs on equal terms. In 1537 he launched a major attack on southern Italy, expecting a promised French attack in the north, with the objective of a joint conquest of Italy. But France, fearing a hostile European reaction to its alliance with the infidel, withheld the diversion. Doria then organized and led an allied European naval force against the Ottomans; but it was routed at the Battle of Preveza (September 25–28, 1538), off the Albanian coast. Venice then surrendered its last possessions in the Aegean, the Morea, and Dalmatia, thus assuring Ottoman naval supremacy in the eastern Mediterranean, which remained unbroken for the next three decades.

Süleyman failed to pursue his ambitions in Europe after 1541, largely because of his increasing preoccupation with problems in the East. He ruthlessly suppressed Ṣafavid propagandists and supporters in eastern Anatolia and stimulated the Turkmen Özbek Empire of Transoxania to attack Iran from the East. Iran fell into disorder following the death of Shāh Esmāʿīl and the accession of his infant son Ṭahmāsp I (ruled 1524–76), but Süleyman was able to use this situation only during periods of peace in Europe. He personally led three campaigns into northwestern Iran—in 1534–35, 1548–50, and 1554—but although he captured Ṣafavid territories in the southern Caucasus, Azerbaijan, and Iraq on each occasion, he

The eastern front

never was able to catch the Iranian army to defeat it; and supply problems invariably compelled him to retire to Anatolia during the winter months, allowing the Persians to regain their territories with little difficulty. Süleyman finally despaired of defeating his elusive enemies, so he agreed to the Peace of Amasya (May 29, 1555), by which he retained Iran and eastern Anatolia but renounced Ottoman claims to Azerbaijan and the Caucasus and agreed to allow Shīʿah Persian pilgrims to visit Mecca and Medina as well as their own holy places in Iraq. Thus the same geographical problems that had limited Ottoman conquests in central Europe made western Azerbaijan the viable limit of Ottoman expansion in the East; preventing the final elimination of the Ṣafavid danger.

Süleyman was somewhat more successful in restoring the old international trade routes through his Middle East possessions. To counteract the Portuguese fleet, supplied by the Ṣafavids from their Persian Gulf ports, he built major naval bases—at Suez (1517) and, as soon as he took Iraq, at Basra (1538)—establishing garrisons and fleets that not only resisted the Portuguese naval attacks but also went out against them in the Eastern seas. As a result the old trade route regained some of its former volume in the 16th century; the Ottomans never were able to fully restore it, however, because Portugal still was able to pay higher prices in the East and sell at lower prices in Europe because of their use of a sea route, which avoided the duties and local charges levied on goods sent through Ottoman territory. (It should be noted that it was the Ottomans who fought to keep the old Middle Eastern trade route open; the route was closed only when the Cape route was taken over from the Portuguese by the much more powerful fleets of Great Britain and Holland.)

Classical Ottoman society and administration. During the 16th century the institutions of society and government that had been evolving in the Ottoman dominions for two centuries reached the classical forms and patterns that were to characterize them until modern times.

Social divisions

The basic division in Ottoman society was the traditional Middle Eastern distinction between the small group of rulers who formed the ruling class, and the large mass of subjects. Possession of three attributes was essential for membership in the Ottoman ruling class: (1) profession of loyalty to the sultan and his state; (2) acceptance and practice of the Muslim religion and the system of thought and action that was an integral part of it; and (3) knowledge and practice of the complicated system of customs, behaviour, and language known as the Ottoman Way. Those who lacked any of these attributes were considered to be members of the subject class, *reayas* (*rayahs*), the "protected flock" of the sultan. There was a system of social mobility based on the possession of these definable and attainable attributes; *reayas* able to acquire them could rise into the ruling class, and Ottomans who came to lack any of them would become members of the subject class.

Members of the Ottoman ruling class were considered to be the sultan's slaves, and thus acquired their master's social status. As slaves, however, their properties, lives, and persons were entirely at his disposition; and he could do with them as he wished. Their basic function was to preserve the Islāmic nature of the state and to rule and defend the empire. By Ottoman theory, the main attribute of the sultan's sovereignty was the right to possess all sources of wealth in the empire together with the authority necessary to exploit them. The function of enlarging, protecting, and exploiting that wealth for the benefit of the sultan and his state, therefore, was the main duty of the ruling class. The primary duty of the *reayas* was to produce the wealth—by farming the land or engaging in trade and industry—and then paying a part of the resulting profits to the ruling class in the form of taxes.

The Ottoman state encompassed organizations and hierarchies developed by the ruling and subject classes to carry out their functions in Ottoman society. The ruling class divided itself into four functional institutions: the Imperial (Mülkiye) Institution, led by the Sultan himself, provided the leadership and direction for the other institutions as well as for the entire Ottoman system; the Military (Seyfiye) Institution was in charge of expanding and defending the empire and keeping order and security within the sultan's dominions; the Administrative (Kalemiye) Institution, organized as the Imperial Treasury (Hazine-i Amire), was in charge of collecting and spending the imperial revenues; and the Religious/Cultural (İlmiye) Institution, which included the *ʿulamāʾ* (all Ottomans expert in the religious sciences), was in charge of organizing and propagating the faith and maintaining and enforcing the religious law (Sharīʿah), its interpretation in the courts, its expounding in the mosques and schools, and its study and interpretation.

To cover the areas of life not included within the scope of the ruling class, members of the subject class were allowed to organize themselves as they wished. As a natural manifestation of Islāmic society their organization was determined largely by religious and occupational distinctions. The basic class divisions within the subject class were determined by religion, with each important group organizing into a relatively self-centred autonomous community called a *millet*, under its own laws and forms of internal organization, directed by a religious leader who was responsible to the sultan for the fulfillment of the duties and responsibilities of the *millet* members, particularly those of paying taxes and security. In addition, each *millet* cared for the many social and administrative functions not assumed by the Ottoman state, concerning such matters as marriage, divorce, birth and death, health, education, internal security, and justice.

The *millets*

Within the *millets*, just as in Ottoman society as a whole, there was a social mobility, with persons moving up and down the ladder according to ability and luck. Individuals could pass from one *millet* to another if they wished to convert, but because all the *millets* were extremely antagonistic toward those who left them to convert to another religion, the state discouraged such action as much as possible to preserve social harmony and tranquility, the main object of the system. The *millet* system succeeded for 500 years by keeping the different peoples of the empire as much apart as possible, thus reducing to a minimum the possible sources of conflict and keeping social order in a highly heterogeneous state.

All of the classes, institutions, and communities described above were, in essence, means by which the slaves and subjects of the sultan were divided and organized so as best to fulfill their functions. But there also were means by which they were related to one another and united into the whole of Ottoman society. The principal cement was the sultan, the keystone of the system who alone was the common focus of loyalty of the ruling and ruled subjects alike. But such loyalty was an abstract; it could not have held Ottoman society together without the help of the artisan and religious guilds, which cut across the boundaries and made members of different groups brothers in common organizations based not on class, rank, or religion but on mutually shared values and beliefs, economic activities, and social needs. Through contact and cooperation in such guilds, members of the different groups of Ottoman society were cemented into a common whole, performing many of the social and economic functions outside the scope of the ruling class and of the *millets*, particularly those functions associated with economic regulation and social security, and receiving a more personal religious experience than that provided by the established religious organizations, whether Muslim or non-Muslim.

Within the Ottoman ruling class, the most important unit of organization and action was the *mukataʿa*, in which a member of the ruling class "cut off" a portion of the sultan's revenues and had sufficient authority to exploit it for purposes determined by the sultan. The exact nature of the *mukataʿa* depended on the extent to which the holder remitted his revenues to the treasury and what portion he retained for himself. The *timar* type of *mukataʿa*, traditionally described as a fief, only marginally approached the concept of feudalism as it was known in Europe because it was part of a centralized system and did not involve the kind of mutual rights and obligations

The *mukataʿa*

characteristic of Western feudalism. In return for services to the state, the *timar* holder was given the full profits of the source of revenue for his personal exploitation and profit, which were independent of, and in addition to, those connected with the exploitation of the *timar* itself. For many military and administrative positions, *timar*s normally were given in lieu of salaries, thus relieving the treasury of the trouble and expense of collecting revenues and disbursing them to its employees as salaries. Almost all the 14th- and 15th-century Ottoman conquests in southeastern Europe were distributed as *timar*s to military officers, who in return performed the tasks of administration in peacetime and provided soldiers and military leadership for the Ottoman army in war. Many of the officers of the central government also were rewarded with *timar*s in place of, or in addition to, salaries paid by the treasury.

The second principal form of the *mukataʿa* was the *emânet* (trusteeship), held by the *emin* (trustee, agent). In contrast to the situation of the *timar* holder, the *emin* turned all his proceeds over to the treasury and was compensated entirely by salary, thus being the closest Ottoman equivalent to the modern government official. The legal basis for this arrangement was that the *emin* did no more than administer the *mukataʿa;* he undertook no additional service, and so had no right to share in the profits of the *mukataʿa* that he held. This was the least common type of administrative position in Ottoman times, for the most part used for urban custom houses and market police, which were very close to, and under the supervision and control of, the central government and its agents, and which did not need the profit motive in order to assure efficiency on the part of the holders.

The most common kind of *mukataʿa,* and therefore the most prevalent type of administrative unit in the Ottoman system, was the tax farm (*iltizam*), which combined elements of both the *timar* and *emânet.* As in the former, the tax farmer (*mültezim*) could keep a part—but only a part—of the tax he collected, and, like the *emin,* had to deliver the balance to the treasury. This was because his service consisted only of his work in administering the *mukataʿa,* for which he was given a share of his collection instead of the *emin*'s salary. The tax farmer thus was given the inducement of profit to be as efficient as possible. Most of Anatolia and the Arab provinces were administered in this way because they were conquered at a time when the government's need for cash to pay the salaried Janissary infantry and to supply an increasingly lavish court required the treasury to seek out all the revenues it could find. As the *timar*-based *sipahi* cavalry became less important, and as the Turkish notables who held most of the *timar*s lost most of their political power during the time of Süleyman, the estates gradually fell into the hands of the *devşirme* class.

The dual system of law

The legal and customary bases of organization and action in Ottoman society depended on a dual system of law —the Sharīʿah or religious law, and the *kanun,* or civil law. The former, the basic law of Ottoman society, as it was of all Muslim communities, was considered to be a divinely inspired corpus of political, social, and moral regulations and principles, which were supposed to cover all aspects of life for Muslims. But it was highly developed only in the fields of personal behaviour covered in detail in the early Muslim community and reflected in the Qurʾān and early Muslim tradition. It never was developed in detail in matters of public law, state organization, and administration. Its general principles left room for interpretation and legislation on specific matters by secular authorities; and the Muslim judges of the Ottoman Empire recognized the right of the sultan to legislate in civil laws, so long as he did not conflict with the Sharīʿah in detail or principle. The Sharīʿah, therefore, provided the principles of public law, and covered matters of personal behaviour and status in the Muslim *millet* in the same way that the members of the Christian and Jewish *millet*s were subject to their own religious codes. The Sharīʿah was interpreted and enforced by members of the Cultural Institution—the *ʿulamāʾ*—just as the laws of each non-Muslim *millet* were enforced by its leaders. Strictly speaking, the *ʿulamāʾ* members had the right to invalidate any law they felt contradicted the Sharīʿah, but they rarely did this because, as part of the ruling class, they were under the authority of the sultan and could be removed from their positions. The sultan therefore was relatively free to legislate changes in Ottoman institutions and practices to meet the needs of the time; this was a major factor in the long survival of the empire, even during centuries of decline. It must be noted, however, that with the restricted scope of the Ottoman ruling class and state and the large areas of power and function left to the religious communities and guilds as well as to the Ottoman officials who held the different kinds of *mukataʿa*s, the sultans were never as autocratic as has been commonly assumed. It was only in the 19th century that Ottoman reformers centralized government and society on Western lines and restricted or ended the traditional autonomies, which had done so much to decentralize power in the previous centuries.

DECLINE OF THE OTTOMAN EMPIRE, 1566–1807

Internal problems. The reign of Süleyman the Magnificent marked the peak of Ottoman grandeur, but elements of weakness crept in and began the slow but steady decline that followed. An important factor of decline was the increasing lack of ability and power of the sultans themselves. Süleyman tired of the long military campaigns and arduous duties of administration concentrated in his person, and withdrew more and more from public affairs to devote himself to the pleasure of his harem. To take his place, the office of grand vizier was built up to become second only to the sultan in authority and revenue, including the right to demand and obtain absolute obedience. But while the grand vizier was, indeed, able to replace the sultan in official functions, he could not take his place as the focus of loyalty for all the different classes and groups in the empire. The resulting separation of political loyalty and central authority led to a decline in the government's ability to impose its will.

Decline of the sultans

Triumph of the devşirme. The mid-16th century also saw the triumph of the *devşirme* over the Turkish nobility, which lost almost all of its power and positions in the capital and returned to its old centres of power in southeastern Europe and Anatolia. In consequence, many of the *timar*s formerly assigned to the notables to support the *sipahi* cavalry were seized by the *devşirme* and transformed into great estates, for all practical purposes as private property, thus depriving the state of their services as well as of the revenue they could have produced if they had been transformed into tax farms. Whereas the *sipahi*s did not entirely disappear as a military force, the Janissaries and the associated artillery corps became the most important segments of the Ottoman army.

Corruption and nepotism. Because the sultans no longer could control the *devşirme* by playing the Turkish notables off against them, the *devşirme* gained control of the sultans and used the government for its own benefit rather than for that of the sultan or his empire. In consequence, the evils of corruption and nepotism took hold at all levels of administration. In addition, with the challenge of the notables gone, the *devşirme* class itself broke into countless factions and parties, each working for its own advantage by supporting the candidacy of one or another of the imperial princes and each in close alliance with corresponding palace factions led by the mothers, sisters, and wives of each prince. Following Süleyman, therefore, accession and appointments to positions came less as the result of ability than as a consequence of the political manoeuvrings of the *devşirme*-harem political parties. Those in power found it more convenient to control the princes by keeping them uneducated and inexperienced, and the old tradition by which young princes were educated in the field, was replaced by a system in which all the princes were isolated in the private apartments of the harem and limited to the education its permanent inhabitants could provide. In consequence, few of the sultans after Süleyman had the ability to exercise real power even when circumstances might have enabled them to do so.

But the lack of ability did not end the sultans' desire for power; because they lacked the devices developed by their predecessors to achieve this end, they developed new ones. Selim II (ruled 1566–74), known as the Sot, and Murad III (ruled 1574–95) both gained power by playing off the different factions and by weakening the office of grand vizier, the main administrative vehicle for factional and party influence in the declining Ottoman state. As the grand viziers lost their dominant position following the downfall of Mehmed Sokollu (served 1560–79), power fell first into the hands of the women of the harem, during the "Sultanate of the Women" (1570–78), and then into the grasp of the chief Janissary officers, the *ağa*s who dominated from 1578 to 1625. But no matter who controlled the apparatus of government during this time, the results were the same—a growing paralysis of administration throughout the empire and a pulling apart of the different groups into separate and hostile communities.

Economic difficulties. Under such conditions, it was inevitable that the Ottoman government could not meet the increasingly difficult problems that rose to plague the empire in the 16th and 17th centuries. Economic difficulties began in the late 16th century when the Dutch and British were able to completely close the old international trade routes through the Middle East. As a result, the prosperity of the Arab provinces declined and, in addition, the Ottoman economy was upset by inflation, started by the influx of precious metals into Europe from the Americas and by an increasing imbalance of trade between East and West. As the treasury lost more of its revenues to the depredations of the *devşirme*, it began to meet its obligations by debasing the coinage, heavily increasing taxes, and resorting to confiscations—all of which further worsened the situation. All those persons depending on salaries found themselves underpaid, and the result was further theft and corruption. Holders of the *timar*s and tax farms started using them as sources of revenue to be milked as rapidly as possible, rather than as long-term holdings whose prosperity had to be maintained to provide for the future; political influence and corruption also enabled them to transform these holdings into private property, either as life holdings (*malikâne*) or as religious endowments (*vakf*), without any further obligations to the state. Inflation also hit the traditional industries and trades. Functioning under strict price regulations, the guilds were unable to provide quality goods at prices low enough to compete with the cheap European manufactured goods that entered the empire without restriction due to the capitulations agreements. In consequence, traditional Ottoman industry fell into rapid decline.

The decline of Ottoman industry

Social erosion. These conditions were exacerbated by a large increase of population during the 16th and 17th centuries—part of the general population rise that occurred in much of Europe at this time. The amount of subsistance available was not only unable to expand to meet the needs of the rising population, but in fact fell as the result of the political and economic conditions then prevalent in the Ottoman Empire; social distress increased, and upset resulted. Landless and jobless peasants fled from the lands as did cultivators subjected to confiscatory taxation at the hands of timariots and tax collectors, thus reducing food supplies even more. Many of them fled to the cities, where they added to the problem of food supply and reacted against their troubles by rising against the established order; many more remained in the countryside and joined rebel bands, known as *levend*s and Jelālīs (Celâlis), which took what they could from those who remained to cultivate and trade.

The central government became weaker; and as more peasants joined rebel bands, they were able to take over large parts of the empire, keeping all the remaining tax revenues for themselves and often cutting off the regular food supplies of the cities as well as of the Ottoman armies still guarding the frontiers. Under such conditions, the armies themselves broke up, with most of the salaried positions in the Janissary corps and other corps falling into the hands of urban investors, who used them simply as new sources of revenue, without performing any military services in return. Thus it was that the Ottoman armies came to be composed primarily of fighting contingents supplied by the vassals of the sultan (particularly the Crimean Tartar khans), together with whatever rabble could be dragged from the streets of the cities whenever required by campaigns. The Ottoman army still remained strong enough to curb the most pressing provincial revolts; but the latter proliferated through the centuries of decline, making effective administration outside the major cities still under the government's control almost impossible. In many ways the substratum of Ottoman society—formed by the *millet*s and various economic, social and religious guilds and buttressed by the organization of the Ottoman *'ulamā'*—cushioned the mass of the people and the ruling class itself from the worst effects of this multisided disintegration.

External relations. Despite these difficulties, the internal Ottoman weakness was evident to only the most discerning Ottoman and foreign observers during much of the 17th century. Most Europeans continued to fear the Ottoman army much as it had been feared two centuries before; and whereas its ability was reduced, it remained strong enough not only to prevent the provincial rebels from assuming complete control but also to make a few more significant conquests in both East and West. Whereas the empire now suffered defeats for the first time, it still retained reserve strength sufficient for it to recoup when needed and to prevent the loss of any integral parts of the empire. Although the Ottoman navy was destroyed by the fleet of the Holy League at the Battle of Lepanto (1571), it was yet able to rebuild and regain naval mastery in the eastern Mediterranean through most of the 17th century, taking Tunis from the Spanish Habsburgs (1574), Fez from the Portuguese (1578), and Crete from Venice (1669). In consequence, so long as Europe continued to fear the Ottomans, no one tried to upset the precarious peace treaties concluded in Süleyman's later years; and the Ottomans were shielded from the results of their own weakness.

Superiority of Ottoman navy

Military campaigns. Despite the upsets then disturbing the Ottoman body politic, the Porte undertook new campaigns. When the rising Principality of Moscow conquered the last Mongol states in Central Asia and reached the Caspian, thus posing a threat to the Ottoman position north of the Black Sea and in the Caucasus, Murad II conquered the Caucasus and, taking advantage of anarchy in Iran following the death of Shāh Ṭahmāsp I in 1576, seized the long-coveted Azerbaijan. He thus brought the empire to the peak of its territorial extent and added wealthy new provinces whose revenues, for a half century at least, rescued the Ottoman treasury from the worst of its financial troubles and gave the empire a respite in which it could at least try to remedy its worst problems.

Reform efforts in the 17th century were undertaken by sultans Osman II (ruled 1618–22) and Murad IV (ruled 1623–40) and by a famous dynasty of Köprülü grand viziers who served under Sultan Mehmed IV (ruled 1648–1687)—Mehmed Köprülü (served 1656–61) and Ahmed Köprülü (served 1661–76). Each of these early reformers rose as the result of crises and military defeats that threatened the very existence of the empire. Each was given the power needed to introduce reforms because of the fears of the ruling class that the empire, on which its privileges depended, was in mortal danger.

Exposure of Ottoman weakness. In a war between the Ottomans and the Habsburgs (began 1593) the Austrians were able to take much of central Hungary and Romania, and only an accidental Ottoman triumph in 1596 enabled the Sultan to sufficiently recoup so that the Habsburgs agreed to the Treaty of Zsitvatorok (1606), by which agreement Ottoman rule of Hungary and Romania was restored. The treaty itself, however, like the events that led up to it, for the first time showed Europe the extent of Ottoman weakness and so exposed the Ottomans to considerable new dangers in subsequent years. In the East, anarchy in Iran was brought to an end by Shāh 'Abbās I (ruled 1587–1629), who not only restored Iranian power but drove the Ottomans out of Azerbaijan and the Cauca-

sus (1603), conquered Iraq (1624), and threatened to take the entire Ottoman Empire. Though Murad IV was able to retake Iraq (1638), Iran remained a major threat. Finally, a long war with Venice (1645–69), occasioned by Ottoman efforts to capture Crete, exposed Istanbul to a major Venetian naval attack. Although finally pushed back in a naval campaign culminating in the Ottoman conquest of Crete (1669), the Venetians still posed a major threat, which, like those that had occurred earlier in the century, stimulated the ruling class to accept needed reforms.

Reforms. The Ottoman reforms introduced during the 17th century were too limited in nature and scope to permanently arrest the Ottoman decline. Basically, they were no more than efforts to restore to the state the inherited system of government and society that had operated successfully in the past. Corrupt officials were executed. Efforts were made to restore the *timar* and tax-farm systems as the basis of the administration and army. Provincial revolts were suppressed, peasants were forced back to the lands, and cultivation was increased. Debased coins were replaced by coins of full face value. Industry and trade were encouraged, and corruption and insubordination were driven out.

Such reforms were sufficient to end the immediate difficulties. But they were successful only for the moment because the reformers were allowed only to act against the results of the decay and not its cause—the selfish rule of the ruling class—which continued. As soon as the worst consequences of decay had been alleviated, the old groups returned to power and resumed their old ways.

Failure to assess new Europe

Moreover, the reformers really did not understand that the Europe now faced by the Ottomans was far more powerful than that which the great sultans of the past had defeated; thus even if the reforms had been more permanently successful, they could not have corrected the increasing Ottoman weakness relative to the powerful nation-states then rising in Europe. Such an understanding was to come to the Ottoman reformers only in the 19th century.

Military defeats, 1683–1792. The traditional 17th-century reforms did, however, produce at least a semblance of revival. By 1681 the Ottoman army seemed so strong that Grand Vizier Kara Mustafa Paşa (served 1676–83), brother-in-law of Ahmed Köprülü, was emboldened to move again into central Europe and besiege Vienna (July–September 1683). His effort quickly overextended the fragile bases of the Ottoman revival.

Anti-Ottoman coalition. The aroused defenders, stimulated by the Polish king Jan Sobieski (ruled 1674–96), not only held out but also built a major European coalition that moved to bring destruction to the Ottoman Empire during the subsequent century: the Habsburgs set out to reconquer Hungary, Serbia, and the Balkans; Venice hoped to regain its naval bases on the Adriatic and in the Morea and to resume its naval and commercial power in the Levant; Russia worked to extend its lands to the open seas. Only the enemies of the coalition in Europe, led by France and Sweden, tried to support the Ottomans; neutral Britain and Holland, to guard the commercial privileges they had secured from the sultan through the Capitulations, worked to prevent any nation from gaining control of the entire Ottoman Empire and from becoming, thereby, preponderant in Europe.

Russia and Austria fought the Ottomans not only by direct military attack but also by fomenting dissatisfaction and revolt on the part of the non-Muslim subjects of the Sultan. Against such an attack, the Ottomans could only conciliate their subjects where possible and repress them when conciliation was rejected, taking advantage, at every opportunity, of every rivalry that arose between the Habsburgs and Russians for predominance in the Balkan provinces of the empire.

European wars. In consequence of this situation, the Ottoman Empire was at war with European enemies for 41 years between the second siege of Vienna (1683) and the Treaty of Jassy (1792). From 1683 to 1699 it fought the armies of the Holy League in a disastrous war culminating in the Treaty of Carlowitz (1699). In 1710 and 1711 it fought Russia again and at the Treaty of the Pruth (1711) regained some territories previously lost. The war of 1714–18 with Venice and Austria was concluded by the Treaty of Passarowitz (1718); and three wars with Russia and Austria—in 1736–39, 1768–74, and 1787–92—culminated in the treaties of Belgrade (1739), Küçük Kaynarca (1774), and Jassy (1792). As a result of these wars, the Ottomans lost Hungary, the Banat of Temisvar, Transylvania, and Bukovina in Europe, establishing their boundary on the Danube, where it had last been early in the 16th century. To Russia they lost all their possessions on the northern coasts of the Black Sea from the principalities to the Caucasus, including Bessarabia, Podolya, and the Crimea, the soldiers of which had provided the strongest element in the Ottoman army during the previous century. In addition, the Ottomans were compelled to allow the Russians and Austrians to intervene legally on behalf of the sultan's Christian subjects in a manner that opened the way to an increased European influence in internal Ottoman affairs.

Loss of territory

Imperial decline in the 18th and early 19th centuries. The manifestations of decline were only continuations and elaborations of earlier conditions. But a new factor of decline was added—the weakness of the central government resulted in the loss of control of most of the provinces to local rulers, called notables (*a'yan*), who took more or less permanent control of large areas, creating a situation that in many ways resembled European feudalism much more than the traditional Ottoman *timar* system ever did.

Rise of local rulers. These notables were able to build up their power and maintain control not only because the Sultan's government lacked the military resources to suppress them but also because the local populations themselves preferred the notables' rule to that of the corrupt and incompetent Ottoman officials of the time. In the Balkans and Anatolia, local rulers solidified their positions by taking advantage of currents of local nationalism that were arising among the Balkan Christians and the Muslim Turks. The notables formed private armies of mercenaries and slaves, with which they sometimes provided important contributions to the Ottoman armies in return for recognition of their autonomy by the sultans. These rulers were able to exercise almost complete authority, collecting taxes for themselves and sending only nominal payments to the treasury, thus further increasing its problems. The central government maintained its position by playing off the local rebels against each other, using the leverage of Ottoman support to its own advantage, and securing considerable payments of cash and military contributions when needed. The treasury, therefore, did not suffer as much from these provincial revolts as might be imagined; but the revolts did disrupt the established food supplies of the empire sufficiently for large-scale famines to arise in the major cities on a regular basis. In response, the urban populace became a restless, misruled, and anarchic mass that broke loose at the slightest provocation, responding to unemployment, famine, the plague, and the like with riots and summary executions of the officials considered responsible. Such violence, while manifesting Ottoman difficulties, did not remedy them and, in fact, made things worse. Remedy lay only in the hands of the ruling class; but its reaction was quite different.

Urban anarchy

Resistance to change. Most Ottomans saw little need for the empire to change because they actually benefitted financially from the existing anarchy and lack of control by the sultan. In addition, the ruling class was completely isolated from developments outside its own sphere; it assumed that the remedies to Ottoman decline lay entirely within Ottoman practice and experience. This resulted from the basic belief of Ottoman society in its own superiority over anything the infidel could possibly produce —a belief that had far more basis in the 16th century, when it was developed, than in the 18th century. All of the developments in industrial and commercial life, in science and technology, and particularly in political and military organization and techniques that had occurred in Europe since the Reformation were simply unknown to

the Ottomans. The only direct Ottoman contacts with Europe were on the battlefield, where most Ottomans still assumed that their military reverses were caused not by the superiority of Western armies as such but rather by Ottoman failure to apply fully the techniques that had worked so well in the past. So the 18th-century reforms largely paralleled those of the traditional Ottoman reformers of the 17th century, with only occasional efforts to add new military organizations and to make use of specific European weapons and techniques of undeniable superiority.

Contacts with the West. For a few Ottomans this isolation was at least partially broken down when some channels of contact opened with the West during the 18th century—a few Ottoman ambassadors went to Europe to participate in negotiations and sign treaties; more and more European merchants, travellers, and consuls came into the Ottoman Empire; a very few Ottoman men of science and philosophy began to correspond with their Western counterparts; members of the Ottoman minorities entered into correspondence with their relatives in the West. But such contacts had very limited effects; only a small number of Ottomans experienced them, and even when they did learn something, the effect was quite superficial because the resulting information did not fit into the patterns of thought of even the most educated Ottomans. Those few who did understand something of what they heard usually were only voices in the wilderness, and their efforts to apply and spread what they knew had little overall effect.

Such contacts did lead to changes in the modes of living of a few upper class Ottomans and to some military innovations, but to nothing more. Starting in the so-called Tulip Period (1717–30; see below) some Ottomans under the influence of Grand Vizier İbrahim Paşa began to dress like Europeans, and the palace began to imitate European court life and pleasures. Sultan Ahmed III (ruled 1703–30) built several lavish summer residences on the Bosporus and the Golden Horn (an inlet that forms the harbour of Istanbul), and these were imitated by members of his immediate entourage, who held frequent garden parties in imitation of the pleasures of Versailles. The sultan and his ministers were no longer confined behind the walls of the palace; and the new era was celebrated by Nedim, the court poet, who reflected a considerable awareness of his environment and an appreciation of nature. Growing tulips, as a mark of westernization, became an obsession with rich and poor alike, and the flower gave its name to the period. In 1727 Turkish language books were printed for the first time in the empire by a Hungarian convert named İbrahim Müteferrika; and while the press was closed by the government at times, during the remainder of the century it provided a number of books on history and geography that further opened the minds of those who saw and read them.

The Tulip Period

Military reforms. As a result of contact with European armies and the influence of European renegades in Ottoman service, a few attempts were made during the century to adopt Western-type uniforms, weapons, and tactics. Because the members of the established military corps could not and would not surrender their old ways, entirely new corps were formed to handle the new weapons under the direction of European instructors. But the new corps had no effect at all on the Janissaries and the other older corps that continued to form the bulk of the army; the latter accurately saw that the new ways were threats to their privileges and security. The new corps were, therefore, no more than special mercenary bodies built up under the stimulus of individual Ottomans, lasting only so long as the latter remained in power. The most successful and lasting Ottoman military reform during this time came in the navy, which was modernized by Grand Admiral Gazi Hasan Paşa (served 1770–89), with the support and encouragement of Sultan Abdülhamid I (ruled 1774–89); this success came largely because the Ottoman naval establishment was wiped out at the Battle of Çeşme (1770) by a Russian fleet, and there was no such inbred resistance as that which stifled significant reforms elsewhere. Important reforms introduced into the army under Grand Vizier Halil Hamid Paşa (served 1782–85), with the help of Western technicians, were limited to new corps especially created for the purpose, while the bulk of the Ottoman army continued to remain without change.

Selim III (ruled 1789–1807). These 18th-century reform efforts culminated during the reign of Sultan Selim III, who is thought by some to have been the originator of modern reform in the Ottoman Empire. While still a prince, Selim developed plans for modernizing the Ottoman army. He came to the throne during the 1787–92 war with Austria and Russia and had to postpone serious reform efforts until its completion. Early efforts to modernize the Janissary corps created such opposition that Selim thereafter concentrated on creating a new European-style army, called the *nizam-ı cedid* ("new order"), using the modern weapons and tactics developed in Europe. This new force, never numbering more than 10,000 active soldiers, was trained in Istanbul and in certain Anatolian provincial centres by officers and military experts sent by the different European powers that were competing for the sultan's support at the time. So as not to disturb the established Ottoman institutions, it was financed by an entirely new treasury, called the *irad-i cedid* ("new revenue"), whose revenues came from taxes imposed on previously untaxed revenue sources and from the confiscation of some *timar*s whose holders were not fulfilling their military and administrative duties to the state. Under the guidance of European technicians, factories were constructed to manufacture modern weapons and ammunitions, and technical schools were opened to train Ottoman officers. Limited efforts also were made to rationalize the Ottoman administrative machinery, but largely along traditional lines.

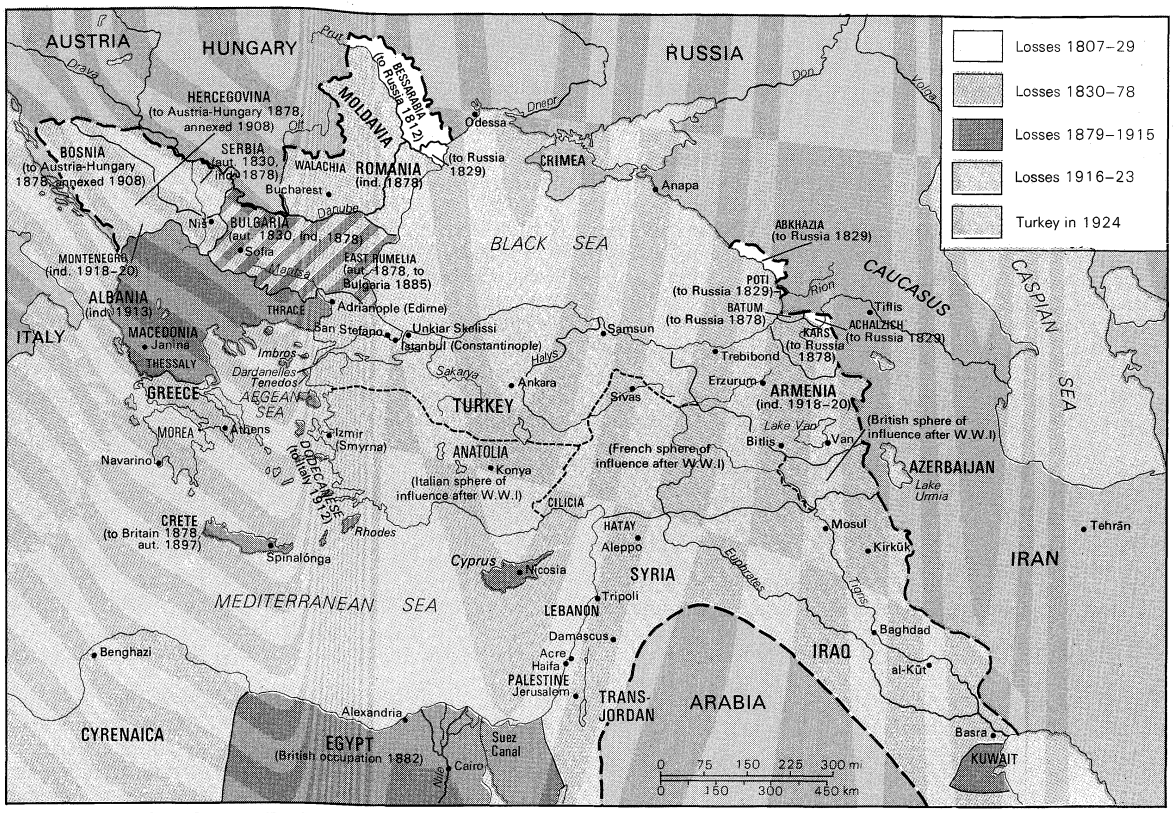
Modern reform

The older military corps, however, remained intact and hostile to the new force, and Selim was therefore compelled to limit its size and use. At the same time, much of his energy was diverted by the rise of powerful autonomous notables in southeastern Europe, Anatolia, and the Arab provinces, as well as by a French Expedition to Egypt (1798–1801), which eventually drew him into alliances with Great Britain and Russia. The rise of nationalism among Ottoman subject peoples stimulated by agents of Russia, Austria, and Revolutionary France, and culminating in the beginning of a Serbian revolution (1804) and a new war with Russia (1806–12) also made it impossible for Selim to resist the wishes of the Janissaries, who still formed the bulk of his army. Finally, the sultan's personal weakness, which led him to desert the reformers and the new army whenever opposition became strong, left him with little significant support in 1807, when he was attacked and overthrown by a conservative coalition. While Selim was imprisoned in the palace, a conservative reaction under Sultan Mustafa IV (1807–08) ended the reforms and massacred most of the reformers. An effort to restore Selim led by the Bulgarian notable Bayrakdar Mustafa Paşa led to Selim's death and, after the short rule of Mustafa IV, the accession of his reforming cousin, Mahmud II (ruled 1808–39). Selim's reforms were largely abandoned for some time. But the greatly increased knowledge of the West in the Ottoman Empire, made possible by the schools established for the *nizam-ı cedid* as well as by the increased numbers of Westerners present in Istanbul during the era of the French Revolution, began the process by which the old Ottoman isolation was finally and definitively broken. This set the stage for more significant reforms, which transformed the empire during the remainder of the 19th century. (S.J.S.)

II. European domination and the establishment of a Turkish national state

THE EMPIRE, FROM 1807 TO 1920

The triumph of the antireform coalition, which had overthrown Selim III, was interrupted in 1808 when the surviving reformers within the higher bureaucracy found support among the *a'yan* (local notables) of Rumelia (Ottoman possessions in the Balkans), who were worried



The dissolution of the Ottoman Empire, 1807–1924.

Adapted from R. Treharne and H. Fullard (eds.), *Muir's Historical Atlas: Ancient, Medieval and Modern*, 9th edition (1965); George Philip & Son Ltd., London

Interruption of conservative rule

by possible threats to their own position. These were led by Bayrakdar (Standard Bearer) Mustafa Paşa. The forces of Mustafa and the grand vizier Çelebi Mustafa Paşa together recovered Istanbul; deposed Mustafa IV; set up Mahmud II, the son of Abdülhamid I; and recommenced some of the reforming policies that had been initiated by Selim.

The *aʿyan* took care to protect their own interests by securing in the Covenant of Union (Sened-i Ittifak) a definition and guarantee of their own rights against the central government. Their victory, however, was short-lived. A further Janissary uprising in November 1808 led to the death of the Bayrakdar and to the re-establishment of conservative rule.

Mahmud II (ruled 1808–39). The Ottoman situation at the end of 1808 appeared desperate. Within the empire the authority of the central government was minimal. Control of North Africa had long since faded. In Egypt, the Ottoman Viceroy Muḥammad ʿAlī (*q.v.*) was laying the foundations for independent power. In Iraq the Georgian Mamlūk pashas paid only lip service to the authority of the Porte, as did various independent local governors in Syria. In Arabia the Wahhābīs mocked Ottoman pretensions. In all Anatolia only two provinces were really under central control, while in the European provinces power had fallen into the hands of such formidable local notables as Ali Paşa, who controlled southern Albania, and Osman Pasvanoğlu, who had dominated northern Bulgaria until his death in 1807. Serbia under the leadership of George Petrović (Karageorge) had been in revolt since 1804; at first the Serbs had risen in desperation against the terrorist policies of the Janissaries, who had usurped the power of the local governor, but they had subsequently demanded autonomy and, in 1807, allied with Russia.

The external threat to the empire was no less ominous. Selim III had hoped to enlist French aid in order to recover territory lost to Russia; as a result, the Ottomans found themselves at war with both Russia, which invaded the Principalities in November 1806, and Britain, which attempted to seize the Dardanelles with a naval force (February 1807) and invaded Egypt (March 1807). Meanwhile Napoleon, through the agreements of Tilsit (July 7, 1807) and Erfurt (October 12, 1808), had not only abandoned active opposition to Russia but had accepted its occupation of the Principalities.

The preoccupation of the European powers with other interests helped the Ottomans mend the international problems. Britain made peace on January 5, 1809 (Peace of the Dardanelles); and Russia, on May 28, 1812 (Treaty of Bucharest), by which the Principalities were returned to Ottoman rule although Russia retained most of Bessarabia.

Internal reform. Mahmud II was then able to concentrate on internal reform. The basic element in Mahmud's reforms was the reconstruction of the army to make it a fit instrument for preserving the Ottoman Empire against both the encroachments of European powers and the centrifugal ambitions of local potentates. This policy brought him into head-on collision with the Janissaries. In 1826 Mahmud set out his proposals for a new European style army; on June 15 the Istanbul Janissaries mutinied in protest and were promptly and efficiently massacred by the sultan—an episode known to later Ottoman historians as "The Auspicious Incident."

Reconstruction of the army

As a tactician, Mahmud proved to be superior to Selim. He had the support of most of the higher *ʿulamāʾ*. Whereas in 1807 the Janissaries had enjoyed the approval of the population of Istanbul, in 1826 only two guilds gave them active help. Mahmud had built up a cooperative group among the Janissary officers and had carefully arranged to have loyal troops at hand. Perhaps most important of all, Mahmud presented his proposals not as dangerous and infidel innovations but as a restoration of the military system of the Ottoman golden age.

The destruction of the old army was completed in 1831 by the final abolition of the feudal (*timar*) system. The remaining fiefs were resumed by the government. Although the new army was dressed, equipped, and trained in the style of European armies and helped by a succession of European advisers (including the future chief of the German General Staff, Helmuth von Moltke), it differed from the former army in its greater loyalty to the

sultan. It thus became an instrument of political centralization, and it provided the major motive for modernization. The continuing need to pay and equip the army and to train its officers and other specialized personnel in a sustained, but ultimately vain, attempt to keep pace with the European powers, stimulated reform of the political and economic institutions of the Ottoman Empire. The modernization of higher education began with the need to train officers, army doctors, and veterinary surgeons; that of the taxation system began with the need to pay the army; that of the administration, with the need to collect the taxes, and so on. Ultimately the entire system of minimal government—by which political, economic, and social decisions were left to local organizations—was replaced by one in which the state came to centralize decisions in its own hands.

Move toward centralization. Mahmud began by curbing the power of rival claimants. He undermined the influence of the *'ulamā'* and of popular religious organizations. He created a new directorate of *evkâf* (charitable endowments) in 1826, hoping to gain control of the hitherto independent financial base of *'ulamā'* power. To make his power more effective, he built new roads and, in 1834, inaugurated a postal service.

Administrative reorganization

The central administration was reorganized. New European-style ministries were created to replace the ancient bottleneck of power caused by the universal administrative responsibility of the grand vizier. New councils were established to assist in long term planning; one—the Supreme Council of Judicial Ordinances (1838)—subsequently became the principal legislative body. Bureaucrats were given greater security by abolishing the practice of confiscating their property at death, while the opening of a translation bureau (1833) and the reopening of embassies abroad gave some the opportunity to learn European languages and encounter European ideas.

The reformed army and administration became the agents by which the sultan extended his authority over the semi-independent governors, local notables, valley lords (*derebeys*), and other groups, that had wielded political power in various parts of the empire. This process had begun immediately after 1812. The Serbian revolt had been temporarily suppressed in 1813, although it broke out again in 1815. Firm Ottoman governmental control was established over Anatolia, Iraq, and much of Rumelia.

The only local ruler who succeeded in asserting his own authority, unaided against the Porte, was Muḥammad 'Alī of Egypt, who was carrying through a still more radical program of modernization. In 1831 Egyptian forces invaded Syria, routed the Ottomans at Konya (December 27, 1832) and threatened Istanbul. Mahmud was forced to seek Russian aid, and on July 8, 1833, he signed the Treaty of Hünkâr İskelesi (Unkiar Skelessi); Muḥammad 'Alī was, for a time, left in possession of Syria, but Mahmud had not abandoned his claims. In 1839 he attacked the Egyptians; once more the Ottomans were defeated (June 24, 1839). With the help of the European powers (except France) through the Treaty of London (July 15, 1840), the Ottomans recovered Syria and eventually consolidated their authority there; but Muḥammad 'Alī obtained recognition as hereditary ruler of Egypt (1841).

Attempts to extend Ottoman control in the European provinces, notably in Greece, Serbia, and the Principalities, were frustrated. The Greek revolt was the product of the economic prosperity of the Napoleonic Wars, exposure to western European ideas, and a reaction against Ottoman centralization. It had two sources. The first lay in the opposition of peasants and bandits; the second, in the plots of certain intellectuals organized through the Philikí Etairía (a political society) and led by Alexander Ypsilantis, who invaded Moldavia in March 1821. Ypsilantis was defeated, but a rising began in the Peloponnesus. A stalemate developed; but the Ottomans were reinforced, in 1825, by Egyptian troops and threatened to put down the revolt. The destruction of the combined Ottoman and Egyptian fleets by Russian, French, and British naval forces at Navarino, in the southwestern Peloponnese (October 20, 1827) prevented the Muslims from supplying their armies and made Greek independence inevitable. The Ottomans were forced to recognize first Greek autonomy (1829), and then independence (1832).

Similarly, Ottoman efforts to regain control of Serbia and the Principalities were obstructed by Russian opposition, leading to the Russo-Ottoman War (1828–29). By the Treaty of Edirne (Adrianople) on September 14, 1829, the Ottomans ceded to Russia the mouth of the Danube and important territories in eastern Asia Minor and conceded new privileges to the Principalities and Serbia. Serbian autonomy was recognized in 1830 and was extended over the full area of the state in 1833.

Changes in the empire

By the death of Mahmud II (July 1, 1839) the Ottoman Empire was diminished in extent; it was more consolidated and powerful but increasingly subject to European pressures, with Russia supporting and Britain opposing separatist movements and the other powers oscillating between. The cure, however, had begun. Mahmud had established the "respectability of change," and its symbol was the replacement of the turban with the fez (1828).

The Tanzimat (1839–76). The "Tanzimat" is the name given to the series of Ottoman reforms promulgated during the reigns of Mahmud's sons Abdülmecid I (ruled 1839–61) and Abdülaziz (ruled 1861–76). The best known of these reforms are the Hatt-ı Şerif of Gülhane (Noble Edict of the Rose Chamber; November 3, 1839) and the Hatt-ı Hümayun (Imperial Edict, February 18, 1856).

Purpose of the Tanzimat. The Tanzimat has been the subject of much controversy. Many Western writers have dismissed the promises of reform as an Ottoman desire to win European diplomatic support at critical moments, and some features of the Tanzimat appear to support such a view. The promises of equality for Christian subjects were not always implemented—for example, it was proposed in 1855 to end the poll tax paid by non-Muslims and to allow them to enter the army; but the old poll tax was merely replaced by a new exemption tax levied at a higher rate, and Christians were still excluded from the army. It is also true that the timing of reform announcements coincided with crises—the 1839 edict came when the Ottomans needed European help against Muḥammad 'Alī; the 1856 edict when the Ottomans needed European acceptance in the wake of the Crimean War; and the 1876 constitution when European pressure for reforms was mounting.

This view of the Tanzimat is based, however, upon a misconception of its purpose. Europeans, who were principally concerned with winning better conditions for Ottoman Christians, looked first at those elements of the Tanzimat that appeared to be directed toward this goal (*e.g.*, a proclamation in the 1839 edict of the principles of individual liberty, freedom from oppression, and equality before the law and a substantial section of the 1856 edict that was concerned with the rights of Christians). But to the Ottomans the purpose of reform was to preserve the Ottoman state. While it was necessary to make some concessions to European powers and to their own non-Muslim subjects, and, although some Tanzimat statesmen saw equality as an ultimate goal, it was preservation that required the mobilization of resources for modernization. The central reforms, therefore, were in the army (notably major reorganizations of 1842 and 1869, the latter following the pattern of the successful Prussian conscript system); in the administration, both at the centre and in the provinces; and in society, through changes in education and law.

Reform in education. Before the reforms, education in the Ottoman Empire had not been a state responsibility but had been provided by the various *millets*; education for Muslims was controlled by the *'ulamā'* and was directed toward religion. The first inroads into the system had been made with the creation of naval (1773), military (1793), engineering, medical (1827), and military science (1834) colleges. In this way specialized Western-type training was grafted onto the traditional system to produce specialists for the army. Similar institutions for diplomats and administrators were founded, including

the translation bureau (1833) and the civil service school (1859); the latter was reorganized in 1877 and eventually became the faculty of political science in the University of Ankara (1950) and the major training centre for higher civil servants.

Comprehensive education plan

In 1846 the first comprehensive plan for state education was made. It provided for a complete system of primary and secondary schools leading to university, all under the Ministry of Education. A still more ambitious educational plan, inaugurated in 1869, provided for free and compulsory primary education. Both schemes progressed slowly because of a lack of money, but they provided a framework within which development toward a systematic, secular pattern could take place.

By 1914 there were more than 36,000 Ottoman schools, although the great majority were small, traditional primary schools. The development of the state system was aided by the example of progress among the non-Muslim *millet* schools, in which the education provided was more modern than in the Ottoman schools; by 1914 these included more than 1,800 Greek schools with about 185,000 pupils and some 800 Armenian schools with more than 81,000 pupils. Non-Muslims also used schools provided by foreign missionary groups in the empire; by 1914 there were 675 U.S., 500 French Catholic, and 178 British missionary schools, with more than 100,000 pupils among them. These foreign schools included such famous institutions as Robert College (founded 1863), the Syrian Protestant College (1866; later the American University of Beirut), and the Université Saint-Joseph (1874).

Reforms in law. Law too, to a large extent, had been the responsibility of the various *millets*. The Capitulations exempted foreigners and those Ottoman citizens on whom foreign consuls conferred protection from the application of criminal law. The Tanzimat reformers had two objects in the reform of law and legal procedure—to make Ottoman law acceptable to Europeans so that the capitulations could be abolished and sovereignty recovered, and to modernize the traditional Islāmic law. Their efforts resulted in the promulgation of a number of codes —a commercial code (effective in 1850), a commercial procedure (1861), a maritime code (1863), and a penal code (1858). French influence predominated in these, as it did in the civil code of 1870–76. Increasingly, the laws were administered in new state courts, outside the control of the *'ulamā'*. Although they failed to achieve the purposes intended, they provided the basis for future success.

Obstructions to reforms. The Tanzimat reforms moved steadily in the direction of modernization and centralization. The reformers were handicapped by lack of money and skilled men; and they were opposed by traditionalists, who argued that the reformers were destroying the empire's fundamental Islāmic character and who often halted the progress of reform. Centralization, meanwhile, was slowed by European interference, which obstructed the Ottoman attempt to recover power in Bosnia and Montenegro in 1853; forced the granting of autonomy to Mount Lebanon in 1861; and considered, but eventually rejected, intervention to prevent the Ottomans from suppressing a Cretan revolt of 1868. Although Britain and France helped the Ottomans resist Russian pressure during the Crimean War (1853–56), the Ottomans derived no real benefits from the peace settlement; and new arrangements helped to bring about the unification of the Principalities (1859) and paved the way for the emergence of independent Romania.

European interference

The very success of the Tanzimat reformers revealed a novel weakness. The effect of centralization removed the checks on the power of the sultan. After the death of Ali Paşa, Abdülaziz so abused his unrestrained authority as to produce a major crisis in 1875–78.

The 1875–78 crisis. Drought in 1873 and floods in 1874 produced widespread discontent and even famine among an Ottoman peasantry already disturbed by its increased burdens under a landholding system that had spread in the Balkans in the 19th century and by increased taxation and greater liability to conscription as a result of the 1869 military reorganization. The burden of taxation had been aggravated by the Ottoman problems of debt repayment. The first Ottoman foreign loan was in 1854; by 1875 the nominal public debt was £200,000,000 with annual repayments and amortization taking £12,000,000, or more than half the national revenue, but the Ottomans could pay only half the interest due because a world financial crisis in 1873 had made new credit difficult to obtain.

Balkan discontent was fanned by nationalist agitation supported by Serbia and by émigré Slav organizations. It culminated in risings largely of Christian peasants against Muslim lords in Bosnia and Hercegovina (July 1875) and in Bulgaria (August 1876). Ottoman efforts to suppress the risings led to war with Serbia and Montenegro (July 1876) and to attempts by European powers to force Ottoman reforms.

Agreement among the European powers proved impossible; and when the Ottomans rejected its demands, Russia decided to act alone and declared war (April 24, 1877). The war ended in defeat for the Ottomans, but their unexpected resistance at Plevna (modern Pleven, Bulgaria; July–December 1877) allowed other European powers, led by Britain, to intervene. According to the Treaty of San Stefano (March 3, 1878), the Ottomans were to recognize the independence of Romania, Serbia, and Montenegro and cede territory to them; to concede autonomy to an extensive new state of Bulgaria; to cede territory to Russia in the Dobruja (west of the Black Sea) and eastern Asia Minor; and to introduce various administrative reforms and pay an indemnity.

Russo-Turkish War, 1877–78

Diplomatic pressure from other European powers led to the modification of these terms at the Congress of Berlin (June–July 1878). The major changes concerned autonomous Bulgaria, which was substantially reduced in size and divided into two parts—the northern part to have political and the southern (eastern Rumelia) to have administrative autonomy. The independence of Serbia, Montenegro, and Romania was recognized, but their territorial gains were much reduced. Russia retained its acquisitions of Kars and Batum in Asia Minor. Austria-Hungary was given control of Bosnia and Hercegovina and the strategic district of Novi Pazar, in Yugoslavia. By a separate convention Cyprus was put under British rule.

The settlement was a major defeat for the Ottomans. Eastern Rumelia was soon lost, when it united with Bulgaria in 1885. The Ottoman territories in Europe were reduced to Macedonia, Albania, and Thrace, and European influence had attained new dimensions. Britain now proposed to supervise governmental reforms in the Asian provinces, although this was skillfully frustrated by Abdülhamid II (ruled 1876–1909). In addition, the Ottomans were soon forced to accept new financial controls. By the Decree of Muharrem (December 1881) the Ottoman public debt was reduced from £191,000,000 to £106,000,000, certain revenues were assigned to its service, and a European-controlled organization (the Ottoman Public Debt Administration—OPDA) was set up to collect them.

The OPDA subsequently played an important part in Ottoman affairs, acting as agent for the collection of other revenues and as an intermediary with European companies seeking investment opportunities. Its influence should not, however, be exaggerated. It remained under Ottoman political control, and its existence even enabled the Ottomans to add to the debt at the annual rate of £3,000,000 throughout the reign of Abdülhamid; nor was the burden of repayments a major drain on the country's resources. But taken in conjunction with the activities of European-controlled banks and with the tariff limitations imposed on the Ottomans by the capitulations, the result was a distinct restriction on Ottoman ability to guide the allocation of resources.

The constitution, 1876. Perhaps more significant than external changes were the internal political developments that brought about the first Ottoman constitution, December 23, 1876. The Tanzimat had produced three types of criticism within the Muslim community. The first was a simple traditionalist opposition. The second was a more

sophisticated critique elaborated by certain intellectuals, many of whom had bureaucratic training and some knowledge of Western ideas. The third was a determination to depose the sultan (see below).

The Young Ottomans

The intellectuals were known as the Young Ottomans. Although some had taken part in a secret society (the "Patriotic Alliance") in 1865 and had some similarity of background, the Young Ottomans were not an organized political party; they are considered as a group largely through the accident of their assembly in Paris and London in 1867–71. Their political views ranged from secular, cosmopolitan revolutionism to profoundly Islāmic traditionalism. Because his views occupied a middle ground among these intellectuals and because of his lucidity of expression, Namık Kemal (1840–88) has often been regarded as the representative figure, although he is no more representative than the others. His views, however, had the greatest effect on later reformers.

Kemal criticized the Tanzimat reformers for their indiscriminate adoption of Western innovations. While admiring much of Western civilization, he believed that the principles underlying its best institutions were to be found in Islām. In particular he derived from early Islāmic precept and practice the idea of a representative assembly that could check the unbridled power of the sultan and his ministers. He helped to form and popularize the idea of a constitution and of loyalty to the Ottoman fatherland. Like others, he was assisted by the development of an Ottoman press, which had its origins in the 1830s but began to express opinions, occasionally critical of the government, in the 1860s, which saw the establishment of the *Tercüman-ı Ahval* (1860) and the *Tasvir-i Efkâr* (1862), both of which, along with later newspapers, became the vehicles for Young Ottoman ideas.

But it was the third line of criticism that was most important. Arising within the higher Ottoman bureaucracy itself, it was led by Midhat Paşa. Midhat and others became convinced by their own exclusion from power and by the disastrous results of Abdülaziz's policies that some check was needed upon the sultan's power. The traditional check was deposition, and this was accomplished (May 30, 1876) following a riot by theological students and the removal of the hated grand vizier Mahmud Nedim Paşa. A new Cabinet was formed, which included Midhat and other partisans of reform. A new sultan, Murad V (ruled 1876), with a reputation for liberalism was installed; but Murad became insane and was deposed in August and replaced by Abdülhamid II. The experience convinced Midhat of the necessity of a permanent check upon the power of the sultan, such as could be provided by a representative assembly that would give ministers a basis of support independent of the sultan. Accordingly, Abdülhamid was persuaded to agree to a constitution.

First Ottoman constitution

Although earlier documents had had constitutional implications and although the development of councils—particularly provincial councils with their elected elements—had had parliamentary aspects, the December 23 document was the first comprehensive Ottoman constitution and, except for a Tunisian organic law of 1861, the first in any Islāmic country. The constitution was derived entirely from the will of the ruler, who retained full executive power and to whom ministers were individually responsible. In legislation the sultan was assisted by a two-chamber Parliament—the lower indirectly elected and the upper nominated by the ruler. Rights of ruler and ruled were set out, but the system it established might best be described as attenuated autocracy. Midhat has been criticized for accepting certain amendments demanded by Abdülhamid, including the then notorious article 113, which gave the sultan the right to deport persons harmful to the state; but it is clear that the majority of Midhat's colleagues were content with these amendments, and that the amendments made little difference, so great were the sultan's powers within and outside the constitution. The Parliament summoned under the constitution in March 1877 was dissolved in less than a year and was not recalled until 1908. The liberals were exiled; some, incuding Midhat, were put to death.

Abdülhamid II (ruled 1876–1909). The reign of Abdülhamid II is often regarded as having been a reaction against the Tanzimat; but, insofar as the essence of the Tanzimat reforms was centralization rather than liberalization, Abdülhamid may be seen as its fulfiller rather than its destroyer. The continued development of the army and administration, the formation of a gendarmerie, the growth of communications—especially the telegraph and railways—and the formation of an elaborate spy system enabled the sultan to monopolize power and crush opposition. His brutal repression of the Armenians in 1894–96 earned him the European title red sultan. But Abdülhamid's reign also made positive advances in education (including the foundation of the University of Istanbul in 1900); in legal reform, led by his grand vizier Mehmed Said Paşa; and in economic development, through the construction of railways in Asia Minor and Syria with foreign capital and of the Hejaz Railway from Damascus to Medina with the help of international Muslim subscriptions.

Pan-Islāmism. The Hejaz Railway constituted one element in Abdülhamid's Pan-Islāmic policies. Political Pan-Islāmism had made its first appearance in Ottoman policy at the Treaty of Küçük Kaynarca (1774) with Russia, when the Ottoman sultan had put forward claims to religious jurisdiction over Muslims outside his territories (particularly those in the Crimea). Some years later the theory was elaborated by the addition of the baseless legend that in 1517 the ʿAbbāsid caliphate had been transferred to the Ottoman sultan. With the extinction of many independent Muslim states and their absorption into the empires of European powers, this myth of the caliphate became a useful weapon in the Ottoman diplomatic armoury and was exploited by Abdülhamid as a means of deterring European powers from pressing him too hard, lest he create dissension within their own territories. In addition, stress on popular Islām through the press and other publications and through the Sultan's patronage of dervish orders served to rally Muslim opinion within the empire behind him.

Power of the myth of the caliphate

Preservation of the empire. Abdülhamid had reasonable success in preserving the empire after 1878. Apart from eastern Rumelia, no further territories were lost until 1908 (Ottoman authority in Tunisia, occupied by France in 1881, and that in Egypt, occupied by Britain in 1882, was already insignificant). In Crete the Ottomans suppressed revolts and defeated Greece when it intervened in 1897 in support of the Cretans. The European powers, however, forced Abdülhamid to concede autonomy to Crete. He was more successful in obstructing the efforts of the powers to force the introduction of substantial reforms in Macedonia. In Arabia the Ottomans continued the expansion of their power that had begun in the early 1870s.

The Young Turk Revolution, 1908. Several conspiracies took place against Abdülhamid. In 1889 a conspiracy in the military medical college spread to other Istanbul colleges. These conspirators came to call themselves Union and Progress (İttihad ve Terakki). When the plot was discovered, some of its leaders went to reinforce Ottoman exiles in Paris, Geneva, and Cairo, where they helped to prepare the ground for a revolution by developing a comprehensive critique of the Hamidian system. The most noteworthy among these were Murad Bey, Ahmed Rıza, and Prince Sabaheddin. As editor of *Mizan* ("Balance"), published first in Istanbul (1886) and later in Cairo and Geneva, Murad Bey preached liberal ideas combined with a strong Islāmic feeling; this last may have contributed to his defection and return to Istanbul in 1897. Ahmed Riza in Paris edited *Meşveret* ("Consultation"), in which he set out ideas of reform, strongly flavoured by positivism. His advocacy of a strong central government within the Ottoman Empire and the exclusion of foreign influence led to a major split within the Young Turk exiles at the 1902 Paris Congress; Ahmed Rıza clashed with Sabaheddin, who, with Armenian support, favoured administrative decentralization and European assistance to promote reform. Sabaheddin set up the League of Private Initiative and Decentralization.

The émigrés could supply literary sustenance to dissidents, but Abdülhamid could not be overthrown so long as the army remained loyal. The real origin of the Young Turk Revolution of 1908 lay in the discontent within the 3rd Army Corps in Macedonia, where officers acted quite independently of the Committee of Union and Progress (CUP) in Paris. It is still unclear if a coordinated conspiracy existed in Macedonia or if a number of separate centres of disaffection, linked haphazardly through individuals, dervish orders, Freemason lodges, etc., coalesced in July 1908 under the banner of the CUP through the pressure of events. On July 3, 1908, Maj. Ahmed Niyazi (1873–1912), apparently fearing discovery by an investigatory committee, decamped from Resne with 200 followers, including civilians, leaving behind a demand for the restoration of the constitution. The Sultan's attempt to suppress this rising failed, and rebellion spread rapidly. Unable to rely on other troops, on July 24 Abdülhamid announced the restoration of the constitution.

Goals of rebel leaders

The young officers who had made the revolution, like their civilian supporters, were primarily concerned with preserving the Ottoman Empire; they feared that Hamidian policies and European interventions were endangering its existence. Personal grievances concerned with pay, promotion, etc., also played a part. Though some writers have argued that a new type of officer of lower than usual social origins influenced this discontent, there is little evidence to support such a theory. It is clear, however, that the officers had not thought much beyond their demand for the restoration of a constitution that had proved ineffectual in 1877–78. They had no program of action and were content to leave government to the established bureaucrats.

In April 1909, however, an army mutiny in Istanbul (known because of the Julian calendar as the "31st March Incident") exposed the weakness of the CUP and at the same time gave it a new opportunity. The mutiny resulted from the discontent of ordinary soldiers arising from their conditions and their neglect by college-trained and politically ambitious officers, and from what they regarded as infidel innovations. They were encouraged by a religious organization—the Mohammedan Union. Government weakness allowed the mutiny to spread; and although order was eventually restored in Istanbul and more quickly elsewhere, a force from Macedonia (the Action Army) led by Mahmud Şevket Paşa marched on Istanbul and occupied the city (April 24).

Dissolution of the empire. Abdülhamid was deposed and replaced by Sultan Mehmed V (ruled 1909–18), son of Abdülmecid. The constitution was amended to transfer real power to the Parliament. The army, and particularly Şevket Paşa became the real arbiters of Ottoman politics.

Rise of the CUP. Although the removal of many of its political opponents had allowed the CUP to move into a more prominent position in government, it was still weak. It had a core of able, determined men but a much larger collection of individuals and factions that wore the Unionist label so lightly that they easily merged into other parties. Although the CUP won an overwhelming majority in the election of April 1912, its support rapidly melted away following military losses to Italy. Evidence of army hostility finally forced the CUP out of office in July 1912, to be succeeded by a political coalition called the Liberal Union.

The Liberal Union, too, lost support following defeats in the Balkans. This provided the opportunity for a small group of CUP officers and soldiers to stage a coup (January 23, 1913), known as the Sublime Porte Incident, to force the resignation of the grand vizier Mehmed Kâmil Paşa and establish a new cabinet under Şevket. Şevket, however, was not a Unionist and it was only after his assassination (June 11, 1913) that the CUP at last succeeded in establishing a Unionist-dominated government under Said Halim Paşa.

The Sublime Porte Incident

Internal developments. The disastrous results of the Young Turks' external policies have overshadowed the important internal developments of the years 1908–18. Further administrative reforms, particularly of provincial administration in 1913, led to more centralization, although by European standards the central Ottoman government remained relatively weak, particularly in the provinces distant from Istanbul. The burden of taxation was well below that of European powers.

The Young Turks were the first Ottoman reformers to promote industrialization, with a Law for the Encouragement of Industry (1909, revised 1915). Although they had little success, they did build a framework for later state-directed economic planning. Considerable attention was given to education, especially to the neglected area of primary education. The process of secularization of the law was carried much further. A major development in national journalism took place, and the position of women improved. The whole period was one of intense social and political discussion and change.

Turkish nationalism. The basic ideologies of the state remained Ottomanism and Islām, but a new sense of Turkish identity began to develop. This new concept was fostered by educational work of the Turkish Society (1908) and the Turkish Hearth (1912). A political twist was given by the adherents of Pan-Turkism and Pan-Turanianism. Pan-Turkism, which aimed at the political union of all Turkish-speaking peoples, began among Turks in the Crimea and on the Volga. Its leading exponent was Ismail Bey Gasprinski (Gaspirali; 1851–1914), who attempted to create a common Turkish language. Many Pan-Turkists migrated to Ottoman lands, especially after 1905. One of them, Yusuf Akçuraoğlu, argued in *Üç tarz-ı siyaset* (1903; "Three Kinds of Policy") that Turkism provided a better basis for the Ottoman Empire than either Islām or Ottomanism. Pan-Turanianism developed from a now much-disputed 19th-century theory of the common origin of Turkish, Mongol, Tungus, Finnish, Hungarian, and other languages; in certain very limited circles it looked forward to a great political federation of speakers of these languages, extending from Hungary to the Pacific.

Pan-Turkism and Pan-Turanianism

These ideas, however, found little support within the Ottoman government. The accusation that the Young Turks pursued a deliberate policy of Turkification within the empire so as to alienate non-Turks and promote the rise of Arab and Albanian nationalism is an oversimplification. The extension of government activity inevitably brought with it the language of government—Turkish. This produced some reaction from speakers of other languages, but the evidence suggests that it did not override basic feelings of Muslim solidarity, except among some small minorities. It was among the Christian groups that distinct separatist ideas were developed.

Foreign relations. The foreign relations of the Ottoman Empire under the Young Turks led to disaster. The 1908 revolution provided an opportunity for several powers to press home their designs upon the empire. On October 5, 1908, Austria annexed Bosnia and Hercegovina and Bulgaria proclaimed its independence. Italy seized Tripoli and occupied the Dodecanese, a group of Aegean islands; by the Treaty of Lausanne (October 18, 1912) Italy retained the former but agreed to evacuate the Dodecanese. In fact, however, it continued to occupy them.

The two Balkan Wars (1912–13) almost completed the destruction of the Ottoman Empire in Europe. In the first (October 1912–May 1913), the Ottomans lost almost all of their European possessions, including Crete, to Bulgaria, Serbia, Greece, Montenegro, and the newly created state of Albania (Treaty of London, May 30, 1913). In the second (June–July 1913), fought between Bulgaria and the remaining Balkan states, including Romania, over the division of Macedonia, the Ottomans intervened against Bulgaria and recovered part of eastern Thrace including Edirne (Adrianople). The Ottomans had lost 83% of the territory and 69% of the population of their European provinces.

The people. In 1914 the total population of the Ottoman Empire was approximately 25,000,000, of which about 10,000,000 were Turks; 6,000,000, Arabs; 1,500,000 each, Kurds and Greeks; and between 1,000,000 and 1,500,000 Armenians. The population of the empire (excluding such virtually independent areas as Egypt, Romania, and Serbia) in the period immediately prior to the

Population of the empire

losses of 1878 is estimated to have been about 26,000,000. Natural increases and Muslim immigration from Russia and the Balkans virtually made up the losses, and in 1914 the population was more homogeneous in religion and, though less so, in language.

World War I, 1914–18. The Ottoman entry into World War I resulted from an overly hasty calculation of likely advantage. German influence was strong, but not decisive; Germany's trade with the Ottomans still lagged behind that of Britain, France, and Austria and its investments, which included the Baghdad railway, were smaller than those of France. A mission to Turkey led by the German military officer Otto Liman von Sanders in 1913 was only one of a series, and Liman's authority was much more limited than contemporaries supposed. Except for the interest of Russia in Istanbul and the Straits, no European power had really vital interests in the Ottoman Empire. The Ottomans might have remained neutral, as a majority of the cabinet wished to do, at least until the situation became clearer. But the opportunism of the minister of War Enver Paşa, early German victories, friction with the Triple Entente (France, Russia, and Great Britain) arising out of the shelter given by the Ottomans to German warships, and basic hostility to Russia combined to produce an Ottoman bombardment of the Russian Black Sea ports (October 28, 1914) and a declaration of war by the Entente against the Ottoman Empire.

The Ottomans made a substantial contribution to the Central Powers' war effort. Their forces fought in eastern Asia Minor, Azerbaijan, Mesopotamia, Syria and Palestine, and at the Dardanelles, as well as on European fronts, and they held down large numbers of Entente troops. In September 1918 they dominated Transcaucasia. During the war the Young Turks also took the opportunity to attack certain internal problems—the Capitulations were abolished unilaterally (September 1914), the autonomous status of Lebanon was ended, a number of Arab nationalists were executed in Damascus (August 1915 and May 1916), and the Armenian community in eastern Asia Minor and Cilicia was massacred or deported as part of a deliberate policy of eliminating one cause of European interference. Possibly a million Armenians either fled or were killed (principally by Kurdish irregulars) or deported.

From the end of 1916 army desertions took place on a massive scale, and economic pressures became acute. The surrender of Bulgaria (September 28, 1918), which severed direct links with Germany, was the final blow. The CUP Cabinet resigned on October 7, and a new government was formed under Ahmed Izzet Paşa (1864–1937) on October 9. On October 30 the Ottomans signed the Armistice of Mudros.

Allied war aims and the proposed peace settlement. Partition of Ottoman territories

Entente proposals for the partition of Ottoman territories were formulated in a number of wartime agreements. By the Istanbul Agreements (March–April 1915) Russia was promised Istanbul and the Straits; and France, a sphere of influence in Syria and Cilicia. Britain had already annexed Cyprus and declared a protectorate over Egypt. By the Anglo-French Sykes-Picot Agreement (January 3, 1916) the French sphere was confirmed and extended eastward to Mosul in Iraq. A British sphere in Mesopotamia as far north as Baghdad, control of Haifa and Acre, and a linking sphere of influence were recognized. Palestine was to be placed under an international régime. In compensation, the Russian gains were extended (April–May 1916) to include the Ottoman provinces of Trabzon, Erzurum, Van, and Bitlis in eastern Asia Minor. By the London Agreement (April 26, 1915) Italy was promised the Dodecanese and a possible share of Asia Minor. By the Agreement of St.-Jean-de-Maurienne (April 1917) Italy was promised a large area of southwestern Anatolia, including Izmir (Smyrna) and a further sphere to the north. Britain made various promises of independence to Arab leaders, notably in the Ḥusayn-MacMahon correspondence, 1915–16, and, in the Balfour Declaration (November 2, 1917), promised to support the establishment of a national home for the Jewish people in Palestine.

The Russian withdrawal in 1917 and postwar bargaining led to some modifications of these agreements, and the Allied terms were not finally presented until 1920. By the Treaty of Sèvres (August 10, 1920) the Ottomans retained Istanbul and part of Thrace, but lost the Arab provinces and a large area of Asia Minor to a newly created Armenian state with access to the sea, surrendered the islands of İmroz and Bozcaada to Greece, and accepted arrangements that implied the eventual loss of Izmir to Greece. The Straits were internationalized, and strict European control of Ottoman finances was established. An accompanying tripartite agreement between Britain, France, and Italy defined the extensive spheres of influence of the latter two powers. The treaty was ratified only by Greece and was abrogated by the Treaty of Lausanne (July 24, 1923) as the result of a determined struggle for independence waged under the leadership of the outstanding Ottoman wartime general Mustafa Kemal, later known as "Atatürk."

The War of Independence, 1919–23. Although the legal Ottoman government in Istanbul, under the 36th, and last, Ottoman sultan Mehmed VI Vahideddin (ruled 1918–22) had decided that resistance to Allied demands was impossible, pockets of resistance remained in Asia Minor after the armistice. These included bands of irregulars and deserters, certain intact Ottoman forces, and various societies for the defense of rights. Resistance was stimulated by the Greek occupation of Izmir (May 15, 1919). At this time Mustafa Kemal was sent on an official mission to eastern Asia Minor, landing at Samsun on May 19. He immediately began to organize resistance, despite official Ottoman opposition. Through the Association for the Defence of the Rights of Eastern Anatolia (founded March 3, 1919), he summoned a congress at Erzurum (July–August), followed by a second congress at Sivas (September) with delegates representing the whole country. A new Association for the Defence of the Rights of Anatolia and Rumelia was established, and an executive committee with Mustafa Kemal as chairman was created to conduct resistance. Congress at Erzurum

The official government yielded to Kemalist pressure. The unpopular grand vizier, Damad Ferid Paşa, resigned and was replaced by the more sympathetic Ali Rıza Paşa. Negotiations with the Kemalists were followed by the election of a new parliament, which met in Istanbul in January 1920. A large majority in parliament was opposed to the official government policy and passed the "National Pact," which embodied the political aims of independence roughly within the October 1918 armistice lines and which had been formulated at Erzurum and Sivas. The Allies countered by extending the occupied area of Istanbul (March 16, 1920) and by arresting and deporting many deputies. Damad Ferid became grand vizier again on April 5 and, with religious support, set out to crush the Kemalists.

The Kemalists were now faced with local risings, official Ottoman forces, and the Greeks. The first necessity was to establish a legitimate basis of action. A parliament (the Grand National Assembly) met at Ankara on April 23 and asserted that the sultan's government was under infidel control and that it was the duty of Muslims to resist foreign encroachment. In the Fundamental Law of January 20, 1921, the assembly declared that sovereignty belonged to the nation and that the assembly was the "true and only representative of the nation." The name of the state was declared to be "Turkey" ("Türkiye"); and executive power was entrusted to an executive council, headed by Mustafa Kemal, who could now concentrate on the war.

TURKEY SINCE 1920

The uprisings and Ottoman forces were both defeated, principally by irregular forces, who at the end of 1920 were brought under Kemal's control. In 1920–21 the Greeks made major advances, almost to Ankara, but were defeated at the Battle of the Sakarya River (August 24, 1921) and began a long retreat that ended in the Turkish occupation of Izmir (September 9, 1922).

The Kemalists had already begun to gain European rec-

ognition. On March 16, 1921, the Soviet-Turkish Treaty gave Turkey a favourable settlement of its eastern frontier by restoring Kars and Ardahan. Domestic problems induced Italy to begin withdrawal from the territory it occupied; and, by the Treaty of Ankara (Franklin-Bouillon Agreement, October 20, 1921), France agreed to evacuate Cilicia. Finally, by the Armistice of Mudanya the Allies agreed to Turkish reoccupation of Istanbul and eastern Thrace.

New Turkish frontier

A comprehensive settlement was eventually achieved at the Lausanne Conference (November 1922–July 1923). The Turkish frontier in Thrace was established on the Maritsa River, and Greece returned the islands of İmroz and Bozcaada. A compulsory exchange of populations was arranged, as a result of which an estimated 1,300,000 Greeks left Turkey in return for 400,000 Turks. The question of Mosul was left to the League of Nations, which in 1925 recommended its retention by Iraq. The Lausanne Treaty also provided for the apportionment of the Ottoman Public Debt, for the gradual abolition of the Capitulations (Turkey regained tariff autonomy in 1929), and for an international régime for the Straits. Turkey did not recover complete control of the Straits until the 1936 Montreux Convention.

The result of the war and the peace settlement created a state in which the great majority spoke Turkish. Though there has been a tendency to see this as the almost inevitable consequence of the rise of Turkish and Arab nationalisms, it seems in fact to have been the accident of war that broke off the Arab provinces. Whatever the views of Mustafa Kemal himself, it is clear that the majority of his followers thought of themselves primarily as Muslims; in the elaborate religious ceremony that preceded the opening of the Grand National Assembly there was no word of Turks or Turkey but only of the need to save "religion's last country." The creation of a sense of Turkish nationhood was the product of a long effort in which Mustafa Kemal played the dominant role.

Kemalism, 1922–38. Construction of a new political system began with the abolition of the sultanate and the declaration of a republic. Loyalty to the Ottoman dynasty was strong even among Kemalists; but Vahideddin's identification with the Allies weakened his support.

Abolition of the sultanate and caliphate. An Allied invitation to the sultan to nominate representatives to Lausanne aided Kemal—a split Turkish delegation would be self-defeating. With a brilliant mixture of threats and persuasion, Kemal was able, therefore, to induce the Assembly to abolish the sultanate (November 1, 1922). Vahideddin left Turkey, and his cousin Abdülmecid (died 1944) was installed as the first and last Ottoman caliph who was not also sultan. The caliphate was finally abolished on March 3, 1924, and all members of the Ottoman dynasty were expelled from Turkey. Before that

Republic of Turkey

the assembly had declared Turkey to be a republic and had elected Mustafa Kemal as first president (October 29, 1923). A full republican constitution was adopted on April 20, 1924; it retained Islām as the state religion, but in April 1928 this clause was removed and Turkey became a purely secular republic.

Kemal's government. The assembly was the instrument of Kemal's will. The first assembly had contained large factions hostile to his policies, including religious conservatives, merchants, and former members of the CUP. In opposition to his 197 acknowledged supporters, who were known as the "First Group," there were 118 opponents, members of a "Second Group." The first Assembly was dissolved on April 16, 1923, and Mustafa Kemal took care to keep his opponents out of the second assembly; only three of the Second Group were returned. Kemal's own party, which became the Republican People's Party (RPP), dominated all assemblies until 1950; this period saw a heavy preponderance in the Assembly of urban professional men and of officials with university education. With an outlook different from that of the illiterate Turkish peasant they carried out a revolution from the top.

Opposition. There was little opposition to Mustafa Kemal—the small Progressive Party (founded, November 17, 1924; dissolved, June 5, 1925) had only 29 members and was suppressed because Kemal feared that its leading members, who included some of his most notable associates in the war of independence, might have too much influence in the army; the short-lived Liberal Republican Party (August 12–December 18, 1930) was an abortive attempt by Kemal to organize a moderate opposition to his own party. Otherwise Kemal ruled quite autocratically. A plot against his life in 1926 gave him the chance to deal with his rivals, who were tried by a special court and many of them sentenced to death, imprisonment, or exile. Opposition outside the assembly, of which the most dangerous were the Kurdish revolts of 1925, 1930, and 1937, was suppressed vigorously.

Kemalist policies. The bases of Mustafa Kemal's policies were enshrined in the Republican People's Party program of 1931 and subsequently (February 1937) written into the Turkish constitution. These were the six principles of republicanism (the creation of the republic), nationalism, populism, statism (*devletçilik*), secularism, and revolution. Revolution was implicit in the whole radical reorganization of the political, social, and economic systems. Populism was the effort to mobilize popular support from the top through such characteristic devices as the People's Houses (1931–51), which spread the new concept of a national culture in provincial towns, and the village institutes, which performed the same educational and proseletyzing role in the countryside. The creation of a sense of nationalism was encouraged by changes in school curricula, by the rewriting of history to glorify the Turkish past, by the "purification" of the language by a reduction of the number of words of foreign origin (sometime later, this effort appeared to be redundant in the light of a promulgation that all languages were descended from Turkish), and by the renunciation of Pan-Islāmic, Pan-Turkish, and Pan-Ottoman goals in foreign policy.

Mustafa Kemal's six principles

Statism was the movement toward state-controlled economic development; the shortage of skilled labour and entrepreneurs (particularly owing to the reduction of the Greek and Armenian communities, which in 1914 had controlled 80% of Ottoman finance, industry, and commerce), the lack of capital, and the intense nationalist desire for industrial self-sufficiency that would banish foreign influence, all stimulated a movement in the 1930s towards state ownership or control. This was achieved through investment banks, monopolies, state industrial enterprises, and planning. A five-year plan was instituted in 1934. Although the immediate results were disappointing, the policy of state-inspired economic growth was important for future economic advance.

Under secularism is included the reform of law, involving the abolition of religious courts and schools (1924) and the adoption of a purely secular system of family law. The substitution of Latin for the Arabic alphabet in writing Turkish was a significant step toward secularism and made learning easier; other measures included the adoption (1925) of the Gregorian calendar that had been jointly used with the Hijrī calendar since 1917; the replacement of Friday by Sunday as the weekly holiday (1935); the adoption of surnames (1935); and, most striking of all, the abolition of the fez (1925). The wearing of clerical garb outside places of worship was forbidden in 1934.

These changes, coupled with the abolition of the caliphate, and the elimination of the dervish orders after a Kurdish revolt in 1925, dealt a tremendous blow to Islām's position in social life, completing the process begun in the Tanzimat. With secularism there came a steady improvement in the status of women, who were given the right to vote and to sit in the parliament.

Vital as the changes made were, in many cases they were primarily matters of appearance and style. Structural changes in society took longer. At the first census, in 1927, the population was put at 13,600,000, of which 24% was urban. In 1940 the population was 17,800,000, but the urban percentage was almost unchanged. In 1938 the per capita income and literacy were both below comparable figures for developed countries.

Foreign policy was subordinated to internal change. The loss of Mosul was accepted (June 5, 1926). Hatay, however, was recovered. It was given internal autonomy by France in 1937, occupied by Turkish troops in 1938, and incorporated into Turkey in 1939. Turkey followed a neutralist policy, supported the League of Nations, which it joined in 1932, and sought alliances with other minor powers, leading to the Balkan Entente (1934) and the Sa'dābād Pact with Iran, Iraq, and Afghanistan (1937).

Neutralist policy

World War II and the postwar era, 1938–50. Atatürk's autocratic, dominating, and inspiring personality had directed and shaped the Turkish republic. At this death in 1938, his closest associate İsmet İnönü (born 1884) was elected president. With the approach of war, foreign affairs assumed greater importance. An alliance with Britain and France (October 19, 1939) was not implemented because of Germany's early victories. After Germany's invasion of Russia (June 22, 1941), there was popular support for an alliance with Germany, which seemed to offer prospects of realizing old Pan-Turkish aims. Although a nonaggression pact was signed with Germany (June 18, 1941), Turkey clung to neutrality until an Axis defeat was seen to be inevitable, when it entered the war on the Allied side (February 23, 1945). The great expansion of Soviet power exposed Turkey, in June 1945, to demands that control of the Straits be given to Black Sea powers and for the cession of territory in eastern Asia Minor. It was also suggested that a large area of northeastern Anatolia be ceded to Soviet Georgia and a more democratic government established in Turkey. This caused Turkey to seek U.S. assistance; it received U.S. military aid beginning in 1947 (providing the basis for a large and continuing flow of military aid) and economic assistance beginning in 1948.

Economic problems. The war also brought changes in domestic policy. The army had been kept small throughout the Atatürk period, and defense expenditure had been reduced to 28% of the budget. The army was rapidly expanded in 1939, and defense expenditures rose to between 50% and 60% of the budget for the duration of the war. Substantial deficits were incurred, imposing a severe economic strain, which was aggravated by shortages of raw materials. By 1945 agricultural output had fallen to 70% of the 1939 figure; per capita income, to 75%. Inflation was strong; official statistics show a rise of 354% between 1938 and 1945, but this probably understates the fall in the value of money, which in 1943 was less than one-fifth its 1938 purchasing power. One means chosen by the government to raise money was a capital levy introduced in 1942, arranged to fall with punitive force upon the non-Muslim communities and upon the Dönmes (Jewish converts to Islām). The war did provide some stimulus to industry, however, and enabled Turkey to build up substantial foreign credits, which could be used to finance postwar economic development.

Political changes. The most notable change in the postwar years was the liberalization of political life. The investment in education was beginning to show some return, and literacy had risen to 29% by 1945. A growing class of professional and commercial men demanded more freedom. The Allied victory had made democracy more fashionable; accordingly, the government made concessions allowing new political parties, universal suffrage, and direct election.

Political liberalization

From a split within the Republican People's Party (RPP), the Democrat Party (DP) was founded in 1946 and immediately gathered support. Despite government interference, the DP won 61 seats in the 1946 general election. Some elements in the RPP, led by the prime minister Recep Peker (served 1946–47), wished to suppress the DP, but they were prevented by İnönü. In his declaration of July 12, 1947, İnönü stated that the logic of a multiparty system implied the possibility of a change of government. Prophetically, he renounced the title of "National Unchangeable Leader," which had been conferred upon him in 1938. Peker resigned and was succeeded by more liberal prime ministers in Hasan Saka (served 1947–49), and Şemseddin Günaltay (served 1949–50).

Other restrictions on political freedom, including press censorship, were relaxed. The period saw the establishment of the first mass circulation, independent newspapers. Trade Unions were permitted in 1947, although they were not given the right to strike until 1963. A far-reaching measure of land redistribution was passed in 1945, although little was done to implement it before 1950. Other political parties were established including the conservative National Party (1948); Socialist and Communist activities, however, were severely repressed.

In the more open atmosphere the DP was able to organize in the villages. The RPP, despite its Village Institutes, had always been the government party and had had little real grass-roots organization. The Democrats were much more responsive to local interests. The DP won a massive victory in the 1950 elections, claiming 54% of the vote and 396 out of 487 seats. The RPP won 68 seats; the National Party, one. The DP victory has been attributed variously to U.S. influence, to social change, to a desire for economic liberalization, to better organization, to religious hostility to the RPP, and to a bad harvest in 1949. Perhaps the basic reason was that in 27 years the RPP had made many enemies.

Turkey under the Democrats, 1950–60. In the DP government Celâl Bayar became president; Adnan Menderes, prime minister (a post which for the first time came to surpass that of the president in importance).

The economy. The Democrats were pledged to a program of economic growth, to be achieved through a reduction of state interference. At first they had much success; in 1950–53 Turkey's average annual rate of economic growth was 6%, and annual per capita income rose at a rate of 3%. Good harvests in 1950 and 1953 and the Korean War boom assisted. But problems appeared after 1953. In 1954 a poor harvest obliged Turkey to import wheat again. A shortage of foreign exchange limited the purchase of essential materials and parts, which handicapped industry. After a sudden favourable surge in the early 1950s, the terms of trade moved steadily against Turkey. Inflation, which averaged 15% or more annually, became a serious problem. The government attempted unsuccessfully to check the price rise by legislation, but its policies contributed to inflation as the result of a continued rise in public expenditure. Despite the problems, the DP achieved considerable success over the decade 1950–60. The proportion of the gross national product going to investment was raised (with overseas aid) from about 10% to between 12% and 15%, and annual growth averaged 5%.

Political repressions. The political fortunes of the Democrat government closely reflected the economic changes. In the 1954 elections—the Democratic peak—the DP took 57% of the vote and gained 503 out of 541 seats; the RPP took 35% and 31 seats. Subsequent economic difficulties led to mounting criticism within and outside the DP, to which the government replied with increasing repression. In 1953 much of the property of the RPP was confiscated, forcing the closure of the People's Houses. The RPP newspaper presses in Ankara were seized. In 1954 the National Party was dissolved because of its opposition to Kemalist principles (it was immediately reformed as the Republican Nation's Party and united in 1958 with the Peasants' Party to form the Republican Peasants' Nation Party). Laws passed in 1954 provided for heavy fines on journalists who damaged the prestige of the state or the law; several prominent journalists were prosecuted under this law, which was made more severe in 1956, in which year other laws substantially abridged the independence of civil servants (including university teachers) and judges. In October 1955 critics within the DP were expelled; these critics subsequently formed the Freedom Party (December 1955), which later (1958) merged with the RPP. In 1956 limitations were placed upon public meetings.

Seizure of RPP newspaper presses

The DP's loss of popularity was reflected in the elections of October 1957, when they won only 48% of the vote and 424 out of 610 seats. The RPP took 41% and 178 seats. The three opposition parties had attempted to form an electoral coalition, but a law passed that September

had declared such coalitions illegal. The combined opposition vote was 52% of the total, and many persons believed that this government action had deprived them of victory. Opposition attacks upon the DP became stronger, and it was accused of unconstitutional action. At the same time, the Democrats, fearing a revolution, redoubled control. In December 1959 an alleged plot (the so-called Nine Officers' Plot) was unearthed; some of the accused were so clearly innocent that punishment fell upon the accuser, but it appears that there had been a conspiracy of some sort.

A charge on which the RPP laid great stress was that the DP was reversing the principles of secularism and favouring conservative religious organizations. The DP had relaxed some of the secularist policies of pure Kemalism, following in the steps of the RPP in the years 1945–49. Religious instruction in schools had been extended and the organization of religious schools permitted. Arabic had been reinstated for the call to prayer, and radio readings of the Qur'ān had been allowed. These, however, were modest concessions in themselves, and the Democrats had clearly demonstrated their unwillingness to tolerate religious influence in politics by suppressing the activities of dervish orders in 1950–52.

Growth of economic problems

The years 1958–60 saw a further worsening of the economic situation as the government reluctantly introduced restrictive measures. Returns on new investment fell and inflation continued. Serious problems of housing, unemployment, etc., were emerging in the large towns, whose population had been growing annually at the rate of about 10%, so that by 1960 the proportion of urban population to the whole had risen, to 32%. RPP attacks became more bitter, and the government's response stronger. In April 1960 the government ordered the army to prevent İnönü from campaigning in Kayseri and followed this by forming a committee to investigate the affairs of the RPP. It was widely believed that the next action would be to close the RPP. Student demonstrations followed, and martial law was declared on April 28. The army had been brought directly into the political arena.

Military takeover. Relatively neglected from 1923 to 1939, the army during the war had undergone a rapid expansion and a considerable modernization subsequently with the aid of U.S. advisers. Many officers feared that the DP threatened the principles of the secular, progressive Kemalist state. Some younger officers saw the army as the direct instrument of unity and reform. On May 3, 1960, the commander of the land forces, Gen. Cemal Gürsel, demanded political reforms, and resigned when they were refused. On May 27 the army acted; an almost bloodless coup was carried out by officers and cadets from the Istanbul and Ankara War colleges. The leaders established a 38-man "National Unity Committee" with Gürsel as chairman. The Democrat leaders were imprisoned.

The National Unity Committee, 1960–61. From the outset, a clear division existed among the officers who had carried out the coup. One group, predominantly of younger officers, believed that to restore national unity and carry out major social and economic reforms it was necessary to retain power for some years; this group included both those who supported a nationalistic and Islāmist policy and those who favoured accelerated secularization. Another group, which included most of the senior officers, wanted to withdraw the army from politics as soon as possible. In November 1960 the dispute was decided against the first group, and 14 were expelled from the committee and and sent into diplomatic exile.

Work of the National Unity Committee

The main work of the National Unity Committee was to destroy the DP and to prepare a new constitution. Substantial purges took place—5,000 officers including 235 out of 260 generals were dismissed or retired; 147 university teachers left their jobs; 55 wealthy landowners were banished from eastern Anatolia, their lands confiscated. The DP was abolished (September 29, 1960), and many Democrats were brought to trial at Yassi Ada (October) on charges of corruption, unconstitutional rule, and high treason. Of 601 tried, 464 were found guilty. Three former ministers, including Menderes, were executed; 12 others, including Bayer, had their death sentences commuted to life imprisonment.

Work on the new constitution began immediately after the coup, when a committee of five law professors was appointed to prepare a draft. This was submitted to the National Unity Committee on October 18. The Committee appointed a second committee to redraft the constitution; the new draft was presented to a Constituent Assembly, which met in January 1961. The constitution was completed in May and approved by 61% of the voters at a referendum in July.

The new constitution established a two chamber parliament—consisting of the Senate and the National Assembly. A separate electoral law provided for proportional representation. The president was elected by the Senate and National Assembly together. The constitution also provided for a Constitutional Court and a State Planning Organization. The first elections were held in October 1961. The army then withdrew from direct political involvement, although the members of the National Unity Committee retained some influence as life members of the Senate.

Period of transition, 1961–65. No party won a majority in October 1961. The RPP won 38% of the votes and 173 of the 450 Assembly seats. The newly formed Justice Party (JP), led by retired general Rağip Gümüspala, received 35% and 158 seats. The remaining seats were divided between two smaller parties—the Republican Peasants' Nation Party, which took 54 seats, and the liberal New Turkey Party, which gained 65. The results demonstrated the enduring popularity of the old Democrat Party. Its votes had been divided among the three smaller parties, the majority of them going to the Justice Party, which had also emerged as the largest party in the Senate. The RPP had failed to hold all its 1957 vote and had suffered by identification with the army coup.

The new Grand National Assembly elected General Gürsel as president. The RPP leader İnönü formed a coalition government with the JP, but the coalition survived only until June 1962, when it broke up over the question of an amnesty for the imprisoned Democrats. After some delay and splits within the parties (which led to the formation of the Nation Party by dissidents who withdrew from the Republican Peasants' Nation Party), the RPP formed a coalition with the two smaller parties. This accelerated the tendency for former Democrat voters to turn to the JP.

In the local elections of 1963 the JP made extensive gains at the expense of the two smaller parties. This led to the breakup of the coalition, and because the JP was unable to form a government, İnönü formed a minority government from his own party alone, but with voting support from the New Turkey Party. The RPP government resigned after a defeat on the budget in February 1965 and was replaced by a coalition of all the other parties under the leadership of an independent, Suat Hayri Ürgüplü; this coalition acted as caretaker until the elections of October 10, 1965.

The national remainder principle

In December 1964 a new electoral law had introduced the principle of "the national remainder," by which a certain number of seats were distributed to parties according to their proportion of the vote. The law had been intended to operate in favour of the smaller parties and against the JP, but in the election the JP won a surprising majority with 53% of the votes and 240 seats. The RPP received 29% and 134 seats; and the smaller parties, 76 seats. The new JP leader Süleyman Demirel, a former engineer, was able to form a government.

Political moderation had triumphed in the years 1961–65. The army had stood aloof while power came gradually to a party that drew its main support from the same groups and areas as the Democrats and that espoused a similar philosophy. Attempts to restore army rule had failed. Intervention proposed by senior officers in October 1961 had been rejected by others. Two projected coups had been foiled in February 1962 and May 1963. In December 1962, 11 air force officers were removed for their radical views. Members of another secret society within the army—the Young Kemalists—were arrested in April 1963. Criticism of the 1960 revolution was

made illegal in 1962; army leaders contented themselves with occasional warnings.

The trend was toward liberalization—283 Democrat prisoners were released in October 1962, and the remainder in October and November 1964. The landowners and university teachers were allowed to return to their lands and classes. Some army radicals were absorbed into the political system; the secularists went into the RPP, and the Islāmists in July 1965 took over the Republican Peasants' Party (one of the two halves into which the old Republican Peasants' Nation Party had split in 1962; the other section was the Nation Party). For the first time the Marxist-influenced left wing attained a degree of political respectability—the Turkish Worker's Party (TWP), founded in 1961, was taken over by Marxists, in 1962; with a policy of socialism and neutralism, it won 3% of the vote and 15 seats in 1965.

Much of the credit for this peaceful evolution must go to İnönü, who used his prestige with the army to ward off further intervention by the army and to prevent his own party from advocating policies that would have made parliamentary government impossible, all the while permitting power to seep to the JP.

Moderation and compromise had, however, produced uncertainty in handling such major issues as taxation and land reform. Under Atatürk the process of shifting the burden of taxation from peasant to town dweller had begun with the abolition of the agricultural land tax in 1925. Agricultural subsidies had been greatly increased under the DP. In consequence the agricultural sector was large and inefficient, and it made a small contribution to the national revenue. In 1962 the land tax was equal to only 0.2% of the value of the produce. Although a substantial amount of state land had been distributed under the 1945 land reform, most of the large private estates, particularly those in eastern Anatolia, were still intact. The State Planning Organization proposed a graduated land tax; but only a modest agricultural income tax was introduced (1964), and land reform was discussed but not enacted.

JP policy

Political development, 1965–70. The JP's policy emphasized industrialization and welfare legislation. The latter was partly inspired by the need to counter the appeal of left-wing groups, represented in politics by the TWP and elsewhere by more active trade unions and militant students. Beginning in 1960 demonstrations by left-wing students and counter demonstrations by traditionalist and nationalistic student groups became a feature of Turkish life and led to violent clashes. These and other manifestations of violent unrest had produced increasing problems of internal security.

Relations with the RPP were also bitter. The opposition accused the JP of favouring illegal religious organizations, and in fact the JP did depend upon support from religious conservatives and peasants as well as from professional and business elements. Nevertheless, Demirel exercised considerable skill in keeping the JP on the path of moderation and, by his balancing, managed to ward off a second military intervention, although the new president, Gen. Cevdet Sunay, whose term began in March 1966, maintained careful vigilance on behalf of the army. Military threats forced the withdrawal in 1969 of a bill passed by the assembly to restore political rights to the former Democrat leaders. This bill eventually became law in 1969.

While Demirel held the JP factions together in a moderate policy, the RPP moved steadily to a position described as "left of centre." From the Atatürk period it had been an alliance of urban intellectuals, often with official backgrounds, and wealthy landowners from eastern Anatolia. This had made it difficult for the RPP to support radical land reform. In 1966 the party adopted a policy of democratic Socialism, which caused a number of conservatives to leave the RPP and form the Reliance Party (May 1967). Their loss was partly balanced by an agreement between İnönü and Celâl Bayar (May 1969) that attracted some former Democrat support to the RPP.

A new election was held in October 1969. The national remainder system, which had favoured the smaller parties, was dropped and the election was decided by simple proportional representation. The representation of the small parties was thus cut substantially, and the two main parties won more seats with fewer votes. The JP won 47% of the vote and 254 seats, and the RPP 27% and 143 seats; the third largest was the Reliance Party, with 7% and 15 seats. In general the extremist parties of left and right did badly, but neither of the two major parties could be too gratified. The drop in the JP vote was the first swing against the party since 1961. The factional struggle within it increased; some rightists were excluded from the new Cabinet (November 1969). The right retaliated by bringing down the government in February 1970, and the new government formed by Demirel lacked a dependable majority and was in a weak position to deal with the growth in violence. The RPP was also unhappy with the lowest share of the vote in its history.

Exclusion of rightists

Economic and social development, 1960–70. Financial problems and the continued deterioration in the terms of trade made it seem unlikely that the 5% rate of growth achieved under the Democrats could be maintained in 1960. But, in fact, Turkey averaged over 6% throughout the decade. Although some members of the State Planning Organization resigned in 1962–63 because they doubted that resources were available to meet the 1962–67 five-year plan target of 7% a year, the final result of 6.6% was very satisfactory. The next plan (1968–72) had a similar target. Even with the annual rate of population growth (nearly 3% per annum in the 1950s, about 2½% in the 1960s) a substantial growth of per capita income was achieved. Agriculture failed to achieve its target, but important structural changes took place in this sector. The percentage of population dependent upon agriculture fell to 71% (compared with 82% in 1927), and the expansion of cultivation since the early 1950s produced a substantial fall in the ratio of population to cultivated land. Industrial output increased at the annual rate of 9%, to which the private sector contributed more than its target.

Statism in Turkey was strengthened by the establishment of the State Planning Organization and by the favourable attitudes of both the RPP and the JP toward the leading role of the state. The financing of economic development, however, raised problems because of the large amount of outside investment—an Aid Turkey Consortium was established by western European countries (July 31, 1962) and later joined by the U.S. and the World Bank, falling U.S. aid was, to some extent, balanced by new assistance from the U.S.S.R.; another valuable source of foreign exchange was the remittances of Turkish workers in Europe, which became an important feature after 1958 and helped to alleviate the problems of underemployment in Turkey. But as a result Turkey suffered sustained inflation and was forced to devalue in August 1960 and again in August 1970. One important motive for rapid growth was provided by the desire for association with the European Economic Community (EEC). An agreement was negotiated (September 12, 1963) on associate membership, which envisaged full membership in about 25 years.

Rise in literacy

Significant social changes also took place. By 1968 the percentage of literacy had risen to 49%; but because of a relatively low rate of expenditure on education in the 1950s and early 1960s (2–3% of GNP), secondary education was not expanded sufficiently and there was some reversion to illiteracy. With increased urbanization and industrialization, trade unionism developed in the 1950s and grew more rapidly after 1960. Strikes took place for the first time in 1962, and the right to strike was conceded in 1963. A great expansion in the number of government employees took place, rising from 66,000 in 1939 to 246,000 in 1955 and 400,000 in 1960. Whereas older bureaucrats had tended to come from bureaucratic families in western Turkey, many new entrants came from eastern and central Anatolia.

Foreign policy, 1950–70. There were no substantial divisions between the major parties regarding Turkish foreign policy. The 1946 Soviet challenge confirmed the

Turkish inclination toward the West. Early RPP efforts to join the North Atlantic Treaty Organization (NATO) were rewarded under the Democrats in September 1951, although Turkey did not achieve full membership until February 1952. Menderes had already demonstrated firm support for the West by sending a Turkish contingent to assist the UN forces in Korea in 1950. Turkey continued to rest primarily upon NATO throughout the 1950s and gave further demonstrations of its European proclivities by joining the Organization for European Economic Co-operation (OEEC) and the Council of Europe.

Turkey also revived its policy of the 1930s of seeking regional alliances. A friendship agreement with Greece and Yugoslavia (February 1953) blossomed into the subsequent defense (Balkan) pact of August 1954. A mutual aid pact with Pakistan (April 1954) became the forerunner of the Baghdad Pact, inaugurated by the Turko-Iraqi agreement of February 1955, which later included Britain, Iran, and Pakistan. A separate bilateral defense agreement with the United States was signed in March 1959. Relations with the Arab countries before 1958, except for those with Iraq, were not warm. Several differences occurred with Egypt, often arising from Turkey's Western alignment. Hatay remained a source of dispute with Syria, and in 1956 and 1957 Turkish forces took up threatening postures on the borders. At the time of a revolution in Iraq (July 1958) Turkey was apparently dissuaded from armed interference against the government only by U.S. pressures. In 1962 there was considerable friction arising from Iraqi claims that Turkey was giving assistance to the Kurdish rebels in Iraq.

The problem of Cyprus

The most vexing problem was that of Cyprus, arising in the 1950s as a result of agitation by the majority Greek Cypriot community for the end of British rule and for union with Greece. Strife ensued involving the Turkish Cypriot community, which appealed to Ankara for support. This led to friction between Greece and Turkey (September 1955); and anti-Greek riots, resulting in considerable destruction of property, took place in Istanbul, Izmir, and Ankara. An agreement was reached, however, in Zürich and London in February 1959 between Britain, Greece, Turkey, and representatives of the two Cypriot communities and was guaranteed by the three states concerned. This provided for an independent republic of Cyprus, safeguards for the Turkish minority, and small Greek and Turkish garrisons.

Cyprus became independent in 1960 but tensions again developed between the two communities, leading to civil war in December 1963. Public opinion in Turkey supported military intervention on behalf of the Turkish Cypriots, and in August 1964 Turkish aircraft attacked coastal areas of Cyprus. Greek citizens were expelled from Turkey in 1964–65. The U.S. warned Turkey and Greece against intervention, and UN action restored peace. There was still no agreement, however, and in December 1965 the UN again called upon Greece and Turkey to exercise restraint. An attempt by Demirel and the Greek prime minister to reach agreement at a meeting in September 1967 failed. That November Turkey again threatened military intervention, but war was averted by UN and United States pressure, which procured the withdrawal of Greek regular troops that had been introduced into Cyprus.

The Cyprus problem had several ramifications in Turkish politics. Resentment of United States intervention was expressed in riots and attacks on United States property at Adana in March 1966. Left-wing hostility to United States economic and alleged political influence also became more vocal. In this climate and in the context of the general East-West détente Turkey made an attempt to improve its relations with the Soviet Union. Menderes had planned to visit Moscow in July 1960 and, although the 1960 revolution caused a temporary setback, in August 1964 the Turkish foreign minister visited the Soviet Union. This was the first of a series of high-level visits, which led to economic and cultural agreements. The improvement of relations with the Soviet Union was assisted by the softening of Soviet support for Greece over Cyprus and by the desire for closer economic relations following the United States decision to phase out economic aid to Turkey by 1972. Nonetheless, the thaw in Turkey's relations with the Soviet Union represented only a slight shift in Turkey's continued broad identification with the West. (M.E.Y.)

BIBLIOGRAPHY

General studies: J.F. VON HAMMER-PURGSTALL, *Geschichte des osmanischen Reiches*, 10 vol. (1827–35, reprinted 1963); French trans. by J.J. HELLERT, *Histoire de l'Empire Ottoman depuis son origine jusqu'à nos jours*, 18 vol. (1835–43), a classic study of Ottoman history to 1774, based on Turkish source materials; J.W. ZINKEISEN, *Geschichte des osmanischen Reiches in Europa*, 7 vol. (1840–63, reprinted 1962), a study of Ottoman diplomatic and military history in Europe, based on extensive use of European diplomatic archives and travellers' reports; L.S. STAVRIANOS, *The Balkans Since 1453* (1958), concentrates on European and Balkan aspects of Ottoman history, based on study of Western-language materials.

The Ottoman State to 1481: (Origins and expansion of the Ottoman state, 1280–1402): PAUL WITTEK, *The Rise of the Ottoman Empire* (1938, reprinted 1965), a classic study of Ottoman origins in 13th- and early 14th-century Anatolia, emphasizing the importance of the *gazi* tradition; CLAUDE CAHEN, *Pre-Ottoman Turkey: A General Survey of the Material and Spiritual Culture and History, c. 1071–1330* (1968), the most up-to-date study of 13th-century Anatolia, based on extensive research in Turkish and Greek sources, with emphasis on the economic and social background of the rise of the Ottomans; ERNST WERNER, *Die Geburt einer grossmacht—Die Osmanen (1300–1481)* (1966), a general survey, based on examination of Turkish and Western sources; HALIL INALCIK, "Ottoman Methods of Conquest," *Studia Islamica*, 2:103–129 (1954), a description of the development of the vassal system and of toleration of non-Muslim communities as a means of gaining Ottoman conquests in southeastern Europe. (Restoration of the Ottoman Empire, 1402–81): PAUL WITTEK, "De la défaite d'Ankara à la prise de Constantinople," *Revue des Études Islamiques*, 12:1–34 (1938), a description of the Ottoman Interregnum (1402–13), and the means by which the Ottoman Empire was restored in the first half of the 15th century; FRANZ BABINGER, *Mehmed der Eroberer und Seine Zeit: Weltenstürmer einer Zeitenwende*, 2nd ed. (1959), based primarily on European source materials, emphasizing Ottoman relations with Europe under Mehmed II the Conqueror; D.M. VAUGHAN, *Europe and the Turk: A Pattern of Alliances, 1350–1700* (1954), emphasizes diplomatic, economic, and cultural relations between the Ottoman Empire and Europe.

The peak of Ottoman power, 1481–1566: (The Ottoman Empire as the dominant power of southeastern Europe and the Middle East, 1481–1566): S.N. FISHER, *The Foreign Relations of Turkey, 1481–1512* (1948), a discussion of Ottoman relations with Venice, Hungary, the papacy, and Ṣafavid Iran, under Sultan Bayezid II; G.W.F. STRIPLING, *The Ottoman Turks and the Arabs, 1511–1574* (1942, reprinted 1968), on the relations with the Ṣafavid and Mamlūk empires and the conquest of the Arab world under Selim I and Süleyman the Magnificent; F. BRAUDEL, *La méditerranée et le monde méditerranéen à l'époque de Philippe II*, 2nd rev. ed., 2 vol. (1966), a brilliant study of economic problems and development in the Mediterranean area in the mid-16th century, stressing the importance of population problems, the results of influx of precious metals from the new world, and shifts in international trade routes; OMER LUTFI BARKAN, "Les déportations comme méthode de peuplement et de colonisation dans l'Empire Ottoman," *Revue de la Faculté des Sciences Economiques de l'Université d'Istanbul*, 11:67–131 (1949–50), a major study of Ottoman social movements in the 15th and 16th centuries by a leading Turkish economic historian; S.A. FISCHER-GALATI, *Ottoman Imperialism and German Protestantism, 1521–1555* (1959), on the role of the Ottoman threat in the development of the Reformation; G.E. ROTHENBERG, *The Austrian Military Border in Croatia, 1522–1747* (1960), a study of Ottoman-Habsburg military relations; R.C. ANDERSON, *Naval Wars in the Levant, 1558–1853* (1952); HALIL INALCIK, "Capital Formation in the Ottoman Empire," *Journal of Economic History*, 29:97–140 (1969), a fundamental study of internal Ottoman economic organization and development. (Classical Ottoman society and administration): J.F. VON HAMMER-PURGSTALL, *Des Osmanischen Reichs Staatsverfassung und Staatsverwaltung*, 2 vol. (1815, reprinted 1963), a detailed study of Ottoman administrative organization in the 16th century; H.A.R. GIBB and HAROLD BOWEN, *Islamic Society and the West*, 1 vol. in 2 pt.

(1950–57, reprinted 1965), emphasizes Ottoman organization in 18th century, but adds considerable information on earlier periods based on examination of Turkish and Western sources; S.J. SHAW, "The Ottoman View of the Balkans," in C. and B. JELAVICH (eds.), *The Balkans in Transition*, pp. 56–80 (1963), a new interpretation of Ottoman administrative organization, based on an examination of Ottoman archives; S.J. SHAW, *The Financial and Administrative Organization and Development of Ottoman Egypt, 1517–1798* (1962), a detailed study of the Ottoman provincial system, as applied in Egypt, based on exhaustive research in Ottoman archives; summarized in Shaw's "Landholding and Land-tax Revenues in Ottoman Egypt," in P.M. HOLT (ed.), *Political and Social Change in Modern Egypt*, pp. 91–103 (1968); A.D. ALDERSON, *The Structure of the Ottoman Dynasty* (1956), a detailed study of the Ottoman Imperial Institution and the development of the Ottoman dynasty; ROBERT MANTRAN, *Istanbul dans la second moitié du XVII^e siècle* (1962), an exhaustive study of Ottoman political, economic, and social life in the 17th century, based on research in archives and contemporary travellers' accounts; BERNARD LEWIS, *Istanbul and the Civilization of the Ottoman Empire* (1963); RAPHAELA LEWIS, *Everyday Life in Ottoman Turkey* (1971).

Decline of the Ottoman Empire, 1566–1807: PAUL RYCAUT, *The History of the Turkish Empire from the Year 1623 to the Year 1677*, 2 vol. (1680) and *The Present State of the Ottoman Empire* (1668), a contemporary Western description of Ottoman institutions and society in the age of decline; W.L. WRIGHT, JR., *Ottoman Statecraft: The Book of Counsel for Vezirs and Governors . . . of Sari Mehmed Pasha* (1935), a 17th-century Ottoman analysis and description of decline; T.M. BARKER, *Double Eagle and Crescent: Vienna's Second Turkish Siege and its Historical Setting* (1967), a detailed study of the Eastern Question relative to the Ottoman Empire in the late 17th century; LADY MARY WORTLEY MONTAGU, *The Complete Letters of Lady Mary Wortley Montagu*, ed. by ROBERT HALSBAND, 3 vol. (1965–66), an early 18th-century Western account of Ottoman society and institutions in the age of decline; M.L. SHAY, *The Ottoman Empire from 1720 to 1734* (1944, reprinted 1968), a description of Ottoman life during the Tulip Period, based on reports of Venetian consuls in Istanbul; BARON F. DE TOTT, *Memoires du Baron de Tott, sur les Turcs et les Tartares*, 4 pt. (1784; Eng. trans., 2 vol., 1785), an account of Ottoman military organization, problems, and reforms in the late 18th century, written by the most famous of the European military officers in Ottoman service at that time; M.S. ANDERSON, *The Eastern Question, 1774–1923* (1966), an authoritative study of the Eastern Question, relating to the Ottoman Empire, based entirely on Western sources and studies; S.J. SHAW, *Between Old and New: The Ottoman Empire under Sultan Selim III, 1789–1807* (1971), a detailed study of Ottoman reform effort in the late 18th and early 19th century, with an account of diplomatic and military relations with Europe, and problems in the Balkan, Anatolian, and Arab provinces.

The Ottoman Empire and Turkey (from 1807 to the present): The best general history covering this period is B. LEWIS, *The Emergence of Modern Turkey*, 2nd ed. (1968). N. BERKES, *The Development of Secularism in Turkey* (1964), covers similar ground but concentrates on the development of ideas. R.E. WARD and D.A. RUSTOW (eds.), *Political Modernization in Japan and Turkey* (1964), has valuable essays on general themes. For diplomacy the best outline is M.S. ANDERSON, *The Eastern Question, 1774–1923* (1966). Useful for the European provinces is L.S. STAVRIANOS, *The Balkans Since 1453* (1958), which has an excellent bibliography. For the Tanzimat, see R.H. DAVISON, *Reform in the Ottoman Empire, 1856–1876* (1963). Although old, E.P. ENGELHARDT, *La Turquie et Le Tanzimat*, 2 vol. (1882–84), is still a valuable source of information. R. DEVEREUX, *The First Ottoman Constitutional Period: A Study of the Midhat Constitution and Parliament* (1963), is a careful study of the 1876 crisis and the establishment of the first Ottoman Parliament. For the Young Ottomans, see S. MARDIN, *The Genesis of Young Ottoman Thought* (1962); and for the Young Turks, E. RAMSAUR, *The Young Turks: Prelude to the Revolution of 1908* (1957); and F. AHMAD, *The Young Turks: The Committee of Union and Progress in Turkish Politics, 1908–14* (1969). The Ottoman public debt is considered in D.C. BLAISDELL, *European Financial Control in the Ottoman Empire* (1929). AHMED EMIN, *Turkey in the World War* (1930), is still the only account, although relations with Germany can be followed in U. TRUMPENER, *Germany and the Ottoman Empire, 1914–1918* (1968).

Several books consider aspects of Allied war aims in the Near East. Although outdated in parts, H.N. HOWARD, *The Partition of Turkey: A Diplomatic History, 1913–1923* (1931, reprinted 1966), is comprehensive. On the Kemalist movement, see E.D. SMITH, *Origins of the Kemalist Movement and the Government of the Grand National Assembly, 1919–1923* (1959). LORD KINROSS, *Atatürk: The Rebirth of a Nation* (1964), is a good, recent biography. On political developments in Turkey and particularly the rise of the Democrat Party, see K.H. KARPAT, *Turkey's Politics* (1959). R.D. ROBINSON, *The First Turkish Republic* (1963), is a good general account, strongest on economic aspects. F.W. FREY, *The Turkish Political Elite* (1965), is an illuminating and detailed analysis of the membership of the Grand National Assembly. For the 1960 military coup, see W.F. WEIKER, *The Turkish Revolution, 1960–1961* (1963); and for the period from 1961–65, C.H. DODD, *Politics and Government in Turkey* (1969). The chronologies by G. JASCHKE are very useful for Turkish history since 1918. The earlier ones appeared mostly in *Die Welt des Islams* and subsequent ones are *Die Türkei in den jahren 1935–1941* (1943), *Die Türkei in den jahren 1942–1951* (1955), and *Die Türkei in den jahren 1952–1961* (1965). An outline of foreign policy is in ALTEMUR KILIC, *Turkey and the World* (1959). For economic developments, see M.W. THORNBURG, G. SPRY, and G. SOULE, *Turkey: An Economic Appraisal* (1949); and Z.Y. HERSHLAG, *Turkey: The Challenge of Growth* (1968). For education, see A.M. KAZAMIAS, *Education and the Quest for Modernity in Turkey* (1966). There are good bibliographies of works in Turkish in the books by Lewis, Karpat, Dodd, and Hershlag.

(S.J.S./M.E.Y.)

Ou-yang Hsiu

The Chinese poet, historian, and statesman Ou-yang Hsiu (in Pin-yin romanization, Ou-yang Xiu) reintroduced the unadorned "ancient style" in literature and as a statesman sought to reform political life through adherence to Classical Confucian principles. He was born in 1007 in what is today Mien-yang, Szechwan Province, where his father was a judge. Orphaned at three, he and his mother went to live with his uncle in Hupeh. Although the story that the family was so poor that he had to learn writing in the sand with a reed is apocryphal, they probably lived in straitened circumstances. In 1030 he placed first in the doctoral examinations and was appointed a judge at the western capital, Lo-yang. He was already known as a brilliant young writer, and at Lo-yang he befriended the renowned essayist Yen Shu and the famous poet Mei Yao-ch'en. These friendships not only enhanced Ou-yang's status but, more important, reinforced his strong preference for the simplicity and clarity of the "ancient style." Some years before, he had read the works of Han Yü, the great master of T'ang dynasty literature, whose pure and easy "ancient style," free of outworn metaphors and allusions, had greatly impressed him. Eventually, his leadership and advocacy of that style paved the way for a new literary movement.

In 1034 he was appointed a collator of texts in the Imperial library at the capital, K'ai-feng. Two years later, when Fan Chung-yen, a government official, was banished, at the insistence of an Imperial counsellor, for speaking out against certain official practices and institutions, Ou-yang did not hesitate to attack the counsellor in writing. As a result, he, too, was banished and demoted to low judicial office in Hupeh and Hunan provinces, where he wrote the *Hsin Wu Tai shih* ("New History of the Five Dynasties"), a period of political chaos lasting through almost the entire 10th century. Ou-yang's strong sense of fairness led him to devote separate sections to political outcasts such as martyrs, rebels, and traitors, a radical departure from previous dynastic histories.

Writing of the "New History of the Five Dynasties"

Highly recommended by Fan Chung-yen, who was back at the capital, and other high officials, Ou-yang was recalled to the capital in 1043 to become Imperial counsellor. When Fan and others were dismissed for forming a private group of political reformers, Ou-yang, in a notable essay on partisanship, defended associations of gentlemen as politically constructive. His courage and forthright opinions earned the respect of the emperor, Sung Jen Tsung, and he was commissioned to record Jen Tsung's daily life and to draft edicts. His frank opinions and severe criticisms of others created many enemies, and in 1045 he was accused of and tried for having had illicit relations with his niece many years before—a charge to which his romantic life with women and wine,

Career in government

during his days in Lo-yang, lent support. Although he was finally acquitted, his reputation was seriously impaired. He was demoted and sent to Anhwei Province, where he served as magistrate of one country after another. The beautiful countryside intensified his partiality for wine. He called himself the "Old Drunkard," built a pavilion of that name, and wrote an essay about it, which has become one of the most celebrated works in Chinese literature. After a term (1050) as defense commander of the southern capital of Kuei-te, in Honan province, he was recalled to the capital in 1054 to become an academician of the Hanlin Academy. It has been more than nine years since he was exiled from the capital, and the new appointment signified a promotion. As always, his moral courage and outspoken manner did not endear him to his colleagues. He was first ordered to write the *Hsin T'ang shu* ("New History of the T'ang Dynasty"). A year later, with his work only begun, he was sent as ambassador to the Manchurian Khitans, who ruled most of Northern China. In 1057 he was placed in charge of civil service examinations. He favoured those who wrote in the "ancient style" but failed those who employed literary embellishments. For thus imposing his own ideas of literature on the traditional examination system, he was physically attacked by disgruntled candidates. He survived, however, and the literary style championed by him set a new course for Chinese literature. He praised and promoted brilliant young writers such as Wang An-shih and Su Tung-p'o.

Champion of the "ancient style"

When the "New History" was finished in 1060, he was rapidly promoted to the highest councils of state, leaving a remarkable record in social, financial, and military affairs. Although never original in political ideas, he had been an advocate of progressive ideas. Now in his 50s, however, he became more cautious; and this approach probably contributed not a little to these solid accomplishments. But age did not temper his independence, directness, and critical attitude. Eventually his position at court became untenable, and at 60 he was approaching the end of his political career. Someone holding a grudge against him accused him of having an affair with his daughter-in-law, and although the woman's own father came to his defense and the accusation was found to be sheer rumour, Ou-yang's prestige was injured and he became increasingly isolated in the capital. He repeatedly asked to be relieved, but instead the new emperor sent him to be magistrate successively in Anhwei, Shantung, and Honan. In Shantung he opposed the reforms of his former protégé Wang An-shih, particularly a system of loans to farmers at a low interest rate, and refused to carry them out in his districts. Perhaps Ou-yang was opposed not so much to the measures themselves as to the excessive bureaucracy that had become a burden to the people. Nevertheless, it is clear that he had become a disappointed conservative. In 1071 he was retired with the title of grand preceptor of the crown prince. He intended to make his permanent home in beautiful Anhwei, the place of his Old Drunkard Pavilion (Ts'ui Weng T'ing), but within months he died in his 66th year.

His library consisted of 10,000 books and a large collection of literary remains and archaeological records from ancient times. He was honoured posthumously with the title Wen-chung (literary and loyal).

BIBLIOGRAPHY. JAMES T.C. LIU, *Ou-yang Hsiu: An Eleventh-Century Neo-Confucianist* (1967; orig. pub. in Chinese, 1963), an excellent biography; CHANG CARSUN, *The Development of Neo-Confucian Thought*, vol. 1, pp. 91–92, 137–138 (1957), for a brief account of Ou-yang's thought; SHOU-YI CH'EN, *Chinese Literature*, pp. 355–359 (1961), for an appraisal of his literary merits and influence; HERBERT A. GILES, *A History of Chinese Literature*, pp. 212–216 (1923), for excerpts of essays.

(W.-t.C.)

Ovid

One of the greatest poets of classical Rome and second to none in the influence of his poetry on European literature, Ovid is the arch-poet of love, a theme that in his work achieved a new range and significance, being raised in his masterpiece, the *Metamorphoses,* to epic stature. In technical accomplishment he is unexcelled, and the poetical resources of the Latin language were permanently enriched by his craftsmanship.

By courtesy of the Galleria degli Uffizi, Florence

Ovid, marble bust. In the Uffizi, Florence.

Early years. Publius Ovidius Naso was, like most Roman men of letters, a provincial. He was born on March 20 in 43 BC at Sulmo (modern Sulmona), a small town about 90 miles (140 kilometres) east of Rome. More than once in his poetry he refers affectionately to the green and pleasant fields of his native land. His family was old and respectable, sufficiently well-to-do for his father to be able to send him and his elder brother to Rome to be educated. At that time Ovid must have been about 12 years old, having passed rapidly through the primary and secondary stages of his education. At Rome he embarked, under the best teachers of the day, on the third stage, that of the schools of rhetoric.

In these establishments Roman boys studied to achieve absolute mastery of the art of fluent extempore speaking on any subject. For the majority, since philosophy tended to be despised by practical men, this was the final and most important part of their education. The method used was that of formal declamation on prescribed themes, usually of a legal or pseudo-legal character. These themes were based on imaginary situations that were often improbable and sometimes bizarre. In contrast to his brother, whose early death robbed the Roman bar of a promising recruit, Ovid had no liking for the legalistic exercises most in vogue. The elder Seneca (*c.* 55 BC–*c.* AD 37), who has left us a valuable account of Ovid at this period of his life, remarks that he was irked by formal argument and preferred the themes called "ethical," in which the point at issue turned on moral and psychological considerations. There is no doubt that Ovid's time in the schools profoundly affected his poetry. Oratory as there practiced was "rhetorical" in the most extreme sense, dominated by constant striving for effect, by the search for point, paradox, and antithesis and by the ambition to outdo the previous speaker by some fresh and ingenious perversion of probability. The intensely competitive atmosphere was calculated to foster Ovid's natural gift for effective expression and to encourage his interest in the communication of emotion.

As a member of the Roman knightly class (whose rank lay between the commons and the Senate) Ovid was marked by his position, and intended by his father, for an official career. First, however, he spent some time at Athens (then a favourite finishing-school for young men of the upper classes) and travelled in Greek lands with his friend Pompeius Macer, like himself a poet. This experience bore fruit in his poetry, in the form of mytho-

logical associations that are firmly rooted in the classical landscapes. Afterward, Ovid dutifully held some minor judicial posts, the first steps on the official ladder; but he soon decided that public life did not suit him. One legacy of this period would seem to be the predilection for legal phrases and metaphors that he shows in his earlier poems. From now on Ovid abandoned business to cultivate poetry and the society of poets, including his contemporary Sextus Propertius and the somewhat older Horace. Virgil he had not met and Tibullus died before he could get to know him well. It is interesting and may be significant that the warmest tribute Ovid pays to a contemporary poet is to Tibullus, in a funeral elegy on his death. Tibullus, like himself, was a member of the literary circle of the patron Marcus Valerius Messala, whereas Virgil, Horace, and Propertius all belonged to the group around Gaius Maecenas, which was more representative of what might be called the Augustan establishment. Certainly the type of poetry he wrote during the first period of his career reflects a view of life, love, and poetry somewhat at odds with the sober "official" moral attitudes encouraged by Augustus. As a witty and irreverent skeptic whose only declared commitment was to poetry, he could not be expected to regard the official myths of Augustan politics as highly as did, for instance, his older contemporaries Virgil and Horace. They had lived through the chaos of civil war until 31 BC, when Octavian (later named Augustus) won the decisive battle at Actium. Ovid was only 12 when this battle was fought, so he had no emotional reason to share the special gratitude of an older generation toward Augustus for introducing peace.

Publication of his first love elegies

His first poems, the *Amores* (*Loves*), were published at intervals, beginning about 20 BC, in five books (reduced in number to 50 poems, they were afterward republished in a three-book edition). They form a series of short poems in the elegiac metre depicting the various phases of a love affair with a woman called Corinna (almost certainly a figment of the poet's imagination). Inspiration and treatment, however, are more intellectual than emotional, and the collection verges on a tongue-in-cheek mockery of the genre. A poem that was later to play a part in his downfall, the *Ars amatoria* (*Art of Love*), was published *c.* 1 BC. The message of this brilliant treatise on the art of seduction and intrigue was essentially subversive of the official program of moral reforms then being fostered by Augustus, and it cannot have been well received by those who were seriously committed to the goals and aspirations of Augustanism. It also included a number of references, in their contexts both flippant and tactless, to symbols of Augustus' personal prestige. A mock recantation that soon followed, the *Remedia amoris* (*Cure for Love*), cannot have helped matters. During this period Ovid also published his *Epistulae Heroidum* (*Letters from Heroines*), a series of clever, though in sum perhaps monotonous, dramatic soliloquies. While the theme of these character sketches is love, the material is mythological; and here can be seen, in embryo, the *Metamorphoses*.

Maturity: the major poems. As a writer of love elegy, Ovid had spoken in character, claiming the respect and attention due to experience and even revelation. Perhaps the fruits of some wild oats may be detected in the *Amores, Ars amatoria,* and *Remedia amoris,* but the biographer must discount most of Ovid's protestations as properly belonging to a conventional literary pose. In fact Ovid seems to have been, by contemporary standards, a respectable family man. He was first married for a short time, when very young, to a wife of whom he says no more than that she was unsuitable. To his second wife he imputes no blame, but this marriage also was brief. His third union, to a member of an important noble family, was stable and apparently based on mutual affection. We hear also of one daughter, perhaps of the second marriage. Having published a substantial body of highly original work, Ovid was by now established as the leading living poet of Rome (Horace, the last of the great Augustans, had died in 8 BC, and there were no other possible rivals). In the last poem of the *Amores,* Ovid had promised that he would soon turn to more ambitious themes.

Leading poet of Rome

This promise he now redeemed by the publication of work in three major genres.

His tragedy *Medea* has been lost. It was praised by the critic Quintilian and the historian Tacitus and can hardly have failed to influence Seneca's play on the same theme.

The Fasti. Ovid's *Fasti* (*Calendar*) is an account of the Roman year and its religious festivals, consisting of 12 books, one to each month, of which only the first six survive. The various festivals are described as they occur and are traced to their legendary origins. Such "aetiological" poetry was characteristic of poets of the Hellenistic Age (after 323 BC), especially of Callimachus; and in choosing this genre Ovid was deliberately challenging Propertius' claim to be "the Roman Callimachus." Thus the *Fasti* was a national poem, intended to take its place in the Augustan literary program and perhaps designed to rehabilitate its author in the eyes of the ruling dynasty. It contains a good deal of flattery of the imperial family and much perfunctory patriotism, for which the undoubted brilliance of the narrative passages does not atone.

The Metamorphoses. The *Metamorphoses* ("Transformations") must also be interpreted against its contemporary literary and political background. Even Virgil and Horace, the most "Augustan" of the Augustans (both were personal friends of Augustus), had shown reluctance to produce on demand the sort of official poetry that is apt to be expected of a poet laureate—martial epics celebrating important victories and the like. When Virgil did finally write his epic, the *Aeneid,* it was a very different poem from what contemporaries had been led to expect. The *Aeneid* is an astonishing tour de force: a recognizably Augustan poem that is yet something much greater and more universal. This unique character of Virgil's poem, which immediately became canonized as the national epic, posed an acute problem for his successors. The *Aeneid* was both spur and stumbling block; and most subsequent aspirants to epic honours stumbled. After Virgil, a straightforward historical or mythological epic would represent an anticlimax. Ovid was warned against this pitfall alike by his instincts and his intelligence; he chose, as Virgil had done, to write an epic on a new plan, unique and individual to himself.

Themes and rhetoric of the *Metamorphoses*

The *Metamorphoses* is a long poem in 15 books, written (unlike all Ovid's other surviving works) in hexameter verse. It is a collection of mythological and legendary stories in which metamorphosis (transformation) plays some part, however minor. The stories are told in chronological order from the Creation (the first metamorphosis, of chaos into order) to the death and deification of Julius Caesar (the culminating metamorphosis, again of chaos —that is, the Civil Wars—into order—that is, the Augustan Peace). The importance of metamorphosis is more apparent than real; the essential theme of the poem is passion, and this gives it more unity than all the ingenious linking and framing devices the poet uses. The erotic emphasis that had dominated Ovid's earlier poetry has now been broadened and deepened into an exploration of nearly every variety of human emotion—for his gods are nothing if not human. This undertaking brought out, as his earlier work had not, Ovid's full powers: his wit and rhetorical brilliance, his mythological learning, his gifts of narrative and description, and the peculiar qualities of his fertile imagination. Like most classical Latin poetry, the *Metamorphoses* is a highly literary work. The vast quantities of verse in both Greek and Latin that Ovid had read and assimilated are transformed, through a process of creative adaptation, into original and unforeseen guises. The style also represents an individual modification of inherited usage. It is designed to respond unobtrusively to the many variations of genre and tone within the poem and also to carry the reader at an equable and agreeable pace through its considerable length. In both material and treatment the *Metamorphoses* is not merely un-Augustan: it is, language apart, as much Greek as Roman. By AD 8, if not yet formally published, it was complete; and it was at that precise moment, when Ovid seemed securely placed on a pinnacle of successful achievement, that there fell, with stunning suddenness, a blow that shattered his life.

Banishment and death. Ovid was at the time on the island of Elba. He was recalled to Rome for a personal interview with Augustus, in which, after a severe dressing down, he was ordered into immediate banishment at Tomis on the Black Sea. The reasons for his punishment are mentioned several times by Ovid himself: a poem—the *Ars amatoria*—and another unnamed offense. What the latter was he does not specify, beyond insisting that it was an indiscretion (Latin *error*), not a crime. The precise relationship between the two charges is obscure: the *Ars amatoria* had been published several years previously; that it was brought up against its author now seems to imply that it was connected in some way with the more recent *error*. Whatever this was, it was taken as a personal affront to Augustus and his family and bitterly resented as such. This is shown not only by Augustus' own attitude but also by that of his adopted son and successor Tiberius, who paid no attention to Ovid's appeals. Here it is relevant to recall the generally irritant, if not actually subversive, quality of Ovid's poetry. His view of the world was that of an artist, a skeptic, and an individualist. No successful man lacks enemies, as he himself had remarked in the *Remedia amoris*. Not only in the *Ars amatoria* but also in some parts of the *Metamorphoses* is there apparent disrespect for certain officially approved values that might have been enough, skillfully exploited, to create an image of their author as a social and political dissident. If Augustus' mind had already been poisoned against Ovid, or if he had been abruptly reminded of these earlier pinpricks by Ovid's *error*, whether or not it was serious, he may have grasped at the opportunity to rid Rome of a man who, if not dangerous, was personally offensive to him. Augustus was usually ruthless in such cases, but public opinion might have been aroused by the assassination or judicial murder of the leading poet of Rome. A living death in Tomis was perhaps in the event a more severe punishment.

Causes of his banishment

Ovid's situation was curious. The charge on which he was arraigned was one of high treason (Latin *majestas*), which could cover a multitude of sins. He was tried in the Emperor's private court, and the sentence was personally decided and pronounced by the Emperor. In effect he was to be detained at Augustus' pleasure. He was deprived of neither his citizenship nor his property, nor was he forbidden to write poetry or to communicate with his family and friends. Yet his books were removed from the public libraries, and none of the poems sent from Tomis in the first years of his exile, except those to his wife and to Augustus, is addressed to its recipient by name. Clearly, Ovid was unsafe to know; and until the death of Augustus, at least, his poetry must have circulated privately. The mystery that still surrounds the episode itself shows that Augustus was successful in stifling overt comment.

Toward the end of AD 8 Ovid set out for Tomis, where he arrived in the spring of the following year. The place (modern Constanţa, Romania) was a semi-Hellenized port on the extreme edge of the Roman Empire, exposed to periodic attacks by the surrounding barbarian tribes. Books and civilized society were lacking; little Latin was spoken; the climate was severe. From time to time there was actual danger and, with other able-bodied citizens, Ovid took his turn in the home guard. He was alone, for his wife had stayed in Rome to protect his property and to make what intercession she could through influential friends. She was never to see him again. In his solitude and depression, he turned again to poetry, now of a more personal and introspective sort. The *Tristia* (*Sorrows*) and *Epistulae ex Ponto* (*Letters from the Black Sea*) are all, in various guises, appeals for clemency. All, even those to his wife, are "public" poems, intended to present his case to the Emperor and, eventually, to the world. Though not written as spiritual autobiography, they nevertheless offer a unique document of the life of a poet in exile. Properly understood, they are as revealing of their author as anything he ever wrote. Particularly interesting is the apologia that forms the second book of the *Tristia*: a puzzling poem that may have damaged his case more than it helped because, to whatever flattery and self-abasement he descends in this and the other poems of exile, he never withdraws from the one position with which his personal self-respect was identified—his status as a poet. Again and again it is implied or even stated outright that over poetry the Emperor has no power. Critics who find Ovid's conduct in adversity unmanly have failed to see that, in what mattered most to him, he never yielded.

Poems of exile

That his poetical powers were not as yet seriously impaired is shown by his poem *Ibis*. This, written not long after his arrival at Tomis, is a long and elaborate curse directed at an anonymous enemy. It is a tour de force of abstruse mythological learning, composed largely without the aid of books. But in the absence of any sign of encouragement from home, Ovid lacked the heart to continue to write the sort of poetry that had made him famous, and the later *Epistulae ex Ponto* make melancholy reading. Yet life was not uniformly gloomy. As Ovid settled down and became reconciled to the fact that Tomis was his home, he began to see some virtues in the place and its inhabitants. He took an interest in local history and politics and even wrote poetry in the local language, Getic. He seems to have been genuinely touched by the honours which the Tomitans paid him. Nevertheless, death, when it came in AD 17, cannot have been wholly unwelcome.

Character and significance. Often Ovid has been patronized and pitied. The key to an understanding of his character and of his poetry is that he was a rationalist and an extremely intelligent one; he had, for instance, an instinctive sympathy with the materialist poet Lucretius. He was too skeptical and too intellectually independent to devote himself to any cause except that of poetry. That was the faith in which he lived and died and which pervades his work, from the *Amores* to the poems of exile. To poetry his devotion was absolute; his sensuous appreciation of words and his delight in manipulating language is that of a lover. With these qualities went a rich vein of fantasy and an exuberant creativity. His understanding of human nature, if less profound than Virgil's, was wider and perhaps to ordinary people more appealing and comprehensible. He was a kind friend, a humorous and understanding lover, but above all a man of letters, a creator and craftsman—in the fullest and most exact sense of the word, a poet.

Influence on later poetry

In classical antiquity, Ovid's influence on later poetry was primarily technical. He had perfected both the elegiac couplet and the hexameter as all-purpose metres and as instruments of fluent communication. Even the poets most obviously indebted to Virgil betray Ovid's influence in nearly every line. In the Middle Ages, like most ancient writers, he was regarded as an authority, a source of doctrine and erudition. The *Metamorphoses*, in particular, offered one of the most accessible and attractive avenues to the riches of Greek mythology. But his chief appeal, then and subsequently, stems from the humanity of his writing: its gaiety, its sympathy, its exuberance, its pictorial and sensuous quality. It is these things, together with his apparently genuine liking for women as a sex, that have recommended him, down the ages, to the troubadours and the poets of courtly love, to Chaucer, Shakespeare, Goethe, and to Ezra Pound.

MAJOR WORKS

Amores (published at intervals beginning *c.* 20 BC); *Ars amatoria* (*c.* 1 BC); *Medea* (now lost); *Fasti* (unfinished; AD 1–8); *Metamorphoses* (AD 1–8); *Tristia, Epistulae ex Ponto* (both AD 9–17).

BIBLIOGRAPHY. For current work, see *L'Année Philologique*, ed. by J. MAROUZEAU and J. ERNST (annual); and the periodical surveys: R.J. GARIEPY, "Recent Scholarship on Ovid (1958–1968)," *The Classical World*, 64:37–56 (1970); and WALTHER KRAUS, "Der Forschungskericht: Ovid," *Anzeiger für die Altertumswissenschaft*, 11:129–146 (1958), 16:1–13 (1963), and 18:193–207 (1965), which also provide access to the older literature. For Ovid's posthumous survival, see WILFRIED STROH (comp.), *Ovid im Urteil der Nachwelt* (1969). The standard texts of Ovid's works are those in the "Teubner Series" by RUDOLF EHWALD and F.W. LENZ (1888–1932); in J.P. POSTGATE, *Corpus poetarum Latinorum*, vol. 1 by ARTHUR PALMER *et al.* (1894); in the "Oxford Classical

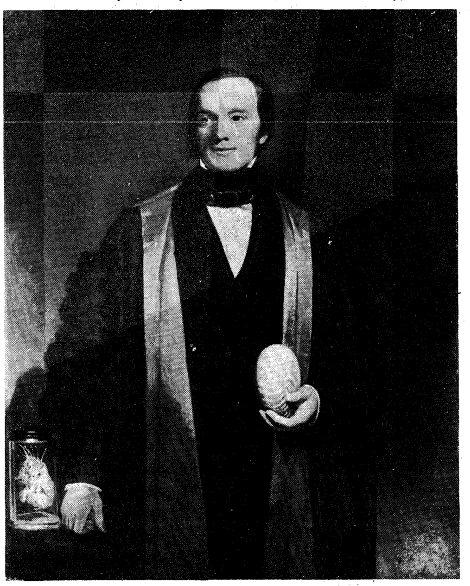
Texts," by S.G. OWEN *et al.* (1915–); and (with English translation) in the "Loeb Series" by GRANT SHOWERMAN *et al.*, 5 vol. (1914–31). The earliest English translation of Ovid is that of the *Metamorphoses* by W. CAXTON (facsimile ed. 1968). Other older versions include the *Metamorphoses* by A. GOLDING, "Shakespeare's Ovid" (1567; new ed. by J.F. NIMS, 1965); of the *Elegies* (*Amores*) by CHRISTOPHER MARLOWE (1596); and of the *Epistles* (*Heroides*), *Ars amatoria*, and *Metamorphoses* by JOHN DRYDEN, ALEXANDER POPE, WILLIAM CONGREVE *et al.* (1680–1717). The numerous modern versions vary greatly in style: see, for instance, A.E. WATTS, *Metamorphoses* (1954), with etchings by Picasso; PAUL TURNER, *Ars amatoria* (1968), prose; GUY LEE, *Amores* (1968). There are many translations into other European languages. The chief source of biographical data is Ovid's poetry, especially his autobiographical elegy (*Tristia* IV. 10). The best general studies are L.P.WILKINSON, *Ovid Recalled* (1955); and HERMANN FRANKEL, *Ovid: A Poet Between Two Worlds* (1945). On Ovid's banishment, see J.C. THIBAULT, *The Mystery of Ovid's Exile* (1964). On the *Metamorphoses*, see BROOKS OTIS, *Ovid As an Epic Poet*, 2nd ed. (1970).

(E.J.Ke.)

Owen, Sir Richard

A British anatomist and paleontologist, Sir Richard Owen is remembered now for his many significant publications, his reconstructions of giant flightless birds, his early recognition of the difference between "homology" (fundamental agreement in structure between parts of different animals regardless of their function) and "analogy" (parts resembling one another only in function), his skill as a lecturer, and his extreme hostility to Charles Darwin and the concept of evolution by natural selection. He was the last British speculative transcendental anatomist who held that the Platonic "idea" of an object was more real than the object itself.

By courtesy of the National Portrait Gallery, London



Owen, oil painting by H.W. Pickersgill, 1845. In the National Portrait Gallery, London.

Richard Owen was born at Lancaster, Lancashire, on July 20, 1804, the son of Richard Owen, merchant in the West Indies trade, and Catherine (*née* Parrin), of a French Huguenot family. Educated at Lancaster Grammar School, he was apprenticed in 1820 to a group of Lancaster surgeons but in 1824 went to Edinburgh to continue medical training. In 1825 he transferred to St. Bartholomew's Hospital in London to work under John Abernethy, the great surgeon, and in the following year was admitted to the Royal College of Surgeons of England, where he was engaged as curator of the Hunterian Collections (made by John Hunter, the great anatomist) and set up in medical practice. In 1830 he met Georges Cuvier, a celebrated French paleontologist, and the following year visited him in Paris, where he studied specimens in the Muséum d'Histoire Naturelle. Elected a Fellow of the Royal Society in 1834, in 1835 he married Caroline Clift (who died in 1873) and had one son, William (1837–86). In 1836 Owen became Hunterian Professor at the Royal College of Surgeons and in 1837 its Professor of Anatomy and Physiology, as well as Fullerian Professor of Comparative Anatomy and Physiology at the Royal Institution. Leaving medical practice and devoting himself to research, he was appointed superintendent of the natural history departments of the British Museum in 1856. From then until his retirement in 1884 he was largely occupied with the development of the British Museum (Natural History) in South Kensington, London.

Attitude to Darwin

Owen's character was not free from ambition for priority, intolerance of competition, and jealousy. This showed itself particularly in the case of the publication by Darwin in 1859 of *Origin of Species*. Darwin had been a good friend for 20 years, but Owen, recognizing that his own pre-eminent position in biology would be eclipsed by Darwin, set about to discredit him. He wrote a very long anonymous review of the book (*The Edinburgh Review*, 1860), on which Darwin commented:

> It is extremely malignant, clever, and I fear will be very damaging. . . . It requires much study to appreciate all the bitter spite of many of the remarks against me. . . . He misquotes some passages, altering words within inverted commas. . . .

As Darwin's thesis began to become more widely known and accepted, it seems that Owen shifted his position somewhat. The *London Review* (no. 12, 1866) apparently received a "communication" from him in which, although he denied the Darwinian doctrine, he admitted the accuracy of its basis, claiming to have been the first to have pointed out the truth of the principle on which it was founded.

On retirement he was created a knight of the Order of the Bath. He died on December 18, 1892, at Sheen Lodge, Richmond Park, London.

Major areas of study

Among his earliest publications were the Descriptive and Illustrated Catalogue of the physiological series of Comparative Anatomy contained in the Museum of the Royal College of Surgeons in London (1833), which enabled him to acquire the great knowledge of comparative anatomy that he used in his numerous researches on living and fossil forms. His *Memoir on the Pearly Nautilus* (1832) was a classic. Other works included *Odontography* (1840–45), *Lectures on Comparative Anatomy and Physiology of the Vertebrate Animals* (1846), *A History of British Fossil Mammals and Birds* (1846), *A History of British Fossil Reptiles* (1849–84), and *On the Anatomy of Vertebrates* (1866–68).

In 1863 he published a description of *Archaeopteryx*, the first known fossil bird (*Philos. Trans. Roy. Soc. London*, pp. 33–47) that he had obtained for the museum; but when this was re-examined in 1954, it was found that Owen had got it upside down, dorsal for ventral, and had missed the two most important features: the breastbone, which was flat, proof that the bird could not fly but glided; and the natural cast of the brain case, which was like that of a reptile. The high reputation that Owen enjoyed needs to be considered critically and objectively. Actually, as early as 1858 his transcendental views on anatomy, particularly the view that the vertebrate skull was only modified vertebrae, was completely demolished by the English biologist Thomas Huxley.

BIBLIOGRAPHY. The only biography of the subject is *The Life of Richard Owen*, 2 vol. (1894), written by his grandson, RICHARD OWEN. It contains Owen's full bibliography.

(G.de B.)

Owen, Robert

Robert Owen, a Welsh manufacturer turned reformer, became one of the most influential of the utopian Socialists of the early 19th century. Friedrich Engels called him "a man of almost sublime and childlike simplicity of character, and at the same time one of the few born leaders of men."

Early life. Owen was born on May 14, 1771, at Newtown, Montgomeryshire. He attended local schools until

Robert Owen, watercolour by A. Hervieu, 1829. In the National Portrait Gallery, London.
By courtesy of the National Portrait Gallery, London

the age of ten, when he became an apprentice to a clothier. His employer had a good library, and young Owen spent much of his time reading. He did so well in business that by the time he was 19 he had become superintendent of a large cotton mill in Manchester, which he soon had made one of the foremost establishments of its kind in Great Britain. Owen made use of the first American sea island cotton (a fine, long-staple fibre) ever imported into the country and made improvements in the quality of the cotton spun. On becoming manager and a partner in the Manchester firm, Owen induced his partners to purchase the New Lanark mills in Lanarkshire.

The success at New Lanark

In the town of New Lanark lived 2,000 people, 500 of whom were young children from the poorhouses and charities of Edinburgh and Glasgow. The children, especially, had been well treated by the former proprietor, but the condition of the people was unsatisfactory. Crime and vice bred by demoralizing conditions were common, education and sanitation alike were neglected, and housing conditions were intolerable. Owen improved the houses and, mainly by his own personal influence, encouraged the people in habits of order, cleanliness, and thrift. He opened a store at which goods of sound quality could be bought at little more than cost price and at which the sale of alcoholic beverages was placed under strict supervision. His greatest success, however, was in the education of the young, to which he devoted special attention. In 1816 he opened the first infant school in Great Britain at the New Lanark mills and thereafter gave it his close personal supervision.

Though Owen at first was regarded with suspicion as an outsider, he soon won the confidence of the people. The mills continued to thrive commercially, but some of Owen's schemes entailed considerable expense, which displeased his partners. Finally, frustrated by the restrictions imposed on him by his partners, who wished to conduct the business along more ordinary lines, he organized a new firm in 1813. Its members, content with a 5 percent return on their capital and ready to give freer scope to his philanthropy, bought out the old firm. Stockholders in the new firm included the legal reformer Jeremy Bentham and the Quaker William Allen.

The reformer. In the same year (1813) Owen published two of the four essays in *A New View of Society, or Essays on the Principle of the Formation of the Human Character*, in which he expounded the principles on which his system of educational philanthropy was based. Having at an early age lost all belief in the prevailing forms of religion, he had thought out for himself a creed that he took to be an entirely new and original discovery.

The chief points in Owen's philosophy were that man's character is formed by circumstances over which he has no control and that he is not a proper subject either of praise or of blame. These convictions led him to the conclusion that the great secret in the right formation of man's character is to place him under the proper influences from his earliest years. The irresponsibility of man and the effect of early influences are the keynote of Owen's whole system of education and social amelioration.

New Lanark as a Mecca for social reformers

For the next few years, Owen's work in New Lanark was to have a national, even European, significance. New Lanark became a place of pilgrimage for social reformers, statesmen, and royal personages. According to the unanimous testimony of all who visited it, the results achieved by Owen were singularly good. Children brought up on his system were generally felt to be graceful, genial, and unconstrained; health, plenty, and relative contentment prevailed. The business was a commercial success.

In 1815 Owen, apparently single-handedly, started agitation for factory reform, with only little effect, and by 1817 his work as a practical reformer had given way to ideas—still vital—that were to make him the forerunner of Socialism and the cooperative movement. Owen argued that the competition of human labour with machinery was a permanent cause of distress and held that the only effective remedy lay in the united action of men and the subordination of machinery to man. His proposals for the treatment of pauperism were based on those principles.

Owen recommended that villages of "unity and cooperation" be established for the unemployed. Each village would consist of about 1,200 persons living on 1,000 to 1,500 acres (400 to 600 hectares); all would live in one large structure built in the form of a square, with public kitchen and messrooms. Each family would have its own private apartment and the entire care of the children till the age of three, after which they would be raised by the community. Parents would have access to them at meals and all other proper times. Such communities, Owen believed, might be established by individuals, by parishes, by counties, or by the state; in each case there would be supervision by duly qualified persons. Work and the enjoyment of its results would be shared in common.

The size of the projected community had been suggested by that of the village of New Lanark, and Owen soon advocated an extension of the scheme to the reorganization of society in general. Under his plan, largely self-contained communities of from 500 to 3,000 would first be set up, mainly agricultural and possessing the most modern machinery. As they increased in number, he wrote, "unions of them, federatively united, should be formed in circles of tens, hundreds, and thousands," until they embraced the whole world in a common interest.

Owen's plans for the cure of pauperism were received with considerable favour until, at a large meeting in London, Owen declared his hostility to the received forms of religion. Many of his supporters believed that this action made him suspect to the upper classes, though he did not lose all support from them. To carry out his plan for the creation of self-contained communities, in 1825 he bought 30,000 acres of land in Indiana from a religious community and renamed it New Harmony. For a time, life in the community was well ordered and contented under Owen's practical guidance, but differences about the form of government and the role of religion soon appeared, and numerous attempts at reconstruction failed to compose them, though it is agreed that an admirable spirit prevailed amid all the dissensions. Owen withdrew from the community in 1828, having lost £40,000—80 percent of his fortune. The other chief Owenite community experiments were in Great Britain—at Queenwood, Hampshire (1839–45), in which Owen took part for three years; at Orbiston, near Glasgow, Lanarkshire (1826–27); and at Ralahine, County Cork (1831–33); he was not directly concerned with either of the later ones.

Community in New Harmony, Indiana

Work with unions. In his "Report to the County of Lanark" (a body of landowners) in 1820, Owen declared that reform was not enough, that a transformation of the social order was required. His proposals for communities attracted the younger workers, brought up under the factory system, and between 1820 and 1830 numerous societies were formed and journals organized to advocate his views. The growth of labour unionism and the emergence

of a working class point of view caused Owen's doctrines to be accepted as an expression of the workers' aspirations, and when he returned to England from New Harmony in 1829 he found himself regarded as their leader. In the unions, Owenism stimulated the formation of self-governing workshops. The need for a market for the products of such shops led to the formation of the National Equitable Labour Exchange in 1832, applying the principle that labour is the source of all wealth.

The unprecedented growth of labour unions made it seem possible that the separate industries and eventually all industry might be organized by these bodies. Owen and his followers carried on ardent propaganda all over the country, with the result that the new National Operative Builders Union turned itself into a guild to carry on the building industry, and the Grand National Consolidated Trades Union was formed (1834). Though the enthusiasm of the unions and the numbers of labourers joining them were remarkable, determined opposition by employers and severe repression by the government and law courts ended the movement within a few months. It was two generations before Socialism, first popularly discussed at this time, again influenced unionism. Throughout these years Owen's community ideas maintained a hold, and ultimately they provided the basis for the worldwide consumers' cooperative movement. After 1834 Owen devoted himself to preaching his educational, moral, rationalist, and marriage-reform ideas. At the age of 82 he became a spiritualist. He died six years later, on November 17, 1858, at Newtown.

Opposition to Owenism

Owen's influence. The great-hearted cotton manufacturer who made such a strong impression on those who knew him failed, despite his arduous labours, to achieve a lasting effect on the industrial world of the 19th century. He was most successful in his early years with the experiment at New Lanark. When he left the life of the factory and set up as a reformer, he exaggerated the degree to which society and individuals can be changed by persuasion. It is said that, when Ralph Waldo Emerson asked Owen in his old age, "Who is your disciple? How many men possessed of your views, who will remain after you, are going to put them in practice?" Owen candidly replied, "Not one."

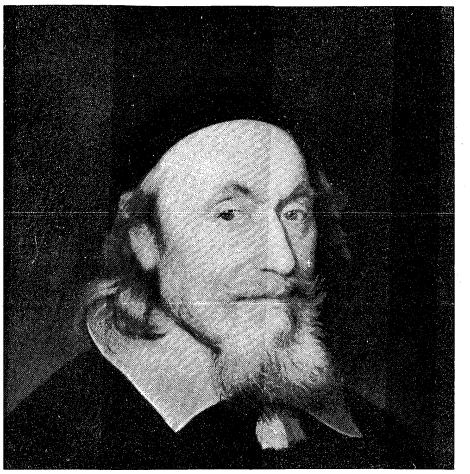
BIBLIOGRAPHY. Owen wrote his own biography, *Life of Robert Owen Written by Himself*, 2 vol. (1857–58). Of his numerous works expounding his ideas, the most important include *A New View of Society, and Other Writings* (1813–21, reprinted 1927); *Two Memorials on Behalf of the Working Classes* (1818); and *Lectures on an Entire New State of Society* (1830). Two lives of Owen are F. PODMORE, *Robert Owen: A Biography* (1906); and G.D.H. COLE, *The Life of Robert Owen*, 3rd ed. (1965). Other works dealing with Owen and his ideas are LEONARD WOOLF, *Co-operation and the Future of Industry* (1919); R.W. LEOPOLD, *Robert Dale Owen* (1940); MARGARET COLE, *Robert Owen of New Lanark* (1953); and J.F.C. HARRISON, *Robert Owen and the Owenites in Britain and America* (1969).

(D.F.Do.)

Oxenstierna, Axel, Count

Chancellor of Sweden for 42 years, Count Oxenstierna was one of his country's greatest public servants. Although a champion of the privileges of the nobility against monarchy, he became the close friend and adviser of King Gustavus II Adolphus (*q.v.*). After the King's death he guided Sweden through the Thirty Years' War and was virtual ruler of the country during the minority of Queen Christina.

Oxenstierna was born near Uppsala on June 16, 1583, of a noble family that had played a considerable part in Sweden's history. After receiving his education at Rostock and other German universities, he was in 1605 appointed to a post in the exchequer; in June 1609, at the early age of 24, he was made a member of the council of state. He soon established an ascendancy in that body, and on the death of Charles IX in 1611, it was he who extorted from the new king, Gustavus II Adolphus, a charter guaranteeing the nation against the royal abuses that had latterly prevailed. One of Gustavus' first acts was to appoint Oxenstierna chancellor (January 1612).



Oxenstierna, detail of an oil painting by David Beck (1621–56). In the Nationalmuseum, Stockholm.
By courtesy of the Nationalmuseum, Stockholm

Oxenstierna had emerged as the champion of the aristocracy against the violence of the monarchy; and the charter might well have initiated a constitutional struggle if strong ties of respect and affection had not soon developed between the King and Chancellor. They became, indeed, ideal collaborators and share the credit for the achievements of the reign. Oxenstierna's contributions were in the spheres of administrative reform and diplomacy. He drafted the *riksdagsordning* ("parliamentary law") of 1617, which stabilized the constitution of the Riksdag; he drew up the ordinance of 1619 on the development of the towns; he carried through a reform in local government in 1623; and he issued a chancery ordinance in 1626 that organized the business of that office. He was mainly responsible for the building of the house of the nobility in Stockholm and for the *riddarhusordning* ("upper house law"; 1626), which divided the nobility into three classes and specified the members of each. As a diplomat he was entrusted with a long succession of major negotiations: the Peace of Knäred (with Denmark, 1613), the Truce of Ogra (with Poland, 1622), the negotiations with Denmark at Sjöaryd (1624). When Gustavus transferred his war against Poland to Prussia in 1626, Oxenstierna was brought over and installed as governor general, and it was he who negotiated the advantageous Truce of Altmark with Poland in 1629. In November 1631 the King called him to Germany (see THIRTY YEARS' WAR).

Disagreements with Gustavus

Oxenstierna had been more reluctant than Gustavus to intervene in Germany and would probably have preferred, in the first instance, a final settlement with Denmark—always, in his view, Sweden's main enemy. Moreover, he disliked the French alliance; considered that Gustavus made a capital error in not marching on Vienna after the Battle of Breitenfeld; disapproved the King's candidature for the Polish throne in 1632; and tacitly opposed the project for marrying Christina to the electoral prince of Brandenburg, Frederick William. His removal to Germany placed the main burden of Swedish diplomacy again on his shoulders; but the King also now entrusted him with military commands, such as the formation and leadership of the army that relieved Gustavus at Nürnberg in August 1632.

The death of Gustavus, in November 1632, put the supreme direction of the Swedish cause in Germany into Oxenstierna's hands. Preserving to himself much of the king's authority and prestige, he negotiated with electors as an equal; and the project of making him elector of Mainz was canvassed. In the League of Heilbronn (1633), he created a *corpus evangelicorum* of the kind that Gustavus had planned, with himself as its director; but he never managed to persuade the North German princes to join it. The disaster at Nördlingen (1634) destroyed his hopes of keeping Sweden's allies loyal, and many of them made peace at Prague in 1635. In the same year the regency at home, in an access of defeatism that

provoked Oxenstierna's indignation, purchased renewal of the truce with Poland by sacrificing the tolls that the Swedes had been levying in the Prussian harbours from 1627. Faced with overwhelming financial and military difficulties, he was for a time the virtual prisoner of his mutinous unpaid soldiery. He hesitated long between buying a peace on the best terms that he could get and accepting a French alliance and indefinite continuation of the war: it was not until 1638 that the intransigence of the Holy Roman emperor, Ferdinand III, forced him to the second alternative. Thenceforward he was the strongest advocate of fighting on until peace could be had on really favourable terms. Political enemies at home accused him of prolonging the war for his own ends; but the terms obtained by Sweden under the Peace of Westphalia, in 1648, justified his obstinacy. Meanwhile, he had launched the sudden attack on Denmark in 1643: the morality of it was dubious, but at Brömsebro in 1645 he was able to dictate a peace that wiped out the humiliations suffered at Knäred in 1613.

Peace of Westphalia

As chancellor, Oxenstierna was one of the five regents who were to govern Sweden during Christina's minority; he also drew up (probably with Gustavus' approval) the *regeringsform* ("form of government") accepted by the Riksdag in 1634. It was not until his return to Sweden in 1636 that he participated in the regency's government; but for the next eight years or so he was the real ruler of the country. His relations with the Queen, after she attained her majority (1644), were never as cordial as with her father: she saw in him the leader of an aristocracy anxious to limit the crown's powers and perhaps even to set up a republic. Others disliked him as the defender of noble privileges and noble encroachments on the liberties of the peasantry, though he repeatedly urged moderation on his colleagues in this regard. He had to face the intrigues of a hostile party at court; and he clashed with Christina on foreign and ecclesiastical policy, on the question of the succession, and on her proposal to abdicate. After 1650, however, relations improved; and he was as firmly established in office as at any time in his career when he died, in Stockholm, on August 28, 1654.

Sagacious, imperturbable, courageous, and industrious, unhurried in negotiation, and not without a pungent humour, Oxenstierna felt the service of the state to be equally congenial and obligatory. Though he never forgot the interests of his class, it may fairly be said that as a rule his single-minded patriotism transcended them.

BIBLIOGRAPHY. No full biography of Oxenstierna exists. NILS AHNLUND's magisterial monograph, *Axel Oxenstierna intill Gustav Adolfs död* (1940), stops short at 1632. WILHELM THAM, *Axel Oxenstierna: Hans ungdom och verksamhet intill år 1612* (1935), is indispensable for the early years. Oxenstierna's enormous correspondence has been in course of publication since 1888. The first series has reached volume 15 and the year 1636; of the second, 12 volumes have so far appeared—*Rikskansleren Axel Oxenstiernas skrifter och brevväxling* (1888–).

(M.Ro.)

Oxidation–Reduction Reactions

In the classification of chemical reactions, the terms oxidation and reduction cover a large and diverse body of processes. Many oxidation–reduction reactions are as common and familiar as fire, the rusting and dissolution of metals, the browning of fruit, and respiration and photosynthesis—basic life functions. To understand the details of oxidation–reduction, a wider knowledge of chemical reactions in general is required than can be provided in this article. For this background, see CHEMICAL REACTIONS; CHEMICAL BONDING; PERIODIC LAW.

GENERAL CONSIDERATIONS OF REDOX REACTIONS

Characteristics of redox processes

Major classifications. Most oxidation–reduction (redox) processes involve the transfer of oxygen atoms, hydrogen atoms, or electrons, with all three processes sharing two important characteristics: (1) they are coupled —*i.e.*, in any oxidation reaction a reciprocal reduction occurs; (2) they involve a characteristic net chemical change—*i.e.*, an atom or electron goes from one unit of matter to another. Both reciprocity and net change are illustrated below in examples of the three most common types of oxidation–reduction reactions.

Oxygen-atom transfer. Carbon reacts with mercury(II) oxide (a compound in which mercury has a bonding capacity expressed as +2; see below *Oxidation-state change*) to produce carbon dioxide and mercury metal. This reaction can be written in equation form:

$$\underset{\text{carbon}}{C} + \underset{\substack{\text{mercury(II)}\\\text{oxide}}}{2HgO} \longrightarrow \underset{\substack{\text{carbon}\\\text{dioxide}}}{CO_2} + \underset{\text{metal}}{Hg}\ .$$

Carbon, receiving oxygen, is oxidized; mercury(II) oxide, losing oxygen, undergoes the complementary reduction; and the net change is the transfer of two oxygen atoms from mercury(II) oxide units to a carbon atom.

Hydrogen-atom transfer. Hydrogen atoms are transferred from hydrazine, a compound of nitrogen and hydrogen, to oxygen in the following reaction:

$$\underset{\text{hydrazine}}{N_2H_4} + \underset{\text{oxygen}}{O_2} \longrightarrow \underset{\text{nitrogen}}{N_2} + \underset{\text{water}}{2H_2O}\ .$$

Hydrazine, losing hydrogen, is oxidized to molecular nitrogen, while oxygen, gaining hydrogen, is reduced to water.

Electron transfer. Zinc metal and copper(II) ion (an ion is any atom that has lost or gained one or more electrons and, therefore, has a positive or negative electrical charge) react in water solution, producing copper metal and an aqueous (denoted by aq) zinc ion according to the equation:

$$\underset{\substack{\text{zinc}\\\text{metal}}}{Zn} + \underset{\substack{\text{aqueous}\\\text{copper(II)}\\\text{ion}}}{Cu^{2+}(aq)} \longrightarrow \underset{\substack{\text{aqueous}\\\text{zinc(II)}\\\text{ion}}}{Zn^{2+}(aq)} + \underset{\substack{\text{copper}\\\text{metal}}}{Cu}\ .$$

With the transfer of two of its electrons, the zinc metal is oxidized, becoming an aqueous zinc ion, while the copper(II) ion, gaining electrons, is reduced to copper metal. Net change is the transfer of two electrons, lost by zinc and acquired by copper.

Because of their complementary nature, the oxidation and reduction processes together are referred to as redox reactions. The reactant that brings about the oxidation is called the oxidizing agent, and that reagent is itself reduced by the reducing agent. In the examples given, mercury(II) oxide, oxygen, and the copper(II) ion are oxidizing agents, and carbon, hydrazine, and zinc are the reducing agents.

General theory. *Stoichiometric basis.* Describing the redox processes as above conveys no information about the mechanism by which change takes place. A complete description of the net chemical change for a process is known as the stoichiometry of the reaction, which provides the characteristic combining proportions of elements and compounds. (Stoichiometry [from the Greek *stoicheion*, "component," and *metron*, "measure"] deals with methods of determining the characteristic combining proportions of reactions from formulas, equations, atomic weights, and molecular weights and with the practical use of the information to determine quickly and easily what and how much is used and produced in chemical processes.) Reactions are classified as redox and nonredox on the basis of stoichiometry; oxygen-atom, hydrogen-atom, and electron transfer are stoichiometric categories.

Oxidation-state change. Comprehensive definitions of oxidation and reduction have been made possible by modern molecular structure theory. Every atom consists of a positive nucleus, surrounded by negative electrons, which determine the bonding characteristics of each element. In forming chemical bonds, atoms donate, acquire, or share electrons. This makes it possible to assign every atom an oxidation number, which specifies the number of its electrons that can be involved in forming bonds with other atoms. From the particular atoms in a molecule and their known bonding capacities, the bonding pattern within a molecule is determined, and each atom is regarded as being in a specific oxidation state, expressed by an oxidation number.

Redox processes defined

Redox processes are defined as reactions accompanied by oxidation-state changes: an increase in an atom's oxidation number corresponds to an oxidation; a decrease, to a reduction. In this generalized theory, three examples of ways in which oxidation-state changes can occur are by oxygen-atom (gain, oxidation; loss, reduction), hydrogen-atom (loss, oxidation; gain, reduction), and electron (loss, oxidation; gain, reduction) transfer. The oxidation-state change definition is usually compatible with the above rules for applying the oxygen-atom-transfer and hydrogen-atom-transfer criteria and always compatible with the electron-transfer criterion when it is applicable. The oxidation state of any atom is indicated by a roman numeral following the name or symbol for the element. Thus: iron(III), or Fe(III), means iron in an oxidation state of +3. The uncombined Fe(III) ion is simply Fe^{3+}.

Historical origins of the redox concept. Of the chemical processes now regarded as redox reactions, combustion was the earliest focus of philosophical and scientific attention. The Greek scientific philosopher Empedocles listed fire as one of the four elements of matter. In more modern times the phlogiston theory enjoyed scientific popularity. First articulated in 1697 by Georg Ernst Stahl, a German physician, the theory asserted that matter releases an elementary constituent, phlogiston, during combustion. Thus, the burning of charcoal was interpreted as the loss of phlogiston from carbon to the air. The theory was also applied to processes other than combustion; in the recovery of a metal from its oxide by heating with charcoal, for example, phlogiston was regarded as being transferred from carbon to the oxide.

Phlogiston saturation was believed to be responsible for the limited ability of air in a closed container to support combustion. A notable consequence of the phlogiston theory was the view that an oxide of a metal, such as mercury(II) oxide (HgO), was a chemically simpler substance than the metal itself: the metal could be obtained from the oxide only by the addition of phlogiston. The theory, however, could provide no acceptable explanation of the gain in weight when an oxide is formed from a metal.

The work of Priestley and Lavoisier

Combustion and oxide formation. Late in the 18th century, the interrelated work of the English chemist Joseph Priestley and his French colleague Antoine Lavoisier led to the overthrow of the phlogiston theory. Lavoisier saw Priestley's discovery of oxygen in 1774 as the key to the weight gains known to accompany the burning of sulfur and phosphorus and the calcination of metals (oxide formation). In his *Traité élémentaire de chimie* (*Elements of Chemistry*), which set forth the systematic foundations of modern chemistry, Lavoisier clearly established that combustion consists of a chemical combination between oxygen from the atmosphere and combustible matter. By the end of the century, his ideas were widely accepted and had been successfully applied to the more complex processes of respiration and photosynthesis. Reactions in which oxygen was consumed were classified as oxidations, while those in which oxygen was lost termed reductions.

Electrochemical reactions. During the 19th century, the evolving field of electrochemistry led to a broadened view of oxidation. It was possible, for instance, to produce the ferric, or iron(III), ion from the ferrous, or iron(II), ion at the anode (positive electrode, where electrons are absorbed from solution) of an electrochemical cell (a device in which chemical energy is converted to electrical energy), according to the equation:

$$Fe^{2+}(aq) \longrightarrow Fe^{3+}(aq) + e^-$$

iron(II) ion — iron(III) ion — electron

Molecular oxygen could effect a similar transformation, according to the equation:

$$4Fe^{2+}(aq) + O_2 + 4H^+ \longrightarrow 4Fe^{3+}(aq) + 2H_2O$$

iron(II) ion — oxygen — hydrogen ion — iron(III) ion — water

The similarity of the two processes led to a precursor of the electron-transfer explanation for redox reactions. After the discovery of the electron, the conviction that oxidation and reduction are accomplished through electron loss and gain became firmly entrenched. Thus, early in the 20th century chemists tended to attribute all redox reactions to the transfer of electrons. Later work on chemical bonding, however, demonstrated the incorrectness of that description. An electronegativity scale (listing of elements in descending order of their tendency to attract and hold bonding electrons) provided a firm basis for the oxidation-state assignments on which oxidation–reduction definitions have become based.

Examples of oxidation–reduction reactions. Molecular oxygen is a conspicuously important oxidizing agent. It will directly oxidize all but a few of the metals and most of the nonmetals as well. Often these direct oxidations lead to normal oxides such as those of lithium (Li), zinc (Zn), phosphorus (P), and sulfur (S).

$$4Li + O_2 \longrightarrow 2Li_2O \text{ (lithium oxide)}$$

$$2Zn + O_2 \longrightarrow 2ZnO \text{ (zinc oxide)}$$

$$4P + 5O_2 \longrightarrow P_4O_{10} \text{ (phosphorus pentoxide)}$$

$$S + O_2 \longrightarrow SO_2 \text{ (sulfur dioxide)}$$

Oxidation of organic foodstuffs

Organic foodstuffs are oxidized to carbon dioxide and water in respiration. The reaction stoichiometry can be illustrated for glucose, a simple sugar:

$$C_6H_{12}O_6 + 6O_2 \longrightarrow 6CO_2 + 6H_2O$$

glucose — oxygen — carbon dioxide — water

Although the oxygen–glucose reaction is very slow at ambient temperatures outside the living cell, it proceeds quickly under the influence of enzymatic catalysis within the body. Essentially all organic compounds react with oxygen under appropriate conditions, but the rates of the reactions at ordinary temperatures and pressures vary greatly.

Many other oxidizing agents serve as oxygen-atom sources. Hydrogen peroxide (H_2O_2), acid chromate ion ($HCrO_4^-$), and hypochlorous acid (HClO) are reagents often used in oxygen-atom-transfer reactions, for example in the following reactions:

$$H_2O_2 + Cl^- \longrightarrow OCl^- + H_2O$$

hydrogen peroxide — chloride ion — hypochlorite ion — water

$$2HCrO_4^- + 3H_3C\overset{\overset{O}{\|}}{C}H + 8H^+$$

acid chromate ion — acetaldehyde — hydrogen ion

$$\longrightarrow 3H_3C\overset{\overset{O}{\|}}{C}OH + 2Cr^{3+}(aq) + 5H_2O$$

acetic acid — chromium(III) ion — water

$$5HClO + I_2 + H_2O \longrightarrow 5HCl + 2HIO_3$$

hypochlorous acid — iodine — water — hydrogen chloride — iodic acid

In the simplest hydrogen-atom transfers, molecular hydrogen serves as the hydrogen-atom source. The hydrogenations (hydrogenation is the adding of hydrogen atoms to a molecule) of ethylene and of molecular nitrogen are illustrative in the following equations:

$$H_2C{=}CH_2 + H_2 \longrightarrow H_3CCH_3$$

ethylene — hydrogen — ethane

$$N_2 + 3H_2 \longrightarrow 2NH_3$$

nitrogen — hydrogen — ammonia

Reactions of molecular hydrogen are characteristically slow at ordinary temperatures. The hydrogenation of molecular nitrogen and of olefins such as ethylene (an olefin is an unsaturated hydrocarbon compound; it has at least two adjacent carbon atoms joined by a double bond to which other atoms or groups of atoms can be joined directly) is a process of extraordinary commercial importance and requires catalysts to occur at useful rates.

Hydrogen-atom transfer in organic oxidation

Hydrogen-atom transfer from an organic molecule to a suitable acceptor is a common mode of organic oxidation. The oxidation of formic acid by permanganate and that of ethanol by acid chromate share stoichiometry that features hydrogen-atom loss by the organic species, as shown in the following equations:

$$3\underset{\text{formic acid}}{\overset{O}{\overset{\|}{HCOH}}} + 2\underset{\text{permanganate ion}}{MnO_4^-} + 2\underset{\text{hydrogen ion}}{H^+} \longrightarrow 3\underset{\text{carbon dioxide}}{CO_2} + 2\underset{\text{manganese dioxide}}{MnO_2} + 4\underset{\text{water}}{H_2O}$$

$$3\underset{\text{ethanol}}{CH_3CH_2OH} + 2\underset{\text{acid chromate ion}}{HCrO_4^-} + 8\underset{\text{hydrogen ion}}{H^+} \longrightarrow 8\underset{\text{water}}{H_2O} + 3\underset{\text{acetaldehyde}}{CH_3CHO} + 2\underset{\text{aqueous chromium(III) ion}}{Cr^{3+}(aq)}$$

The oxidizing agents permanganate and acid chromate, typical of many hydrogen-atom acceptors, undergo complicated changes rather than simple hydrogen-atom addition.

Electron-transfer stoichiometry is usually associated with metal ions in aqueous solution, as shown in the following equations:

$$\underset{\text{iron(II) ion}}{Fe^{2+}(aq)} + \underset{\text{cobalt(III) ion}}{Co^{3+}(aq)} \longrightarrow \underset{\text{iron(III) ion}}{Fe^{3+}(aq)} + \underset{\text{cobalt(II) ion}}{Co^{2+}(aq)}$$

$$\underset{\text{thallium(I) ion}}{Tl^{+}(aq)} + 2\underset{\text{cerium(IV) ion}}{Ce^{4+}(aq)} \longrightarrow \underset{\text{thallium(III) ion}}{Tl^{3+}(aq)} + 2\underset{\text{cerium(III) ion}}{Ce^{3+}(aq)}$$

Many positively charged metal ions have been shown to be bonded to water molecules, so that their electron-transfer reaction occurs between rather complex molecular groups. The iron ion formulas above, for example, are more properly written as $[Fe(H_2O)_6]^{2+}$ and $[Fe(H_2O)_6]^{3+}$ to reflect the presence of six water molecules bonded to the metal ion. Simple electron transfer between free ions is known only in the gas phase, as in this argon–sodium reaction:

$$\underset{\text{argon ion}}{Ar^+(g)} + \underset{\text{sodium}}{Na(g)} \longrightarrow \underset{\text{sodium ion}}{Na^+(g)} + \underset{\text{argon}}{Ar(g)} .$$

Several other types of redox reactions do not fall in the oxygen-atom, hydrogen-atom, or electron-transfer categories. Important among these are reactions of fluorine, chlorine, bromine, and iodine. These four elements known as the halogens, all form diatomic (two-atom) molecules, which are versatile oxidizing agents. The following examples are typical:

$$3F_2 + S \longrightarrow \underset{\text{sulfur hexafluoride}}{SF_6}$$

$$\underset{\text{chlorine}}{Cl_2} + 2\underset{\text{iron(II) ion}}{Fe^{2+}(aq)} \longrightarrow 2\underset{\text{chloride ion}}{Cl^-} + 2\underset{\text{iron(III) ion}}{Fe^{3+}(aq)}$$

$$\underset{\text{bromine}}{Br_2} + \underset{\text{ethylene}}{H_2C{=}CH_2} \longrightarrow \underset{\text{ethylene dibromide}}{BrCH_2CH_2Br}$$

$$\underset{\text{iodine}}{I_2} + 2\underset{\text{thiosulfate ion}}{S_2O_3^{2-}} \longrightarrow 2\underset{\text{iodide ion}}{I^-} + \underset{\text{polythionate ion}}{S_4O_6^{2-}}.$$

Such reactions often qualify as redox processes only in the broad sense that oxidation-state changes occur. The oxidation-state characterization extends oxidation–reduction chemistry to include examples from the reactions of all the chemical elements.

Significance of redox reactions. Oxidation–reduction reactions have vast importance not only in chemistry but in geology and biology as well. The surface of the Earth is a redox boundary between the planet's reduced metallic core and an oxidizing atmosphere. The Earth's crust is largely composed of metal oxides, and the oceans are filled with water, an oxide of hydrogen. The tendency of nearly all surface materials to be oxidized by the atmosphere is reversed by the life process of photosynthesis. Because they are constantly renewed by the photosynthetic reduction of carbon dioxide, life's complex compounds can continue to exist on the Earth's surface.

The Earth's surface as a redox boundary

For similar reasons, much of chemical technology hinges on the reduction of materials to oxidation states lower than those that occur in nature. Such basic chemical products as ammonia, hydrogen, and nearly all the metals are produced by reductive industrial processes. When not used as structural materials, these products are reoxidized in their commercial applications. The weathering of materials, including wood, metals, and plastics, is oxidative, since, as the products of technological or photosynthetic reductions, they are in oxidation states lower than those stable in the atmosphere.

Solar radiation is converted to useful energy by a redox cycle that operates continually on a global scale. Photosynthesis converts radiant energy into chemical potential energy by reducing carbon compounds to low oxidation states, and this chemical energy is recovered either through enzymatic oxidations at ambient temperatures or during combustion at elevated temperatures.

THEORETICAL ASPECTS OF OXIDATION REACTIONS

Oxidation states. The idea of assigning an oxidation state to each of the atoms in a molecule evolved from the electron-pair concept of the chemical bond. Atoms within a molecule are held together by the force of attraction that the nuclei of two or more of them exert on electrons in the space between them. In many cases this sharing of electrons can be regarded as involving electron pair bonds between adjacent nuclei. Electron pair bonding is often diagrammed so as to show all the bonding and nonbonding valence electrons; *e.g.*, the structures of atomic hydrogen, atomic chlorine, and hydrogen chloride shown below (each dot represents one valence electron):

Diagramming electron pair bonding

$$\underset{\text{hydrogen atom}}{H\cdot} + \underset{\text{chlorine atom}}{:\ddot{\underset{\cdot\cdot}{Cl}}\cdot} \longrightarrow \underset{\text{hydrogen chloride molecule}}{H:\ddot{\underset{\cdot\cdot}{Cl}}:}$$

The hydrogen chloride diagram reflects the presence, in the internuclear region, of two electrons that are under the mutual attractive influence of both the hydrogen and chlorine nuclei. Oxidation states for the hydrogen and chlorine in HCl are assigned according to the net charges that remain on H and Cl when the shared electrons are assigned to the atom that has the greater attraction for them. Through physical measurements on isolated atoms and simple molecules, these relative attractive powers have been determined. Table 1 lists the electronegativity values for some important elements.

In the hydrogen chloride molecule the chlorine is more electronegative than hydrogen and is, therefore, assigned both shared electrons. Chlorine has seven valence electrons in its neutral state; having acquired an eighth electron in its reaction with hydrogen, it is considered to have

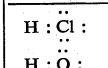
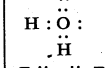
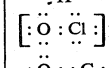
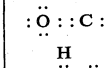
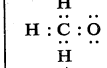
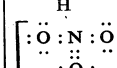
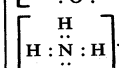
Table 1: Pauling Electronegativities of Selected Elements

Fluorine	4.0
Oxygen	3.5
Nitrogen	3.0
Chlorine	3.0
Bromine	2.8
Sulfur	2.5
Iodine	2.5
Carbon	2.5
Hydrogen	2.1
Phosphorus	2.1
Iron	1.8
Sodium	0.9

Source: L. Pauling, *The Nature of the Chemical Bond.*

an oxidation state of -1, while hydrogen is assigned $+1$, having lost the single valence electron it has in its neutral state. Charges arrived at in this way are the basis for oxidation-state assignments, conventionally represented by roman numerals, as in H(I) and Cl($-$I) for the constituents of HCl. Since determination of oxidation states is simply a method of conceptually distributing shared electrons to individual atoms, the same number of electrons must be accounted for, before and after such assignment. Table 2 includes examples of molecules with multiple bonds. The oxidation states of the atoms involved are added up algebraically in the table, and their sum must always equal the net charge on the molecule. There is no physical reality to oxidation states; they simply represent the results of calculations based on a formal rule.

Table 2: The Oxidation States of the Atoms in Typical Small Molecules

molecular species	oxidation-state assignments	algebraic sums
	H(I), Cl($-$I)	1 − 1 = 0
	H(I), O($-$II)	2(1) − 2 = 0
	Cl(I), O($-$II)	1 − 2 = −1
	C(IV), O($-$II)	4 − 2(2) = 0
	C($-$II), O($-$II), H(I)	4(1) − 2 − 2 = 0
	N(V), O($-$II)	5 − 3(2) = −1
	H(I), N($-$III)	4(1) − 3 = +1

Guidelines for assigning oxidation states

Oxidation states can be assigned for most common molecules with the help of a few guidelines. First, electrons shared by two atoms of the same element are divided equally; accordingly, elements are always in oxidation state of 0, regardless of their allotropic form (allotropic refers to the phenomenon of an element's having two or more forms; *e.g.*, carbon can exist as diamond or graphite and in both cases is in the 0 oxidation state). Second, only fluorine is more electronegative than oxygen. Therefore, except in compounds containing oxygen–oxygen or oxygen–fluorine bonds, oxygen can be reliably assigned the oxidation state -2. Similarly, hydrogen is less electronegative than fluorine, oxygen, nitrogen, chlorine, sulfur, and carbon (F, O, N, Cl, S, and C), so it is in the $+1$ oxidation state in its combinations with those elements. For many common compounds containing only hydrogen, oxygen, and a third element, the third element's oxidation state can be calculated, assuming oxidation numbers of $+1$ for hydrogen and -2 for oxygen. When bonds are present between two elements that differ little in electronegativity, however, oxidation-state assignments become doubtful, and the distinction between redox and nonredox processes is not evident.

There is a general reluctance, particularly regarding organic systems, to assume oxidation-state changes when the reaction results can be accounted for by the transfer or addition of water (H_2O), ammonia (NH_3), the hydroxide ion (OH^-), or the ions of hydrogen (H^+), chlorine (Cl^-), bromine (Br^-), or iodine (I^-), or combinations of these species; *e.g.*, the ammonium ion (NH_4^+), hydrogen chloride (HCl). The reason is that, in these molecules and ions, the elements are present in their most typical oxidation states: hydrogen(I), chlorine($-$I), oxygen($-$II), bromine($-$I), iodine($-$I), and nitrogen($-$III).

Example of oxygen-atom-transfer reaction

The oxidation-state concept clarifies the relationship between oxygen-atom, hydrogen-atom, and electron transfer. The oxygen- and hydrogen-transfer criteria apply only when oxygen and hydrogen occur in their typical oxidation states. An example of an appropriate reaction involving oxygen-atom transfer is the reduction of ferrous oxide by carbon monoxide:

$$\underset{\substack{\text{iron(II) oxide}\\\text{(ferrous oxide)}}}{FeO} + \underset{\substack{\text{carbon}\\\text{monoxide}}}{CO} \longrightarrow \underset{\substack{\text{carbon}\\\text{dioxide}}}{CO_2} + \underset{\text{iron}}{Fe}\,.$$

In terms of oxidation-state changes, this oxygen-atom transfer is equivalent to the two-electron reduction of iron and complementary two-electron oxidation of carbon:

$$Fe(II) + C(II) \longrightarrow C(IV) + Fe(0)\,.$$

Oxygen, which occurs in the oxidation state -2 in both reactants and products in the first equation, is not shown in the second. In transferring, the oxygen atom leaves two electrons behind, causing the reduction of iron, and acquires two electrons from carbon, oxidizing the carbon.

In a similar way, the hydrogenation of ethylene corresponds to a two-electron reduction of the two-carbon skeleton:

$$\underset{\text{hydrogen}}{H_2} + \underset{\text{ethylene}}{H_2C{=}CH_2} \longrightarrow \underset{\text{ethane}}{H_3CCH_3}$$

$$\underset{\substack{\text{hydrogen in}\\\text{oxidation state zero}}}{2H(0)} + \underset{\substack{\text{carbon in}\\\text{oxidation state }-2}}{2C(-II)}$$

$$\longrightarrow \underset{\substack{\text{hydrogen in}\\\text{oxidation state }+1}}{2H(I)} + \underset{\substack{\text{carbon in}\\\text{oxidation state }-3}}{2C(-III)}\,.$$

In this example also, the second equation includes only the atoms that change oxidation states: the four hydrogen atoms initially present in ethylene are in the $+1$ oxidation state in both reactants and products and are therefore omitted. Each of the two neutral hydrogen atoms can be regarded as giving up an electron to, and thereby reducing, one of the carbons. This example also shows that the oxidation that complements the reduction of ethylene is that of the two hydrogen atoms in H_2, from the 0 to the $+1$ oxidation state. General application of the oxidation-state concept leads to a formal viewpoint toward all redox reactions as electron-transfer reactions.

Half reactions. One of the basic reasons that the concept of oxidation–reduction reactions helps to correlate chemical knowledge is that a particular oxidation or reduction can often be carried out by a wide variety of oxidizing or reducing agents. Reduction of the iron(III) ion to the iron(II) ion by four different reducing agents provides an example:

$$\underset{\substack{\text{iron(III)}\\\text{ion}}}{Fe^{3+}(aq)} + \underset{\substack{\text{vanadium(II)}\\\text{ion}}}{V^{2+}(aq)} \longrightarrow \underset{\substack{\text{iron(II)}\\\text{ion}}}{Fe^{2+}(aq)} + \underset{\substack{\text{vanadium(III)}\\\text{ion}}}{V^{3+}(aq)}$$

$$\underset{\substack{\text{iron(III)}\\\text{ion}}}{2Fe^{3+}(aq)} + \underset{\substack{\text{zinc}\\\text{metal}}}{Zn(s)} \longrightarrow \underset{\substack{\text{iron(II)}\\\text{ion}}}{2Fe^{2+}(aq)} + \underset{\substack{\text{zinc(II)}\\\text{ion}}}{Zn^{2+}(aq)}$$

$$2Fe^{3+}(aq) + 2I^- \longrightarrow 2Fe^{2+}(aq) + I_2$$

iron(III) ion; iodide ion; iron(II) ion; iodine

$$2Fe^{3+}(aq) + C_6H_4(OH)_2 \longrightarrow 2Fe^{2+}(aq) + C_6H_4O_2 + 2H^+.$$

iron(III) ion; hydroquinone; iron(II) ion; quinone; hydrogen ion

Production of the same change in the aqueous iron(III) ion by different reductants emphasizes the fact that the reduction is a characteristic reaction of the iron system itself, and, therefore, the process may be written without specifying the identity of the reducing agent in the following way:

$$Fe^{3+}(aq) + e^- \longrightarrow Fe^{2+}(aq) .$$

iron(III) ion; electron; iron(II) ion

Hypothetical equations of this type are known as half reactions. The symbol e^-, which stands for an electron, serves as a reminder that an unspecified reducing agent is required to bring about the change. Half reactions can be written, equally, for the reducing agents in the four reactions with ferric ion:

$$V^{2+}(aq) \longrightarrow V^{3+}(aq) + e^-$$

vanadium(II) ion; vanadium(III) ion; electron

$$Zn(s) \longrightarrow Zn^{2+}(aq) + 2e^-$$

zinc(0) metal; zinc(II) ion; electron

$$2I^- \longrightarrow I_2 + 2e^-$$

iodide ion; iodine; electron

$$C_6H_4(OH)_2 \longrightarrow C_6H_4O_2 + 2e^- + 2H^+$$

hydroquinone; quinone; electron; hydrogen ion

Although hypothetical, half reactions are properly balanced chemical processes. Since $V^{2+}(aq)$ increases its oxidation number by one, from +2 to +3, in the first half reaction, an electron is shown as a product of the change. Similarly, two electrons are produced when the oxidation number of zinc increases from 0 to +2 in the second half reaction. When half reactions for hypothetical isolated oxidations and reductions are combined, the electrons must cancel if the equation for a possible overall chemical reaction is to result.

The use of half reactions is a natural outgrowth of the application of the electron-transfer concept to redox reactions. Since the oxidation-state principle allows any redox reaction to be analyzed in terms of electron transfer, it follows that all redox reactions can be broken down into a complementary pair of hypothetical half reactions. Electrochemical cells (in which chemical energy can be converted to electrical energy, and vice versa) provide some physical reality to the half-reaction idea. Oxidation and reduction half reactions can be carried out in separate compartments of electrochemical cells, with the electrons flowing through a connecting wire and the circuit completed by some arrangement for ion migration between the two compartments (but the migration need not involve any of the materials of the oxidation–reduction reactions themselves).

Redox potentials for common half reactions. Analysis of the electrical potential, or voltage, developed by pairing various half reactions in electrochemical cells has led to the determination of redox potentials for many common half reactions. While a detailed description of redox potentials requires the methods of thermodynamics (the branch of physics concerned with the role of heat in transformation of matter or energy), much useful information can be obtained from redox potentials with minimal recourse to formal theory. Basically, a table of half-cell potentials is a summary of the relative tendencies of different oxidations and reductions to occur. Table 3 lists selected half reactions and their corresponding reduction potentials (symbolized by $E°$).

Table 3: Selected Values of Standard Reduction Potentials

half reactions*	$E°$ (volts)
$F_2(g) + 2e^- \rightarrow 2F^-$ fluorine(0) electrons fluoride(−I) ion	2.87
$MnO_4^- + 8H^+ + 5e^- \rightarrow Mn^{2+}(aq) + 4H_2O$ permanganate ion; hydrogen(I) ion; electrons; manganese(II) ion; water	1.51
$Cl_2(g) + 2e^- \rightarrow 2Cl^-$ chlorine(0) electrons chloride(−I) ions	1.36
$O_2(g) + 4H^+ + 4e^- \rightarrow 2H_2O$ oxygen(0) hydrogen(I) ions electrons water	1.23
$Fe^{3+}(aq) + e^- \rightarrow Fe^{2+}(aq)$ iron(III) ion electron iron(II) ion	0.77
$Cu^{2+}(aq) + 2e^- \rightarrow Cu(s)$ copper(II) ion electrons copper(0)	0.34
$2H^+ + 2e^- \rightarrow H_2(g)$ hydrogen(I) ion electrons hydrogen(0)	0.00
$Zn^{2+}(aq) + 2e^- \rightarrow Zn(s)$ zinc(II) ion electrons zinc(0)	−0.76
$Na^+ + e^- \rightarrow Na(s)$ sodium(I) ion electron sodium(0)	−2.71

*The identifications in parentheses refer to the physical state of the substance: (g), gas; (aq), as a hydrated positive ion in water; (s), as the pure solid.
Source: W. Latimer, *Oxidation Potentials.*

The physical significance of the values is directly linked to several agreements about their use. First, the greater the value of $E°$ (the reduction potential), the greater the tendency of a half reaction to proceed from left to right (as written). The half reactions in Table 3 are listed from top to bottom in order of decreasing $E°$: the higher a reaction's position on the list, the greater the tendency of the reactants to accept electrons. In other words, reagents high on the list, such as fluorine gas (F_2) and permanganate ion (MnO_4^-), are strong oxidizing agents. Second, the reduction of hydrogen ions (H^+) to hydrogen gas (H_2) is arbitrarily assigned the value 0 volts. Half cells with positive reduction potentials involve reactants more easily reduced than H^+; conversely, those with negative potentials, reactants more difficult to reduce than hydrogen ions.

Use of reduction potentials to predict reactions

With the aid of reduction potentials, it is possible to predict whether a particular oxidation–reduction reaction can occur. The predictions require breaking down the overall reaction into two half reactions of known reduction potentials. For example, if a strip of zinc metal is dipped into a solution containing copper(II) ion, the possibility exists for a redox process, which can be regarded as the sum of the half reactions aqueous zinc ion (Zn^{2+}[aq]) to zinc metal (Zn[s]) and aqueous copper ion (Cu^{2+}[aq]) to copper metal (Cu[s]), as follows:

$$Zn(s) + Cu^{2+}(aq) \longrightarrow Zn^{2+}(aq) + Cu(s) .$$

zinc(0) metal; copper(II) ion; zinc(II) ion; copper(0) metal

Combining these two half reactions requires writing the zinc ion to zinc metal half reaction the reverse of the way it appears in Table 3. When the direction of a half reaction is reversed, so that it can be added to another half reaction, the sign of its redox potential is also reversed (in this case, from negative to positive), and the two reduction potential values are then added.

$$Cu^{2+}(aq) + 2e^- \longrightarrow Cu(s) \qquad +0.34 \text{ volt}$$

$$Zn(s) \longrightarrow Zn^{2+}(aq) + 2e^- \qquad +0.76$$

$$Cu^{2+}(aq) + Zn(s) \xrightarrow{\text{spontaneous}} Cu(s) + Zn^{2+}(aq) \qquad +1.10$$

copper(II) ion; zinc metal; copper metal; zinc(II) ion

The resulting $E°$ value for the net reaction, +1.10 volts, measures the tendency of the net reaction to occur. If $E°$ for a particular net reaction is positive, the process may be expected to occur spontaneously when the reactants are mixed at specified concentrations (one mole per litre;

see below *Oxidation–reduction equilibria*). Therefore, it is predicted that copper metal should be deposited on a strip of zinc metal when the latter is immersed in a solution of a copper(II) salt. This reaction is, in fact, readily observed in the laboratory. A more specific physical interpretation of the +1.10 volt value is that it represents the voltage that would be produced by an ideal electrochemical cell based on the copper(II) ion to copper metal and zinc(II) ion to zinc metal half reactions with all the reagents at specified concentrations.

When the same two half cells are combined, with both their directions (and therefore the signs of their redox potentials) reversed, it is predicted that the reverse reaction, the depositing of zinc metal from a zinc(II) ion solution onto a copper strip, will not occur spontaneously. As in the case of $E°$ values for half reactions, those for net redox reactions also change sign when the direction of the reaction is reversed.

$$\begin{array}{lll} \mathrm{Cu(s)} \longrightarrow \mathrm{Cu^{2+}(aq)} + 2e^- & & -0.34 \text{ volt} \\ \mathrm{Zn^{2+}(aq)} + 2e^- \longrightarrow \mathrm{Zn(s)} & & -0.76 \\ \hline \mathrm{Zn^{2+}(aq)} + \mathrm{Cu(s)} \xrightarrow{\text{nonspontaneous}} \mathrm{Cu^{2+}(aq)} + \mathrm{Zn(s)} & & -1.10 \end{array}$$

The results of the copper–zinc system can be applied more generally to the half reactions in Table 3. For example, copper(II) ion in water (Cu^{2+}[aq]) is an oxidant strong enough to force a half reaction lower on the table to proceed spontaneously in the opposite direction of that written. Therefore, not only is copper(II) ion expected to oxidize zinc metal (Zn[s]) to zinc(II) ion (Zn^{2+}[aq]); it is also predicted to oxidize hydrogen gas (H_2[g]) to hydrogen ion (H^+) and sodium metal (Na[s]) to sodium ion (Na^+). Similarly, fluorine gas (F_2[g]), the strongest oxidant listed in Table 3, is predicted to oxidize spontaneously the products of all the other half reactions in the table. In contrast, the strongest reducing agent is solid sodium metal (Na[s]), and it is expected spontaneously to reduce the reactants of all the other half cells.

Oxidation–reduction equilibria. In practice many chemical reactions can be carried out in either direction, depending on the conditions. The spontaneous direction predicted for a particular redox reaction by half-cell potentials is appropriate to a standard set of reaction conditions. Specifically, the temperature is assumed to be 25° C (77° F) with reagents at specified concentrations. Gases are present at one atmosphere pressure and solutes at one mole per litre (one molecular weight in grams dissolved in one litre of solution) concentration (1M). Solids are assumed to be in contact with the reaction solution in their normal stable forms, and water is always taken to be present as the solvent. Many practical problems can be solved directly with standard reduction potentials.

Role of the Nernst equation

The usefulness of reduction potentials is greatly extended, however, by a thermodynamic relationship known as the Nernst equation, which makes it possible to calculate changes in half-cell potentials that will be produced by deviations from standard concentration conditions. In the reaction between zinc metal and copper(II) ion, standard conditions for zinc and copper metal require simply that both solids be present in contact with the solution; the $E°$ values are not affected by either the total or proportionate amounts of the two metals. The calculation that the overall reaction is spontaneous by +1.10 volts is based on standard one mole per litre (1M) concentrations for aqueous zinc(II) ion (Zn^{2+}[aq]) and aqueous copper(II) ion (Cu^{2+}[aq]). Using the Nernst equation it is found that $E°$ for the overall reaction will be +1.10 volts as long as both ions are present in equal concentrations, regardless of the concentration level.

On the other hand, if the ratio of the zinc(II) to copper(II) ion concentrations is increased, the reduction potential ($E°$) falls until, at a very high preponderance of zinc ion, $E°$ becomes 0 volt. At this point, there is no net tendency for the reaction to proceed spontaneously in either direction. If the zinc(II) to copper(II) ion ratio is increased further, the direction of spontaneity reverses, and zinc ion spontaneously oxidizes copper metal. In practice, such high zinc(II) to copper(II) ion concentration ratios are unattainable, which means that the reaction can only be carried out spontaneously with copper(II) ion oxidizing zinc metal. Many reactions with $E°$ values smaller than +1.10 volts under standard conditions can be carried out in either direction by adjusting the ratio of product and reactant concentrations. The point at which $E° = 0$ volt represents a state of chemical equilibrium. When chemical reactions are at equilibrium, the concentrations of the reagents do not change with time, since net reaction is not spontaneous in either direction. Measurements of half-cell potentials combined with Nernst-equation calculations are a powerful technique for determining the concentration conditions that correspond to chemical equilibrium.

Reaction rates. *Predictability.* There are practical limitations on predictions of the direction of spontaneity for a chemical reaction, the most important arising from the problem of reaction rates. An analogy can be made with the simple physical system of a block on a sloping plane. Because of the favourable energy change, the block tends spontaneously to slide down, rather than up, the slope, and, at mechanical equilibrium, it will be at the bottom of the slope, since that is the position of lowest gravitational energy. How rapidly the block slides down is a more complex question, since it depends on the amount and kind of friction present. The direction of spontaneity for a chemical reaction is analogous to the downhill direction for a sliding block, and chemical equilibrium is analogous to the position at the bottom of the slope; the rate at which equilibrium is approached depends on the efficiency of the available reaction processes. Between zinc metal and aqueous copper(II) ion, the reaction proceeds without observable delay, but various other spontaneous redox processes proceed at imperceptibly slow rates under ordinary conditions.

Biological processes. A particularly significant illustration of the role of mechanisms in determining the rates of redox reactions concerns respiration, the central energy-producing process of life. Foodstuffs that are oxidized by molecular oxygen during respiration are quite unreactive with oxygen before ingestion. Such high-energy foods as grains and sugar can resist the atmosphere indefinitely but are rapidly converted to carbon dioxide and water through combination with oxygen during respiratory metabolism. The situation is exemplified by the behaviour of glucose at ambient temperatures.

The significance of the different rate behaviour of high-energy foods inside and outside the cell has been dramatized by Albert Szent-Györgyi, a Hungarian biochemist resident in the United States, a pioneering researcher into the chemical mechanism of respiration:

> You remember the exciting story of the grave of the Egyptian emperor. At its opening the breakfast of the emperor was found unburned though it had been exposed to the action of oxygen during several thousand years at a temperature that was not very different from 37° C [98.6° F]. Had the king risen and consumed his breakfast, as he had anticipated doing, the food would have been oxidized in no time, that is to say the cells of the emperor would have made reactions take place that would not run spontaneously (from Albert V. Szent-Györgyi, *On Oxidation, Fermentation, Vitamins, Health and Disease;* the Williams and Wilkins Company, 1939).

Living systems are able to use respiratory oxidation as an energy source only because the same reactions are slow outside the cell. In return for providing an efficient mechanism for the oxidation of foods, the cell gains control over the disposition of the liberated chemical energy.

Examples such as the chemistry of respiration make clear the importance of determining the rates and mechanisms of redox reactions. Often questions are difficult to answer even in regard to relatively simple reactions. It has been pointed out that many redox processes can be categorized as oxygen-atom-, hydrogen-atom-, or electron-transfer processes. These categories describe the net changes that are involved but provide no insight into the mechanisms of the reactions.

Mechanisms of redox reactions. Some of the problems associated with formulating descriptions of the mechanisms are illustrated by the reaction between two metal

ions that undergo complementary, one-unit changes in oxidation state:

$$M(\text{III}) + N(\text{II}) \longrightarrow M(\text{II}) + N(\text{III}) .$$

There are many different metal ions, designated with the letters M and N, which participate in redox reactions with this basic stoichiometry. To define the possible mechanisms with any precision, it is necessary to identify the groups that are directly bonded to the metals, as well as to specify which particular metals take part in the reactions. One relatively simple example is the chromium(II)–iron(III) reaction. Both chromium and iron aquo ions are surrounded by six water molecules in both the +2 and +3 oxidation states.

$$\underset{\substack{\text{hexaaquoiron(III)}\\\text{ion}}}{Fe(H_2O)_6^{3+}} + \underset{\substack{\text{hexaaquochromium(II)}\\\text{ion}}}{Cr(H_2O)_6^{2+}} \longrightarrow \underset{\substack{\text{hexaaquoiron(II)}\\\text{ion}}}{Fe(H_2O)_6^{2+}} + \underset{\substack{\text{hexaaquochromium(III)}\\\text{ion}}}{Cr(H_2O)_6^{3+}}$$

The reaction appears from its stoichiometry to entail simple electron transfer from Cr(II) to Fe(III). There are, however, other possibilities. Their variety is most evident if it is assumed that the oxidizing agent, hexaaquoiron(III) ($Fe[H_2O]_6^{3+}$), dissociates a hydrogen ion before reacting with hexaaquochromium(II) ($Cr[H_2O]_6^{2+}$). The hexaaquoiron(III) ion is then converted to a hydroxopentaaquoiron(III) ion, known to be an important reactant in the Fe(III)–Cr(II) reaction. Its formation requires no change in the oxidation state of iron and should be regarded as a simple acid–base reaction. With hydroxopentaaquoiron(III) ion, $(H_2O)_5FeOH^{2+}$, as the oxidizing agent, simple electron transfer does provide one plausible means for the reaction. All the steps in the mechanism must add up to give the overall reaction. Intermediate species such as hydroxopentaaquoiron(III) must cancel out in the addition process by occurring as products in one step and reactants in another:

$$\underset{\substack{\text{hexaaquoiron(III)}\\\text{ion}}}{Fe(H_2O)_6^{3+}} \longrightarrow \underset{\substack{\text{hydroxopenta-}\\\text{aquoiron(III)}\\\text{ion}}}{(H_2O)_5FeOH^{2+}} + \underset{\substack{\text{hydrogen}\\\text{ion}}}{H^+}$$

$$\underset{\substack{\text{hydroxopenta-}\\\text{aquoiron(III)}\\\text{ion}}}{(H_2O)_5FeOH^{2+}} + \underset{\substack{\text{hexaaquo-}\\\text{chromium(II)}\\\text{ion}}}{Cr(H_2O)_6^{2+}} \xrightarrow[\text{transfer}]{\text{electron}} \underset{\substack{\text{hydroxopenta-}\\\text{aquoiron(II)}\\\text{ion}}}{(H_2O)_5FeOH^{+}} + \underset{\substack{\text{hexaaquo-}\\\text{chromium(III)}\\\text{ion}}}{Cr(H_2O)_6^{3+}}$$

$$\underset{\substack{\text{hydroxopenta-}\\\text{aquoiron(II)}\\\text{ion}}}{(H_2O)_5FeOH^{+}} + \underset{\substack{\text{hydrogen}\\\text{ion}}}{H^+} \longrightarrow \underset{\substack{\text{hexaaquoiron(II)}\\\text{ion}}}{Fe(H_2O)_6^{2+}}$$

$$Fe(H_2O)_6^{3+} + Cr(H_2O)_6^{2+} \longrightarrow Fe(H_2O)_6^{2+} + Cr(H_2O)_6^{3+}$$

An alternative to simple electron transfer is the transfer of a hydrogen atom from a water molecule bound to chromium(II) to the hydroxide ion (OH^-) bound to iron(III). A third important possibility is the transfer of a neutral OH group (called a hydroxyl radical) from iron(III) to chromium(II). Oxygen and hydrogen occur as oxygen(−II) and hydrogen(I) in both reactants and products. These oxidation states correspond to those present in the hydroxide ion OH^-. Therefore, when an OH group transfers, it must leave an electron behind, reducing the iron, and then reaccept an electron from chromium, oxidizing it. Since both the aqueous iron and chromium ions have a strong tendency to remain surrounded by six oxygen molecules, the most likely method of OH transfer would be the sharing of the transferred oxygen between the two metals at an intermediate stage of the reaction. But iron(III) is not a strong enough oxidizing agent to produce a significant concentration of neutral OH groups from OH^- ions. A more acceptable description of the OH transfer mechanism, therefore, is to regard OH^- as a bridging group through which electron transfer is accomplished as the two metal ion groups share a bonded oxygen atom.

$$[(H_2O)_5\overset{3+}{Fe}-\underset{\displaystyle H}{\underset{|}{O}}-\overset{2+}{Cr}(H_2O)_5]^{4+} \longrightarrow [(H_2O)_5\overset{2+}{Fe}-\underset{\displaystyle H}{\underset{|}{O}}-\overset{3+}{Cr}(H_2O)_5]^{4+}$$

Experimentally, it is difficult to distinguish between the alternative mechanisms posed above. Classification schemes for reaction mechanisms are of little value unless they offer alternatives that can be experimentally verified. For that reason the most successful system for classifying the mechanisms of redox reactions between metal ions has been the inner sphere–outer sphere dichotomy. Outer-sphere reactions are those that take place without breaking any bonds between a metal and a group such as water or hydroxide ion bound to it. Both the simple electron- and hydrogen-atom-transfer mechanisms are outer-sphere processes. If, on the other hand, the oxidation-state changes take place within an intermediate product in which the two metals are directly bonded to a common bridging group (such as shown in the equation above), the mechanism is inner-sphere. Formation of the intermediate requires loss of a bound water molecule by one of the metals. The inner sphere–outer sphere distinction can be tested experimentally in some redox systems.

BIBLIOGRAPHY. R.E. DICKERSON, H.B. GRAY, and G.P. HAIGHT, *Chemical Principles* (1970), a textbook covering most introductory aspects of oxidation-reduction reactions; F.A. COTTON and G. WILKINSON, *Advanced Inorganic Chemistry*, 2nd ed. rev. (1966), a comprehensive reference work with examples of redox reactions; W.M. LATIMER, *Oxidation Potentials*, 2nd ed. (1952), the definitive survey of the redox behaviour of the elements with an emphasis on half-reaction potentials; W.M. CLARK, *Oxidation-Reduction Potentials of Organic Systems* (1960), emphasizes biologically important reactions; L. PAULING, *The Nature of the Chemical Bond and the Structure of Molecules and Crystals*, 3rd ed. (1960), contains a detailed treatment of electronegativities; R. STEWART, *Oxidation Mechanisms* (1964), a concise monograph on organic oxidation-reduction mechanisms; H. TAUBE, *Electron Transfer Reactions of Complex Ions in Solution* (1970), a short monograph on the mechanisms of inorganic oxidation-reduction reactions with primary emphasis on the transition metals; J.B. CONANT, *The Overthrow of the Phlogiston Theory* (1950), a short readable account of the phlogiston theory for the nonscientist; E.I. RABINOWITCH and GOVINDJEE, *Photosynthesis* (1969), includes a good overview of the global redox cycle of respiration and photosynthesis.

(M.V.O.)

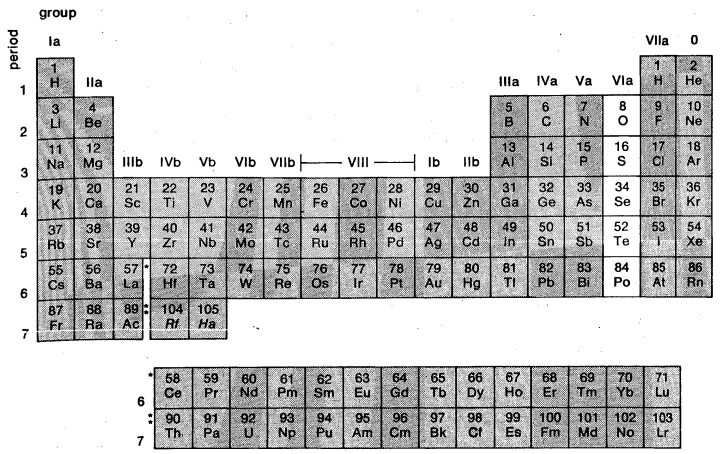
Oxygen Group Elements and Their Compounds

Oxygen, sulfur, selenium, tellurium, and polonium comprise the family referred to in the periodic classification of the chemical elements as Group VIa, or the chalcogens (see the Figure). A relationship among the first three members of the group was recognized as early as 1829; tellurium was assigned its place by 1865, and the discovery of polonium in 1898 completed the group. By the early 1970s, the possibility of finding element 116, the next member of Group VIa, in nature appeared to be negligible.

I. General considerations

OCCURRENCE IN NATURE

Estimates of the proportions of the various kinds of atoms in the universe put oxygen fourth in abundance, after hydrogen, helium, and neon, but the importance of such a ranking is slight since hydrogen atoms account for almost 94 percent of the total and helium for most of the rest. About three atoms out of 10,000 are oxygen, but because the mass of an oxygen atom is approximately 16 times that of a hydrogen atom, oxygen constitutes a larger fraction of the mass of the universe, though still only about 0.5 percent. In the regions ordinarily accessi-



Oxygen group elements in the periodic table.

ble to man, however—*i.e.*, within a few kilometres of the surface of the Earth—oxygen is the most abundant element: in mass, it makes up about 20 percent of the air, about 46 percent of the solid crust of the Earth, and about 89 percent of the water.

Oxygen is represented by the chemical symbol O. In the air, it exists mostly as molecules each made up of two atoms (O_2), although small amounts of ozone (O_3), in which three atoms of oxygen make up each molecule, are present in the atmosphere. Oxygen is a colourless, odourless, tasteless gas essential to living organisms, being taken up by animals, which convert it to carbon dioxide; plants, in turn, utilize carbon dioxide as a source of carbon and return the oxygen to the atmosphere. Oxygen forms compounds by reaction with practically any other element, as well as by reactions that displace elements from their combinations with each other; in many cases, these processes are accompanied by the evolution of heat and light and in such cases are called combustions.

In cosmic abundance, sulfur ranks ninth among the elements, accounting for only one atom of every 20,000–30,000. Sulfur occurs in the uncombined state as well as in combination with other elements in rocks and minerals that are widely distributed, although it is classified among the minor constituents of the Earth's crust, in which its proportion is estimated to be between 0.03 and 0.06 percent. On the basis of the finding that certain meteorites contain about 12 percent sulfur, it has been suggested that deeper layers of the Earth contain a much larger proportion. Seawater contains about 0.09 percent sulfur in the form of sulfate. The most important source is underground deposits of very pure sulfur present in domelike geologic structures where the sulfur is believed to have been formed by the action of bacteria upon the mineral anhydrite, in which sulfur is combined with oxygen and calcium. Deposits of sulfur in volcanic regions probably originated from gaseous hydrogen sulfide generated below the surface of the Earth and transformed into sulfur by reaction with the oxygen in the air.

Sulfur exists under ordinary conditions as a pale yellow, crystalline, nonmetallic solid; it is odourless and tasteless, combustible, and insoluble in water. Its chemical symbol is S. It reacts with all metals except gold and platinum, forming sulfides; it also forms compounds with several of the nonmetallic elements. Several million tons of sulfur are produced each year, mostly for the manufacture of sulfuric acid, which is widely used in industry.

The element selenium (symbol Se) is much rarer than oxygen or sulfur, comprising approximately 90 parts per billion of the crust of the Earth. It is occasionally found uncombined, accompanying native sulfur, but is more often found in combination with heavy metals (as copper, mercury, lead, or silver) in a few minerals. The principal commercial source of selenium is as a by-product of copper refining; its major uses are in the manufacture of electronic equipment, in pigments, and in making glass. The gray, metallic form of the element is the most stable under ordinary conditions; this form has the unusual property of greatly increasing in electrical conductivity when exposed to light. Selenium compounds are toxic to animals; plants grown in seleniferous soils may concentrate the element and become poisonous.

Tellurium is a silvery-white element (symbol Te) with properties intermediate between those of metals and nonmetals; it makes up approximately one part per billion of the Earth's crust. Like selenium, it is less often found uncombined than as compounds of metals such as copper, lead, silver, or gold, and is obtained chiefly as a by-product of the refining of copper or lead. No large use for tellurium has been found.

Polonium (symbol Po) is an extremely rare, radioactive element found in minerals containing uranium. It has some scientific applications as a source of alpha radiation.

HISTORICAL SURVEY

Oxygen. Although several substances now recognized as elements were known in their uncombined states in ancient times, the presently accepted idea of chemical elements dates from 1661, when the English natural philosopher Robert Boyle set forth their nature as simple, primitive, perfectly unmixed bodies, not formed from each other nor from any other bodies. Despite its terrestrial abundance, oxygen was not recognized as belonging to this class before the 1770s. Two chemists, Carl Wilhelm Scheele, in about 1772, and Joseph Priestley, in 1774, both obtained oxygen by heating certain metal oxides. Antoine-Laurent Lavoisier, with remarkable insight, interpreted the role of oxygen in respiration as well as combustion, discarding the phlogiston theory, which had been accepted up to that time; he noted its tendency to form acids by combining with many different substances and accordingly named the element *oxygen* (*oxygène*) from the Greek words for "acid former."

Boyle's definition of chemical elements

Sulfur. The history of sulfur is part of antiquity. The name itself probably found its way into Latin from the language of the Oscans, an ancient people who inhabited the region including Vesuvius, where sulfur deposits are widespread. Prehistoric man used sulfur as a pigment for cave painting; one of the first recorded instances of the art of medication is in the use of sulfur as a tonic.

The combustion of sulfur had a role in Egyptian religious ceremonials as long as 4,000 years ago. "Fire and brimstone" references in the Bible are related to sulfur, suggesting that "hell's fires" are fuelled by sulfur. The beginnings of practical and industrial uses of sulfur are credited to the Egyptians, who used sulfur dioxide for bleaching cotton as early as 1600 BC. Greek mythology includes sulfur chemistry: Homer tells of Odysseus' use of sulfur dioxide to fumigate a chamber in which he had slain his wife's suitors. The use of sulfur in explosives and fire displays dates to about 500 BC in China, and flame-producing agents used in warfare (Greek fire) were prepared with sulfur in the Middle Ages. Pliny the Elder in AD 50 reported a number of individual uses of sulfur and ironically was himself killed, in all probability by sulfur fumes, at the time of the great Vesuvius eruption (AD 79). Sulfur was regarded by the alchemists as the principle of combustibility. Lavoisier recognized it as an element in 1777, although it was considered by some to be a compound of hydrogen and oxygen; its elemental nature was established by the French chemists Joseph Gay-Lussac and Louis Thenard.

Selenium. In 1817 a Swedish chemist named Jöns Jacob Berzelius noted a red substance resulting from sulfide ores from mines of Fahlun (Falun), Sweden. When this red material was investigated in the following year, it proved to be an element and was named after the moon or the moon goddess Selene. An ore of unusually high selenium content was discovered by Berzelius only days before he made his report to the scientific societies of the world on selenium. His sense of humour is evident in the name he gave the ore, *eucairite*, meaning "just in time."

Tellurium. The element tellurium was isolated before it was actually known to be an elemental species. About 1782 Franz Joseph Müller von Reichenstein, an Austrian

mineralogist, worked with an ore referred to as German gold. From this ore he obtained a material that defied his attempts at analysis and was called by him *metallum problematicum*. In 1798 Martin Heinrich Klaproth confirmed Müller's observations and established the elemental nature of the substance. He named the element after man's "heavenly body" Tellus, or Earth.

Polonium. The fifth member of Group VIa, polonium, was the first to be found by taking advantage of the phenomenon of radioactivity, discovered by Henri Becquerel in the last part of the 19th century. Pierre and Marie Curie isolated the element while carrying out analyses on a uranium ore, pitchblende. The very intense radioactivity not attributable to uranium was ascribed to a new element, named by them after Mme Curie's homeland, Poland. The discovery was announced in July 1898.

OXYGEN GROUP ELEMENTS IN THE PERIODIC TABLE

Atomic and molecular nature of matter. *Atoms, nuclei, and electrons.* Of the millions of forms in which matter is known to exist, all are composed of one or more of only about 100 different kinds of tiny particles called atoms. These 100-odd basic varieties of matter are the chemical elements, each of which has distinctive properties that are systematically related to the way in which their atoms are, in turn, built up from even smaller, more fundamental particles. Practically the entire mass of every atom is concentrated in its nucleus, which is made up of protons, each possessing a positive electrical charge, and neutrons, which possess no charge. The number of protons in the nucleus of an atom is called the atomic number and determines the chemical identity of that atom: an atom with eight protons in its nucleus is an oxygen atom; one with 16 protons is a sulfur atom; one with 34 is a selenium atom. The number of neutrons in the nucleus is approximately the same as the number of protons, but for every element, more than one combination of protons and neutrons is possible. Oxygen nuclei, for example, always contain eight protons but may contain from five to 12 neutrons; these different forms of a single element are called isotopes. Some isotopes are unstable, spontaneously changing into others, but of the known elements, 88 have at least one stable isotope, and many have several. Of the eight isotopes of oxygen, for example, those with eight, nine, or ten neutrons are stable, and the one with eight comprises more than 99 percent of the oxygen found in nature.

Protons and neutrons

The nucleus is surrounded by a cloud of electrons, particles each having a single negative electrical charge but very little mass. The number of electrons in an atom is the same as the number of protons in its nucleus, so that the entire atom is electrically neutral. The electrons are distributed in regions of space called orbitals, each with a capacity of two electrons, arranged in shells of increasing distance from the nucleus. Each successive shell contains a larger number of orbitals that differ in shape and orientation. Each electron is attracted to the nucleus but repelled by all the other electrons in the atom, and, as a result of the variations between orbitals, the sum of all these interactions is different for every electron. The particular set of orbitals occupied by electrons, called the configuration of the atom, is determined by the tendency of the total energy involved in the interactions to assume a minimum value.

The electrons in the outermost shell are least strongly bound to the atom and are the only ones affected in ordinary chemical processes; these, called the valence electrons, determine the tendency of an atom to take part in chemical reactions leading to formation of compounds.

Reactions and chemical bonds. A configuration of eight electrons in the outermost shell is an especially stable one, as shown by the extremely small tendency of the noble gases neon, argon, krypton, xenon, and radon, which have this configuration, to take part in chemical reactions. To achieve this condition, atoms that have seven electrons outermost have a strong tendency to gain one more; this gain makes the number of electrons larger than the number of protons, and the resulting negatively charged particle is called an ion. Atoms with six or five valence electrons show the same tendency, but not to the same extent, because each additional electron is increasingly repelled by the negative charge introduced by the first.

Ions and molecules

Atoms that have one electron outermost and eight electrons in the next lower shell have a strong tendency to lose their lone outer electron, forming a positive ion that has a noble gas configuration. Loss of two or three electrons occurs progressively less readily, because in these cases each electron is being removed from the attractive positive charge resulting from loss of the first.

Both these processes are illustrated by the exchange of an electron between atoms of chlorine and potassium. A chlorine atom has two electrons in its first shell, eight in its second, and seven in its third; a potassium atom has two, eight, and eight, respectively, and one in the fourth shell. Loss of the single electron from the fourth shell of the potassium atom leaves the third shell, with eight electrons, outermost, and addition of this electron to the third shell of the chlorine atom gives it eight. Therefore, when potassium and chlorine atoms approach one another, an electron is transferred, resulting in the formation of a positive potassium ion and a negative chloride ion, with each ion having the electron configuration of the noble gas argon.

The attraction between opposite charges constitutes a chemical bond—called ionic or electrovalent—that causes large numbers of these ions to cling together in an orderly array. The formula of the compound is written KCl, which indicates that there is one chloride ion for each potassium ion, but should not be taken to mean that there are molecules—that is, discrete particles in which one potassium ion is specifically associated with one of chlorine. The transfer of electrons between atoms is the feature underlying oxidation and reduction processes. Oxidation was originally defined as the reaction of oxygen with another element, such as hydrogen, carbon, or zinc. Recognition that many other elements reacted in the same general way as oxygen led to extension of the term oxidation to reactions of those elements, such as the oxidation of zinc by chlorine or the oxidation of iron by sulfur. Establishment of the fact that all these oxidation reactions are accompanied by transfer of electrons to the oxidizing agent allowed the generalization that oxidation is any process in which one atom removes electrons from another; the atom that contributes these electrons to the oxidizing agent is called the reducing agent.

Oxidation and reduction

The number of electrons lost in the oxidation or reduction of an element is called the oxidation number. For example, in the oxidation of potassium by chlorine, an atom of potassium loses one electron, and its oxidation number changes from zero to plus one. At the same time, a chlorine atom gains one electron, and its oxidation number changes from zero to minus one. In electrovalent compounds, the oxidation number of an ion is the same as its electrical charge and often is referred to as the valence of the element.

The name of a compound of an element that may assume different oxidation states often specifies the oxidation state by a roman numeral following the name of the element, as in iron(II) oxide, FeO, or tin(IV) chloride, $SnCl_4$.

Atoms in which the outermost electron shell is occupied by three, four, or five electrons do not form ions readily because the gain or loss of the large number of electrons requires too much energy. Such atoms achieve a condition in which the valence shell interacts with eight electrons—without gaining sole possession of them—by sharing electrons with other atoms. Bonds formed in this way are called covalent, and the shared electrons remain in the region of space between the nuclei of the two atoms bonded together. Two atoms may each contribute one, two, or three electrons, forming single, double, or triple covalent bonds, and one atom may share electrons with several other atoms at the same time, linking whole groups of atoms together into molecules as simple as that of hydrogen or as complicated as those of proteins and nucleic acids, in which many thousands of atoms are linked together.

Covalent bonds

General similarities and differences. The elements belonging to Group VIa of the periodic table are characterized by electron configurations in which six electrons occupy the outermost shell. An atom having such an electronic structure tends to form a stable shell of eight electrons by adding two more, producing an ion that has a double negative charge. This tendency to form negatively charged ions, typical of nonmetallic elements, is quantitatively expressed in the properties of electronegativity (the assumption of partial negative charge when present in covalent combination) and electron affinity (the ability of a neutral atom to take up an electron, forming a negative ion). Both these properties decrease in intensity as the elements increase in atomic number and mass proceeding down column VIa of the periodic table. Oxygen has, except for fluorine, the highest electronegativity and electron affinity of any element; the values of these properties then decrease sharply for the remaining members of the group to the extent that tellurium and polonium are regarded as predominantly metallic in nature, tending to lose rather than gain electrons in compound formation.

As is the case within all groups of the table, the lightest element—the one of smallest atomic number—has extreme or exaggerated properties. Oxygen, because of the small size of its atom, the small number of electrons in its underlying shell, and the large number of protons in the nucleus relative to the atomic radius, has properties uniquely different from those of sulfur and the remaining chalcogens. Those elements behave in a reasonably predictable and periodic fashion, as will be shown.

Although even polonium exhibits the oxidation state −2 in forming a few binary compounds of the type MPo (in which M is a metal), the heavier chalcogens do not form the negative state readily, favouring positive states such as +2 and +4. All the elements in the group except oxygen may assume positive oxidation states, with the even values predominating, but the highest value, +6, is not a very stable one for the heaviest members. When this state is achieved, there is a strong driving force for the atom to return to a lower state, quite often to the elemental form. This tendency makes compounds containing Se(VI) and Te(VI) more powerful oxidizing agents than S(VI) compounds. Conversely, sulfides, selenides, and tellurides, in which the oxidation state is −2, are strong reducing agents, easily oxidized to the free elements.

Neither sulfur nor selenium, and most certainly not oxygen, forms purely ionic bonds to a nonmetal atom. Tellurium and polonium form a few compounds that are somewhat ionic; tellurium(IV) sulfate, $Te(SO_4)_2$, and polonium(II) sulfate, $PoSO_4$, are examples.

Stability of compounds

Another feature of the Group VIa elements that parallels trends generally shown in columns of the periodic table is the increasing stability of molecules having the composition $X(OH)_n$ as the size of the central atom, X, increases. There is no compound HO—O—OH, in which the central oxygen atom would have a positive oxidation state, a condition that it resists. The analogous sulfur compound HO—S—OH, although not known in the pure state, does have a few stable derivatives in the form of metal salts, the sulfoxylates. More highly hydroxylated compounds of sulfur, $S(OH)_4$ and $S(OH)_6$, also do not exist, not because of sulfur's resistance to a positive oxidation state but rather because of the high charge density of the S(IV) and S(VI) states (the large number of positive charges relative to the small diameter of the atom), which repels the electropositive hydrogen atoms, and the crowding that attends covalent bonding of six oxygen atoms to sulfur, favouring loss of water:

$$S(OH)_6 \longrightarrow (HO)_2SO_2 + 2H_2O.$$

orthosulfuric acid — sulfuric acid — water

As the size of the chalcogen atom increases, the stability of the hydroxylated compounds increases: the compound orthotelluric acid, $Te(OH)_6$, is capable of existence.

Catenation

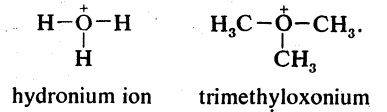
One of the most unusual properties of this family of elements is that of catenation (from Latin *catena*, "chain") or the bonding of an atom to another identical atom. Although oxygen shows this property only in the existence of ozone, sulfur is second only to carbon in exhibiting this mode of combination; the chalcogens beyond sulfur show it to diminishing degrees, polonium having no tendency to catenate. This type of bonding is found in the many ring systems of sulfur and selenium as well as in long zigzag chain structures. Catenation also occurs in the sulfanes and the metal polysulfides, compounds that have the formulas H_2S_x and M_2S_x, in which x may take the values of 2, 3, 4, or more, and M represents a singly charged metal ion. In comparing the catenation of sulfur atoms with that of carbon atoms, it may be noted that the number of molecular species having $(-S-)_x$ structures is very large, as is that of the analogous hydrocarbon compounds $(-CH_2-)_x$. The analogy between molecules containing rings of sulfur atoms and cyclic hydrocarbons is limited because only S_6 and S_8 have sufficient stability to permit proper comparison to be made. The general similarity extends to molecules of the form $Z(-S-)_xZ$ and $Z(-CH_2-)_xZ$, which are represented by compounds in which Z is H, SO_3H, and CF_3.

Covalent links between sulfur atoms have some of the character of multiple bonds—that is, more than one pair of electrons is shared, at least to some extent. Such interactions may involve overlap of p orbitals of one sulfur atom with d orbitals of another. Although not all investigators feel alike on the subject of d-orbital participation in the bonding of sulfur compounds, partial occupation of these orbitals is consistent with certain properties such as the colours of S_8 and S_2 molecules, the rigidity of chains and rings of sulfur atoms, and other features of the chemistry of sulfur compounds.

Similarities of sulfur and oxygen are exhibited in certain compounds in which these elements interchange for one another. Examples include sulfates and thiosulfates (such as Na_2SO_4 and $Na_2S_2O_3$), phosphates and thiophosphates (containing the ions PO_4^{3-}, PO_3S^{3-}, $PO_2S_2^{3-}$, POS_3^{3-}, and PS_4^{3-}), and a similar series of arsenates and thioarsenates.

Ores of heavy metals often are found as both sulfides, MS, and selenides, MSe, or even with MS_xSe_y structures. The similarity in structures as well as properties accounts for the chalcogens' being found together in nature.

The number of atoms to which an element of Group VIa can form covalent bonds increases from oxygen to sulfur. An oxygen atom usually combines with two other atoms, as in the compounds water (H_2O), oxygen fluoride (OF_2), or dimethyl ether ($H_3C-O-CH_3$); the unshared pairs of electrons and the partial negative charge on the oxygen atom in most of these compounds allows bonding to another atom, as in the hydronium ion or trimethyloxonium ion:



hydronium ion — trimethyloxonium ion

Heavier members of the group associate or coordinate with other atoms or groups of atoms in numbers commensurate with the size of both the chalcogen and the coordinating group. Thus, sulfur tetrafluoride (SF_4) and sulfur hexafluoride (SF_6) are stable compounds, although sulfur hexaiodide (SI_6) is not known because of the very large size of the iodine atom. A closely related property is that of anionic complex formation: there is little evidence for the ion SF_6^{2-}, but there are ions such as $TeCl_6^{2-}$, TeF_6^{2-}, and PoI_6^{2-}.

Isotopes

The known isotopes of each of the Group VIa elements are listed in Table 1. Consistent with a generality observed throughout the periodic system, isotopes of even mass number are more abundant than those of odd mass number. Each member of the group except polonium has several stable isotopes; oxygen-18 and sulfur-35 have been used as tracers in chemical analysis, and polonium-210 serves as a convenient source of alpha particles (nuclei of helium atoms) for nuclear reactors and nuclear batteries.

Table 1: Isotopes of the Group VIa Elements

	stable isotopes		unstable isotopes	
	mass	abundance (percentage)	mass	half-life
Oxygen	16	99.759	14	71 seconds
	17	0.037	15	2.07 minutes
	18	0.204	19	29 seconds
			20	14 seconds
Sulfur	32	95.0	30	1.4 seconds
	33	0.76	31	2.7 seconds
	34	4.22	35	88 days
	36	0.014	37	5.1 minutes
			38	2.78 hours
Selenium	74	0.87	70	44 minutes
	76	9.02	71	5 minutes
	77	7.58	72	8.4 days
	78	23.52	73†	42 minutes
	80	49.82		7.1 hours
	82	9.19	75	120 days
			77	17.5 seconds
			79	6.5×10^4 years
			81†	57 minutes
				18.6 minutes
			83†	70 seconds
				23 minutes
			84	3.3 minutes
			85	39 seconds
			87	16 seconds
Tellurium	120	0.089	114	16 minutes
	122	2.46	115	6 minutes
	123	0.87	116	2.5 hours
	124	4.61	117	61 minutes
	125	6.99	118	6 days
	126	18.71	119†	4.7 days
	128	31.79		16 hours
	130	34.48	121†	154 days
				17 days
			123	117 days
			125	58 days
			127†	109 days
				9.4 hours
			129†	34 days
				69 minutes
			131†	1.2 days
				25 minutes
			132	78 hours
			133†	50 minutes
				12.5 minutes
			134	42 minutes
Polonium*			192	0.5 second
			193	4 seconds
			194	0.5 second
			195	3 seconds
			196	6 seconds
			197	54 seconds
			198	1.7 minutes
			199	5 minutes
			200	11 minutes
			201	15.1 minutes
			202	45 minutes
			203	42 minutes
			204	3.6 hours
			205	1.8 hours
			206	8.8 days
			207†	2.8 seconds
				5.7 hours
			208	2.93 years
			209	103 years
			210	138.4 days
			211†	25 seconds
				0.52 second
			212	0.30×10^{-6} second
			213	4×10^{-6} second
			214	164×10^{-6} second
			215	0.0018 second
			216	0.15 second
			217	less than 10 seconds
			218	3.05 minutes

*No stable isotopes. †Two isotopes.

II. The chalcogens and their compounds

OXYGEN

Natural occurrence and distribution. As mentioned earlier, oxygen makes up about 46 percent of the mass of the crust of the Earth. In rocks, it is combined with metals and nonmetals in the form of oxides that are acidic (such as those of sulfur, carbon, aluminum, and phosphorus) or basic (such as those of calcium, magnesium, and iron) and as saltlike compounds that may be regarded as formed from the acidic and basic oxides, as sulfates, carbonates, silicates, aluminates, and phosphates. Plentiful as they are, these solid compounds are not useful as sources of oxygen, because separation of the element from its tight combinations with the metal atoms is too expensive.

Allotropy. Oxygen has two allotropic forms, diatomic (O_2) and triatomic (O_3, ozone). The properties of the diatomic form suggest that six electrons bond the atoms and two electrons remain unpaired, accounting for the paramagnetism of oxygen. The three atoms in the ozone molecule do not lie along a straight line.

Ozone may be produced from oxygen according to the equation:

$$3O_2 \longrightarrow 2O_3.$$

oxygen ozone

Production of ozone

The process, as written, is endothermic (energy must be provided to make it proceed); conversion of ozone back into diatomic oxygen is promoted by the presence of transition metals or their oxides. Pure oxygen is partly transformed into ozone by a silent electrical discharge; the reaction is also brought about by absorption of ultraviolet light of wavelengths around 250 nanometres (nm, the nanometre, equal to 10^{-9} metre); occurrence of this process in the upper atmosphere removes radiation that would be harmful to life on the surface of the Earth. The pungent odour of ozone is noticeable in confined areas in which there is sparking of electrical equipment, as in generator rooms. Ozone is light blue; its density is 1.658 times that of air, and it has a boiling point of −112° C at atmospheric pressure.

Ozone is a powerful oxidizing agent, capable of converting sulfur dioxide to sulfur trioxide, sulfides to sulfates, iodides to iodine (providing an analytical method for its estimation), and many organic compounds to oxygenated derivatives such as aldehydes and acids. The conversion by ozone of hydrocarbons from automotive exhaust gases to these acids and aldehydes contributes to the irritating nature of smog. Commercially, ozone has been used as a chemical reagent, as a disinfectant, in sewage treatment, water purification, and bleaching textiles.

Preparative methods. Production methods chosen for oxygen depend upon the quantity of the element desired. Laboratory procedures include the following:

1. Thermal decomposition of certain salts, such as potassium chlorate or potassium nitrate:

$$2KClO_3 \longrightarrow 2KCl + 3O_2$$

potassium chlorate — potassium chloride — oxygen

$$2KNO_3 \longrightarrow 2KNO_2 + O_2.$$

potassium nitrate — potassium nitrite — oxygen

The decomposition of potassium chlorate is catalyzed by oxides of transition metals; manganese dioxide (pyrolusite, MnO_2) is frequently used. The temperature necessary to effect the evolution of oxygen is reduced from 400°C to 250° by the catalyst.

2. Thermal decomposition of oxides of heavy metals:

$$2HgO \longrightarrow 2Hg + O_2$$

mercury(II) oxide — mercury — oxygen

$$2Ag_2O \longrightarrow 4Ag + O_2.$$

silver(I) oxide — silver — oxygen

Scheele and Priestley used mercury(II) oxide in their preparations of oxygen.

3. Thermal decomposition of metal peroxides or of hydrogen peroxide:

$$2BaO + O_2 \xrightarrow{\text{heat}} 2BaO_2$$

barium oxide — oxygen (from air) — barium peroxide

$$\underset{\text{barium peroxide}}{2BaO_2} \xrightarrow{\text{heat}} \underset{\text{barium oxide}}{2BaO} + \underset{\text{oxygen}}{O_2}$$

$$\underset{\text{barium peroxide}}{BaO_2} + \underset{\text{sulfuric acid}}{H_2SO_4} \longrightarrow \underset{\text{hydrogen peroxide}}{H_2O_2} + \underset{\text{barium sulfate}}{BaSO_4}$$

$$\underset{\text{hydrogen peroxide}}{2H_2O_2} \xrightarrow[\text{catalyst}]{\text{heat or}} \underset{\text{water}}{2H_2O} + \underset{\text{oxygen}}{O_2}$$

An early commercial procedure for isolating oxygen from the atmosphere or for manufacture of hydrogen peroxide depended on the formation of barium peroxide from the oxide as shown in the equations.

4. Electrolysis of water containing small proportions of salts or acids to allow conduction of the electric current:

$$\underset{\text{water}}{2H_2O}. \xrightarrow[\text{electric current}]{H_2SO_4} \underset{\text{hydrogen}}{2H_2} + \underset{\text{oxygen}}{O_2}.$$

Oxygen from the air

Commercial production and use. When required in tonnage quantities, oxygen is prepared by the fractional distillation of liquid air. The process takes advantage of the fact that when a compressed gas is allowed to expand, it cools. Major steps in the operation include the following: (1) Air is filtered to remove particulates; (2) moisture and carbon dioxide are removed by absorption in alkali; (3) the air is compressed and the heat of compression removed by ordinary cooling procedures; (4) the compressed and cooled air is passed into coils contained in a chamber; (5) a portion of the compressed air (at about 200 atmospheres pressure) is allowed to expand in the chamber, cooling the coils; (6) the expanded gas is returned to the compressor with multiple subsequent expansion and compression steps resulting finally in liquefaction of the compressed air at a temperature of −196° C; (7) the liquid air is allowed to warm to distill first the light rare gases, then the nitrogen, leaving liquid oxygen. Multiple fractionations will produce a product pure enough (99.5 percent) for most industrial purposes.

The steel industry is the largest consumer of pure oxygen in "blowing" high carbon steel—that is, volatilizing carbon dioxide as well as other nonmetal impurities in a more rapid and more easily controlled process than if air were used. The treatment of sewage by oxygen holds promise for more efficient treatment of liquid effluents than other chemical processes. Incineration of wastes in closed systems using pure oxygen had become important by the early 1970s. The so-called LOX of rocket oxidizer fuels is liquid oxygen; the consumption of the element in this way depends upon the activity of space programs. One Saturn V engine consumes 590,000 gallons (2,200,000 litres) of liquid oxygen in launching a 6.5-million-pound rocket. Pure oxygen is used in a multitude of breathing devices, submarines, diving bells, and in hospitals.

Chemical properties and reactions. The large values of the electronegativity and the electron affinity of oxygen are typical of elements that show only nonmetallic behaviour. In all of its compounds, oxygen assumes a negative oxidation state as is expected from the two half-filled outer orbitals. When these orbitals are filled by electron transfer, the oxide ion O^{2-} is created. In peroxides (species containing the ion O_2^{2-}) it is assumed that each oxygen has a charge of −1. This property of accepting electrons by complete or partial transfer defines an oxidizing agent. When such an agent reacts with an electron-donating substance, its own oxidation state is lowered. The change (lowering), from the zero to the −2 state in the case of oxygen, is called a reduction. Oxygen may be thought of as the "original" oxidizing agent, the nomenclature used to describe oxidation and reduction being based upon this behaviour typical of oxygen.

Oxygen in combination with itself. As described in the section on allotropy, oxygen forms the diatomic species, O_2, under normal conditions and, as well, the triatomic species ozone, O_3. There is some evidence for a very unstable tetratomic species, O_4. In the molecular diatomic form there are two unpaired electrons that lie in antibonding orbitals. The paramagnetic behaviour of oxygen confirms the presence of such electrons.

The intense reactivity of ozone is sometimes explained by suggesting that one of the three oxygen atoms is in an "atomic" state; on reacting, this atom is dissociated from the O_3 molecule, leaving molecular oxygen.

The molecular species, O_2, is not especially reactive at normal (ambient) temperatures and pressures. The atomic species, O, is far more reactive. The energy of dissociation ($O_2 \rightarrow 2O$) is large at 117.2 kilocalories per mole. The reduction of the molecular species is a one-electron process in nonaqueous solutions but a two-electron process in aqueous solutions:

$$\underset{\text{oxygen}}{O_2} + \underset{\text{electron}}{e^-} \rightleftharpoons \underset{\text{superoxide ion}}{O_2^-}$$

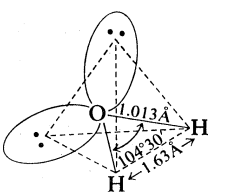
$$\underset{\text{oxygen}}{O_2} + \underset{\text{water}}{H_2O} + \underset{\text{electrons}}{2e^-} \rightleftharpoons \underset{\text{hydroperoxide ion}}{HO_2^-} + \underset{\text{hydroxide ion}}{OH^-}.$$

Reactions of oxygen forming oxides, peroxides, and superoxides may be broadly classified according to periodic table relationships. The strong metals, or those with relatively small ionization energies and large oxidation potentials (Groups Ia and IIa) form ionic oxides (M_xO_y) in which electrons are transferred from the metal to the oxygen. The most active of these metals form peroxides (*e.g.*, Na_2O_2) and superoxides (*e.g.*, KO_2). The metalloids, such as beryllium, silicon, arsenic, antimony, and tellurium, form polymeric oxide species, while the nonmetallic elements (those of Groups VIa and VIIa) form gaseous covalent molecules.

Compounds with hydrogen. The principal compounds containing hydrogen and oxygen are water, H_2O, and hydrogen peroxide, H_2O_2; the hydroxyl radical, OH, is an unstable species that can exist in certain circumstances (see also the article WATER).

Formation of water

The reaction between hydrogen and oxygen takes place extremely slowly at ordinary temperatures and in the absence of catalysts. The reaction may be accelerated by certain catalysts such as the metals of Group VIII. The molecule of water is nonlinear, the angle between the two covalent bonds being 104° 30′. The high electronegativity of oxygen induces a strong negative electrical charge on the oxygen atom with a corresponding positive charge distributed between the two hydrogen atoms. The bent molecule is described in the following diagram, showing the two lone (unshared) pairs of electrons extending in space toward the apices of a tetrahedron.



This model correctly implies that water should possess a dipole moment—*i.e.*, that the centres of positive and negative electrical charge are separated in space. This property, coupled with a large dielectric constant (the ability to decrease the attraction between oppositely charged particles), makes water an effective solvent for substances that are ionic or that themselves possess dipole moments. The ability of water to interact with other molecules in this way is important in understanding biological reactions as well as the structures of large polymer molecules in nature.

Water serves as a medium for characterizing metal and nonmetal oxides. Metal oxides react with water to liberate hydroxide ion, OH^-:

$$MO + H_2O \longrightarrow M^{2+} + 2OH^-.$$

metal oxide — water — metal ion — hydroxide ion

If the final concentration of the hydroxide ion in the solution is greater than the hydrogen ion concentration, the solution is said to be basic. The more soluble the metal oxide, the stronger its basic behaviour.

Nonmetal oxides in general react with water to provide an excess of hydrogen ions, H^+, over hydroxide ions, OH^-:

$$XO + H_2O \longrightarrow XO_2^{2-} + 2H^+.$$

nonmetal oxide — water — oxyanion — hydrogen ion

In addition to its reactions with metal oxides and nonmetal oxides, water also reacts with some metals and nonmetals themselves. The strongly metallic elements (the oxides of which form strong bases) in general react directly with water to liberate hydrogen and form the metal hydroxide as shown by the reaction with the element calcium:

$$Ca + 2H_2O \longrightarrow Ca(OH)_2 + H_2.$$

calcium — water — calcium hydroxide — hydrogen

Nonmetals dissolve, forming acids as illustrated by the nonmetal chlorine:

$$Cl_2 + H_2O \longrightarrow HCl + HOCl.$$

chlorine — water — hydrochloric acid — hypochlorous acid

Less active metals and some nonmetals, such as carbon, react with steam to form hydrogen:

$$C + H_2O \longrightarrow CO + H_2.$$

carbon — water (steam) — carbon monoxide — hydrogen

Hydrogen peroxide

The direct combination of oxygen and hydrogen does not provide a satisfactory method of preparing hydrogen peroxide. Commercial preparation is accomplished through the reaction of water with peroxy salts or peroxides. The following equations represent the electrolytic formation of peroxydisulfuric acid and its hydrolysis to hydrogen peroxide:

$$2H_2SO_4 \xrightarrow[\text{current}]{\text{electric}} H_2S_2O_8 + H_2$$

sulfuric acid — peroxydisulfuric acid — hydrogen

$$H_2S_2O_8 + H_2O \longrightarrow H_2SO_4 + H_2SO_5$$

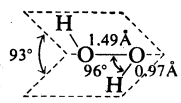
water — peroxymonosulfuric acid

$$H_2SO_5 + H_2O \longrightarrow H_2O_2 + H_2SO_4.$$

hydrogen peroxide

Concentration is accomplished by distillation of the H_2O_2 from the solution of sulfuric acid and then to the pure or nearly pure state by multistage vacuum fractionation. At high purity the compound is sensitive to decomposition by transition metal ions. A nonelectrolytic procedure involves the reaction of anthraquinone with oxygen to form a peroxide that undergoes decomposition to form hydrogen peroxide. The hydrogen peroxide formed is extracted by water from the organic solution containing the anthraquinone.

The structure of hydrogen peroxide is best represented by a twisted H—O—O—H molecule, some 93° out of planarity and with an O—O—H angle of 96°:



Chemically, the compound is both an oxidizing and a reducing agent. With strong oxidizing agents, such as permanganate ion, it is converted to oxygen:

$$5H_2O_2 + 2MnO_4^- + 6H^+ \longrightarrow 2Mn^{2+} + 5O_2 + 8H_2O.$$

hydrogen peroxide — permanganate ion — hydrogen ion — manganese(II) ion — oxygen — water

The more usual behaviour of hydrogen peroxide is as an oxidizing agent. As such it has commercial application in water purification and bleaching. In the pure form it has been used as a rocket propellant fuel. The 30 percent solution is caustic enough to burn skin severely.

Metal oxides

Reactions with metals. The nature of the oxide formed when a metal reacts with oxygen depends upon such factors as the concentration of the oxygen, the ionization energy of the metal, the temperature, and the electron configuration of the metal (*e.g.*, the possibility of polyvalency).

Group Ia metals (lithium, sodium, potassium, rubidium, cesium) all react at room temperature with oxygen to form ionic oxides having the general formula M_2O.

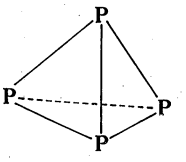
Certain strong metals also form peroxides and superoxides in which oxygen exists in the form of the ions O_2^{2-} and O_2^-. Examples of peroxides are those of sodium and barium, Na_2O_2 and BaO_2, respectively; the superoxides of potassium and cesium, KO_2 and CsO_2, are typical of that group. Acids react with these peroxy species (*i.e.*, compounds containing O—O bonds) to give hydrogen peroxide, H_2O_2, and a superoxide, HO_2, respectively.

Still another of the "poly-oxy" derivative of metals is potassium ozonide, KO_3, formed from ozone and solid potassium hydroxide.

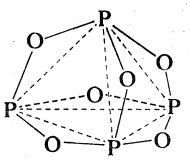
Oxides of the less active metals typified by those of Groups IIIa (aluminum, gallium, indium, thallium), IVa (germanium, tin, lead), Va (arsenic, antimony, bismuth), and VIa (tellurium, polonium) are far less ionic and often variable in their formulation or stoichiometry because of the possibility of more than one valence state. If more than one oxide of a metal is possible, the oxide of the lower oxidation state will be the more ionic, the more soluble in water, the stronger base, and less likely to be soluble in excess of a strong base, such as sodium hydroxide solution. Oxides of such metals as lead, tin, and antimony are likely to be nonstoichiometric, that is, have atomic ratios of metal to oxygen, not those of small whole numbers. Such also is the case for many transition metal oxides. Unique electrical properties may accompany the nonstoichiometric composition. When small amounts of certain metals are added to oxides of transition metals such as tungsten, molybdenum, chromium, or tantalum, unique electrical conducting properties are found, which make the substances useful in the electronics field; the tungsten bronzes, composed of tungsten trioxide, WO_3, and sodium, are examples of such materials.

Reactions with nonmetals. Because atoms of nonmetals have very little tendency to give up electrons in the formation of chemical bonds, the bonding between nonmetals is likely to be covalent as the result of sharing of electrons. Nonmetals of low atomic number, such as oxygen, form multiple bonds with other nonmetals as noted in such oxides as carbon monoxide (CO), nitric oxide (NO), chlorine monoxide (Cl_2O), sulfur dioxide (SO_2), and sulfur trioxide (SO_3). When the number of electrons in a compound is odd, the molecule will of necessity have a lone or unshared electron. Such a molecule is referred to as a radical or odd molecule.

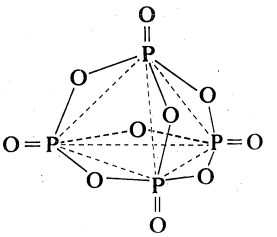
There is little tendency for oxides of light nonmetals to form solid or polymeric systems, in which many small molecules join together to form a very large one. The oxides of the heavier nonmetals and metalloids (*e.g.*, boron, silicon, arsenic, antimony, tellurium, and polonium) are solids and usually polymeric. Each of them attempts to form bonds to four surrounding atoms (fulfill a coordination number of four), often by forming networks or chains in which each metal atom is bonded to four oxygen atoms arranged in a tetrahedron. Although the element phosphorus is strongly nonmetallic, it forms oxides that are solid and in which the tetrahedral orientation of the phosphorus atoms is maintained as shown:



phosphorus, P_4



phosphorus (III) oxide, P_4O_6



phosphorus (V) oxide, P_4O_{10}

Oxides of nonmetals react with water to form acids, as previously mentioned. Certain of the oxides have very strong affinities for water and thus serve as dehydrating agents, rapidly removing water from combination with other substances. Phosphorus(V) oxide, P_4O_{10}, and sulfur(VI) oxide, SO_3, are outstanding examples.

Oxides of the noble gases. The fact that the heavy rare gases (*e.g.*, xenon, radon) will form compounds was definitely established in 1962.

Although only fluorine reacts directly with xenon, oxides form when the fluorides react with water:

$$3XeF_4 + 6H_2O \longrightarrow XeO_3 + 2Xe + \tfrac{3}{2}O_2 + 12HF$$

xenon tetrafluoride — water — xenon trioxide — xenon — oxygen — hydrogen fluoride

$$XeF_6 + 3H_2O \longrightarrow XeO_3 + 6HF.$$

xenon hexafluoride — water — xenon trioxide — hydrogen fluoride

SULFUR

Natural occurrence and distribution. Many important metal ores are compounds of sulfur, either sulfides or sulfates. Some important examples are galena (lead sulfide, PbS), blende (zinc sulfide, ZnS), pyrite (iron disulfide, FeS_2), chalcopyrite (copper iron sulfide, $CuFeS_2$), gypsum (calcium sulfate dihydrate, $CaSO_4 \cdot 2H_2O$) and barite (barium sulfate, $BaSO_4$). The sulfide ores are valued chiefly for their metal content, although a process developed in the 18th century for making sulfuric acid utilized sulfur dioxide obtained by burning pyrite.

Allotropy. In sulfur, allotropy arises from two sources: (1) the different modes of bonding atoms into a single molecule and (2) packing of polyatomic sulfur molecules into different crystalline and amorphous forms. Some 30 allotropic forms of sulfur have been reported, but some of these probably represent mixtures. Only eight of the 30 seem to be unique; five contain rings of sulfur atoms and the others contain chains.

In the rhombohedral allotrope, designated ρ-sulfur, the molecules are composed of rings of six sulfur atoms. This form is prepared by treating sodium thiosulfate with cold, concentrated hydrochloric acid, extracting the residue with toluene, and evaporating the solution to give hexagonal crystals. ρ-sulfur is unstable, eventually reverting to orthorhombic sulfur (α-sulfur).

A second general allotropic class of sulfur is that of the eight-membered ring molecules, three crystalline forms of which have been well characterized. One is the orthorhombic (often improperly called rhombic) form, α-sulfur. It is stable at temperatures below 96° C (205° F). Another of the crystalline S_8 ring allotropes is the monoclinic or β-form, in which two of the axes of the crystal are perpendicular, but the third forms an oblique angle with the first two. There are still some uncertainties concerning its structure; this modification is stable from 96° C to the melting point, 118.9° C (246° F). A second monoclinic cyclooctasulfur allotrope is the γ-form, unstable at all temperatures, quickly transforming to α-sulfur.

An orthorhombic modification, S_{12} ring molecules, and still another unstable S_{10} ring allotrope are reported. The latter reverts to polymeric sulfur and S_8. At temperatures above 96° C, the α-allotrope changes into the β-allotrope. If enough time is allowed for this transition to occur completely, further heating causes melting to occur at 118.9° C; but if the α-form is heated so rapidly that the transformation to β-form does not have time to occur, the α-form melts at 112.8° C (235° F).

Effects of heating sulfur

Just above its melting point, sulfur is a yellow, transparent, mobile liquid. Upon further heating, the viscosity of the liquid decreases gradually to a minimum at about 157° C (315° F), but then rapidly increases, reaching a maximum value at about 187° C (369° F); between this temperature and the boiling point of 444.6° C (832° F), the viscosity decreases. The colour also changes, deepening from yellow through dark red, and, finally, to black at about 250° C (280° F). The variations in both colour and viscosity are considered to result from changes in the molecular structure. A decrease in viscosity as temperature increases is typical of liquids, but the increase in the viscosity of sulfur above 157° C probably is caused by rupturing of the eight-membered rings of sulfur atoms to form reactive S_8 units that join together in long chains containing many thousands of atoms. The liquid then assumes the high viscosity characteristic of such structures. At a sufficiently high temperature, all of the cyclic molecules are broken, and the length of the chains reaches a maximum. Beyond that temperature, the chains break down into small fragments. Upon vaporization, cyclic molecules (S_8 and S_6) are formed again; at about 900° C (1,650° F), S_2 is the predominant form; finally, monatomic sulfur is formed at temperatures above 1,800° C (3,300° F).

Commercial production and use. Elemental sulfur is found in volcanic regions as a deposit formed by the emission of hydrogen sulfide, followed by aerial oxidation to the element. Underground deposits of sulfur associated with salt domes in limestone rock provide a substantial portion of the world's supply of the element. These domes are located in the Louisiana swamplands of the United States and offshore in the Gulf of Mexico.

Frasch process. Herman Frasch, a German-born U.S. chemist, developed a process, known by his name, for extracting and raising pure sulfur from these deposits, which may be anywhere from a few hundred feet to thousands of feet below the surface. Usual underground mining procedures are inapplicable since highly poisonous hydrogen sulfide gas accompanies the element in the domes.

The Frasch process takes advantage of the low melting point of sulfur, about 112° C (234° F). Water heated above this temperature (under pressure) is pumped down one of three concentric pipes, melting the sulfur. Compressed air is then forced down an inner pipe forming a froth of the molten sulfur, which is then forced up through the middle concentric pipe to the surface, where it is either pumped into bins for storage or into barges or ships for transporting to industrial areas for conversion, for the most part, to sulfuric acid. The availability of cheap, very pure sulfur from this process eliminated much of the need to mine sulfur from sulfides and volcanic sulfur-bearing rock for many years. By the mid-20th century the purification of sour (high sulfur-content) petroleum and improved methods for obtaining sulfur from metal sulfides had increased sulfur production from non-Frasch sources.

Other recovery processes. A few of the non-Frasch processes for sulfur production may be mentioned.

(1) Sulfur-bearing rock is piled into mounds. Shafts are bored vertically and fires set at the top of the shafts. The burning sulfur provides sufficient heat to melt the elemental sulfur in the rock layers below, and it flows out at the bottom of the pile. This is an old process, still used to some extent in Sicily. The product is of low purity and must be refined by distillation. The air pollution in the area of the process is so great that its operation is limited

to certain times of the year when prevailing winds will carry the fumes away from populated areas.

(2) Rock bearing sulfur is treated with superheated water in retorts, melting the sulfur, which flows out. This process is a modification of the Frasch method.

(3) Sulfates (such as gypsum or barite) may be treated with carbon at high temperatures, forming the metal sulfides, CaS or BaS (the Chance–Claus process). The metal sulfides can be treated with acid, generating hydrogen sulfide, which in turn can be burned to give elemental sulfur.

(4) Tremendous tonnages of sulfur are available from smelter operations and from power production by combustion of fossil and sour petroleum fuels, some of which contain as much as 4 percent sulfur. Thus, generation of electrical power and heat represent a major source of atmospheric pollution by sulfur dioxide. Unfortunately, recovery and purification of sulfur dioxide from stack gases are expensive operations.

Wherever such metals as lead, zinc, copper, cadmium, or nickel (among others) are processed, much of the sulfuric acid needed in the metallurgical operations may be obtained on the site by converting sulfur dioxide, produced by roasting the ores, to sulfur trioxide, SO_3, and thence to sulfuric acid.

Sulfur available in bulk from commercial production usually is more than 99 percent pure, and some grades contain 99.9 percent sulfur. For research purposes, the proportion of impurities has been reduced to as little as one part in 10,000,000 by the application of procedures such as zone melting, column chromatography, electrolysis, or fractional distillation.

Uses of sulfur. Sulfur is so widely used in industrial processes that its consumption often is regarded as a reliable indicator of industrial activity and the state of the national economy. Approximately six-sevenths of all the sulfur produced is converted into sulfuric acid, for which the largest single use is in the manufacture of fertilizers (phosphates and ammonium sulfate). Other important uses include the production of pigments, detergents, fibres, petroleum products, sheet metal, explosives, and storage batteries; hundreds of other applications are known. Sulfur not converted to sulfuric acid is used in making paper, insecticides, fungicides, dyestuffs, carbon disulfide (a solvent employed in making rayon, cellophane, and industrial chemicals), and numerous other products.

SELENIUM

Occurrence and distribution. The proportion of selenium in the Earth's crust is about 10^{-5} to 10^{-6} percent. It has been obtained mainly from the anode slimes (deposits and residual materials from the anode) in electrolytic refining of copper and nickel. Other sources are the flue dusts in copper and lead production and the gases formed in roasting pyrites. Selenium accompanies copper in the refining of that metal: about 40 percent of the selenium present in the original ore may concentrate in copper deposited in electrolytic processes. About 1.5 kilograms of selenium can be obtained from a ton of smelted copper.

Allotropy. The allotropy of selenium is not as extensive as that of sulfur, and the allotropes have not been studied as thoroughly. Only two crystalline varieties of selenium are composed of cyclic Se_8 molecules: designated α and β, both exist as red monoclinic crystals. A gray allotrope having metallic properties is formed by keeping any of the other forms at 200° –220° C (400°–430° F).

An amorphous (noncrystalline), red, powdery form of selenium results when a solution of selenious acid or one of its salts is treated with sulfur dioxide. If the solutions are very dilute, extremely fine particles of this variety yield a transparent red colloidal suspension. Clear red glass results from a similar process that occurs when molten glass containing selenites is treated with carbon. A glassy, almost black variety of selenium is formed by rapid cooling of other modifications from temperatures above 200° C. Conversion of this vitreous form to the red, crystalline allotropes takes place upon heating it above 90° C (195° F) or upon keeping it in contact with organic solvents, such as chloroform, ethanol, or benzene.

Preparation. Pure selenium is obtained from the slimes and sludges formed in producing sulfuric acid. The impure red selenium is dissolved in sulfuric acid in the presence of an oxidizing agent, such as potassium nitrate or certain manganese compounds. Both selenious acid, H_2SeO_3, and selenic acid, H_2SeO_4, are formed and can be leached from residual insoluble material. Other methods utilize oxidation by air (roasting) and heating with sodium carbonate to give soluble sodium selenite, $Na_2SeO_3 \cdot 5H_2O$, and sodium selenate, Na_2SeO_4. Chlorine may also be employed: its action upon metal selenides produces volatile compounds including selenium dichloride, $SeCl_2$; selenium tetrachloride, $SeCl_4$; diselenium dichloride, Se_2Cl_2; and selenium oxychloride, $SeOCl_2$. In one process, these selenium compounds are converted by water to selenious acid. The selenium is finally recovered by treating the selenious acid with sulfur dioxide.

Selenium is a common component of ores valued for their content of silver or copper; it becomes concentrated in the slimes deposited during electrolytic purification of the metals. Methods have been developed to separate selenium from these slimes, which also contain some silver and copper. Melting the slime forms silver selenide, Ag_2Se, and copper(I) selenide, Cu_2Se. Treatment of these selenides with hypochlorous acid, HOCl, gives soluble selenites and selenates, which can be reduced with sulfur dioxide. Final purification of selenium is accomplished by repeated distillation.

Physical–electrical properties. The most outstanding physical property of crystalline selenium is its photoconductivity: on illumination, the electrical conductivity increases more than 1,000-fold. This phenomenon results from the promotion or excitation of relatively loosely held electrons by light to higher energy states (called the conduction levels), permitting electron migration and, thus, electrical conductivity. In contrast the electrons of typical metals are already in conduction levels or bands, able to flow under the influence of an electromotive force.

Photoconductivity of selenium

The electrical resistivity of selenium varies over a tremendous range, depending upon such variables as the nature of the allotrope, impurities, the method of refining, temperature, and pressure. Most metals are insoluble in selenium, and nonmetallic impurities increase the resistivity.

Illumination of crystalline selenium for 0.001 second increases its conductivity by a factor of 10 to 15 times. Red light is more effective than light of shorter wave length.

Advantage is taken of these photoelectric and photosensitivity properties of selenium in the construction of a variety of devices that can translate variations in light intensity into electric current and thence to visual, magnetic, or mechanical effects. Alarm devices, mechanical opening and closing devices, safety systems, television, sound films, and xerography depend upon the semiconducting property and photosensitivity of selenium. Rectification of alternating electrical current (conversion into direct current) has for years been accomplished by selenium-controlled devices. Many photocell applications using selenium have been replaced by other devices using materials more sensitive, more readily available, and more easily fabricated than selenium.

TELLURIUM

Occurrence and preparation. The demand for tellurium does not match that for selenium. The two elements are found together in many ores; they may be isolated by employing the processes described in connection with selenium, obtaining solutions containing salts of both selenious and tellurous acids, H_2SeO_3 and H_2TeO_3. Upon treatment of these solutions with sulfuric acid, tellurium dioxide, TeO_2, separates because of its low solubility, while the selenious acid remains dissolved. The tellurium dioxide can be converted into elemental tellurium by treatment with sulfur dioxide; an electrolytic process is used to purify the product.

Physical and chemical properties. In tellurium, the covalent bonding necessary to provide large ring- and chain-molecules by catenation is almost nonexistent. The element crystallizes in the rhombohedral form. It is silvery white and isomorphous with gray selenium—that is, the structure and dimensions of the crystals are very similar. It is brittle but not very hard. The tellurium atoms form spiral chains in the crystal with Te–Te distances of 3.74 angstroms. (One angstrom [Å] equals 10^{-10} metre.) There are no good solvents for tellurium, although certain compounds oxidize or reduce the element to soluble substances. The photoconductivity of tellurium is not as pronounced as that of selenium, and tellurium does not have major industrial uses.

POLONIUM

The existence of polonium in pitchblende, an ore of uranium, was noted by the Curies. Polonium is extremely rare, even in pitchblende: 1,000 tons of the ore must be processed to obtain 40 milligrams of polonium. In the chemical isolation, the ore is treated with hydrochloric acid, and the resulting solution is heated with hydrogen sulfide to precipitate polonium monosulfide, PoS, along with other metal sulfides, such as that of bismuth, Bi_2S_3, which resembles polonium monosulfide closely in chemical behaviour, though it is less soluble. Because of the difference in solubility, repeated partial precipitation of the mixture of sulfides concentrates the polonium in the more soluble fraction, while the bismuth accumulates in the less soluble portions. The difference in solubility is small, however, and the process must be repeated many times to achieve a complete separation. Purification is accomplished by electrolytic deposition.

Two modifications of polonium are known, an α- and a β-form, both of which are stable at room temperature and possess metallic character. The fact that its electrical conductivity decreases as the temperature increases places polonium among the metals rather than the metalloids or nonmetals.

COMPOUNDS OF THE CHALCOGENS

The hydrogen chalcogenides. Hydrogen sulfide, H_2S; hydrogen selenide, H_2Se; and hydrogen telluride, H_2Te; are all colourless, evil-smelling, intensely poisonous gases. Their physiological action is one of interference with enzymes that facilitate biologically essential processes, such as respiration. The selenide is a cumulative poison, even in small concentrations, and causes body and respiratory odours that persist for long periods. Table 2 lists some numerical data on these compounds, including water for comparison.

Table 2: Properties of the Hydrogen Chalcogenides

	H_2O	H_2S	H_2Se	H_2Te
Solubility, moles/l 760 mm, 25° C		0.102	0.08415	~0.09
Ionization constant, 25° C				
K_1	$K_w = 1.27 \times 10^{-14}$	1.15×10^{-7}	1.88×10^{-4}	2.27×10^{-3}
K_2		$\sim 10^{-14}$	$\sim 10^{-10}$	$\sim 10^{-5}$
Melting point, °C	0.0	−85.5	−65.9	−49
Boiling point, °C	100.0	−60.7	−41.2	−2
Heat of vaporization, cal/mole	9,715	4,463	4,760	5,700
Heat of formation, $\Delta H°$, cal/mole	−6,835	−4,800	+18,500	+34,200
Heat of fusion, cal/mole (at mp)	1,430	568	—	—
Free energy of formation, $\Delta G°$, cal/mole	−56,720	−78,600	+2,370	+31,000
Critical temperature, °C	366	100.4	137	—
Density, g/cm³ (at bp)	0.958	0.993	2.004	2.650

Hydrogen sulfide

Hydrogen sulfide is preparable by direct combination of the elements, although the reaction is slow. Sulfides of active metals (such as those of small atomic number in Groups Ia, IIa, and IIIa of the periodic table) react vigorously with acids to liberate hydrogen sufide. Iron(II) sulfide, FeS also reacts with nonoxidizing acids such as hydrochloric to form hydrogen sulfide. Hydrogen selenide and telluride are formed similarly.

Aqueous solutions of hydrogen chalcogenides are weak acids; the acidity increases as the atomic number of the chalcogen increases.

Reactions with metals. Metal chalcogenides, though bearing some structural resemblance to oxides, are less ionic. They are also less water-soluble and in general more intensely coloured. Sulfides of the strong metals (Groups Ia and IIa) are somewhat water-soluble, though the process of solution is likely to be slow and accompanied by hydrolysis, a competition between the sulfide ion and the hydroxide ion for the proton:

$$\underset{\text{metal sulfide}}{MS} \rightleftharpoons \underset{\text{metal ion}}{M^{2+}} + \underset{\text{sulfide ion}}{S^{2-}}$$

$$\underset{\text{sulfide ion}}{S^{2-}} + \underset{\text{water}}{H_2O} \rightleftharpoons \underset{\text{hydrosulfide ion}}{HS^-} + \underset{\text{hydroxide ion}}{HO^-}.$$

A solution of an alkali metal sulfide is, therefore, quite strongly basic. These sulfides have the capacity of dissolving powdered elemental sulfur to form polysulfides: $MS + xS \rightarrow MS_{1+x}$, in which x may be a large number, although in the well-characterized polysulfides the value is usually less than 6. When these polysulfides are treated with strong nonoxidizing acids (as mentioned), hydrogen sulfide forms as well as the hydrogen polysulfides (sulfanes), H_2S_x:

$$\underset{\substack{\text{sodium}\\\text{monosulfide}}}{Na_2S} + \underset{\text{sulfur}}{3S} \longrightarrow \underset{\substack{\text{sodium}\\\text{tetrasulfide}}}{Na_2S_4}$$

$$\underset{\substack{\text{sodium}\\\text{tetrasulfide}}}{Na_2S_4} + \underset{\substack{\text{hydrogen}\\\text{ion}}}{2H^+} \longrightarrow \underset{\text{sodium ion}}{2Na^+} + \underset{\substack{\text{hydrogen}\\\text{tetrasulfide}}}{H_2S_4}.$$

The disulfide H_2S_2 has some resemblance to hydrogen peroxide, H_2O_2, in being both an oxidizing and a reducing agent and having a bent or twisted structure. The higher sulfanes are reddish or yellow-red liquids.

Ores of such metals as iron, nickel, copper, cobalt, zinc, and lead often occur as sulfides and occasionally as mixed chalcogenides, having the general formula MS_xSe_y. Their metallurgy provides a considerable fraction of the world sulfur supply through the recovery of sulfur dioxide formed in the process of roasting.

Iron(II) sulfide occurs frequently as both the simple sulfide, FeS, and the disulfide, FeS_2 (pyrite). Just as the atoms in certain metal oxides do not occur in stoichiometric (small whole number) ratios, the sulfides have deficiencies in their structures, leading to nonstoichiometric ratios. Examples are found in the chromium–sulfur system with values ranging from 0.69 to 0.67 atom chromium to 1 of sulfur (close to Cr_2S_3) and in iron(II) sulfide with an Fe:S atomic ratio of 0.858:1.

Oxides of the chalcogens. The only reactions between elements in Group VIa are those of oxygen with the remaining members of the group.

Suboxides. The suboxides of sulfur are disulfur monoxide, S_2O; sulfur monoxide, SO; and sulfur sesquioxide, S_2O_3. The first of these is said to form from sulfur and sulfur dioxide in a glow discharge or from sulfur vapour and active oxygen. The sulfur(II) oxide (sulfur monoxide), SO, is not well characterized and may be merely a mixture of sulfur and sulfur dioxide or of disulfur monoxide, S_2O, and sulfur dioxide, SO_2. There is some evidence for the existence of the sesquioxide, S_2O_3, in mixtures of sulfur powder and anhydrous sulfur trioxide, SO_3. There appear to be no suboxides of selenium or tellurium. A black "oxide," possibly TeO, is reported but is likely to be a mixture of tellurium dioxide, TeO_2, and tellurium, Te.

Dioxides. There are examples of all the chalcogen dioxides; in fact, sulfur dioxide is produced in greater

quantity than almost any other chemical as a precursor of sulfur trioxide and thence sulfuric acid. As an air pollutant it ranks, with carbon monoxide, very high on the list of undesirable species.

Sulfur dioxide is a colourless, pungent gas, boiling at −10° C (14° F) and melting at −72.7° C (−98.9° F). The geometry of the molecule is "bent" with an O—S—O angle of 120° and a sulfur–oxygen bond distance of 1.43 Å. Even at its freezing temperature, the compound exists as discrete molecules.

Selenium dioxide, SeO_2, is a white, solid, chainlike polymeric substance, subliming (passing directly from the solid to the vapour state without melting) readily at 315° C. Both tellurium dioxide, TeO_2, and polonium dioxide, PoO_2, are solid polymeric crystalline materials of which the structures probably contain tetrahedral units. Sulfur burns with a bright blue flame, intensifying in colour as the concentration of oxygen increases, yielding sulfur dioxide. A small percentage of the trioxide is formed at the same time. To make commercial quantities of the trioxide, the dioxide must be oxidized in the presence of a catalyst at elevated temperatures. A major problem in manufacturing sulfur dioxide to be used for further chemical combination is disposal of the large amount of heat evolved in the reaction.

A major source of sulfur dioxide for production of sulfuric acid is the roasting of sulfides of such metals as iron, copper, zinc, nickel, cobalt, and lead. The oxidation of pyrite (FeS_2) is illustrative:

$$4FeS_2 + 11O_2 \longrightarrow 2Fe_2O_3 + 8SO_2.$$

pyrite; oxygen; iron(III) oxide; sulfur dioxide

The pollution of the atmosphere by gases resulting from the combustion of sour (high sulfur content) fossil fuels is a very serious environmental problem.

Sulfur dioxide is easily produced in the laboratory by the reaction of acids with sulfites or the reduction of sulfuric acid with metals. The following reactions are illustrative.

$$2NaHSO_3 + H_2SO_4 \xrightarrow{\text{heat}} Na_2SO_4 + 2H_2O + 2SO_2$$

sodium hydrogen sulfite; sulfuric acid; sodium sulfate; water; sulfur dioxide

$$2H_2SO_4 + Cu \longrightarrow SO_2 + CuSO_4 + 2H_2O$$

sulfuric acid; copper; sulfur dioxide; copper(II) sulfate; water

Uses of the dioxides

In addition to its production for conversion to sulfuric acid, sulfur dioxide is produced as a bleaching agent, as an industrial reducing agent (as for chlorine elimination), for food preservation, and fruit ripening. Sulfur dioxide can be liquefied easily; the liquid has been employed as a solvent in petroleum refining and in organic laboratory operations.

Selenium dioxide, SeO_2, an important reagent in organic chemistry, may be formed by direct union of the elements but is more conveniently prepared by treatment of the red form of the element (as an impure sludge, for instance) with nitric acid, HNO_3. The selenium is first converted to selenious acid, H_2SeO_3, which can be dehydrated to SeO_2. Under similar conditions, tellurium forms a compound formulated $2TeO_3 \cdot HNO_3$, which on heating and neutralization gives tellurium dioxide, TeO_2. Metaperiodic acid, HIO_4, behaves like nitric acid, giving $2TeO_3 \cdot HIO_4$ and then TeO_2 on heating.

Trioxides. Sulfur trioxide is formed only in very small amounts on direct combination of the elements ($S + 3/2O_2 \rightarrow SO_3$). Among the most important catalysts for the formation of SO_3 is vanadium pentoxide, V_2O_5, which is relatively insensitive to such poisons as arsenic oxides and hydrocarbons. Though now of little interest commercially, oxygen and iron(II) ions will convert sulfurous acid or sulfites to sulfuric acid (or sulfates). Oxides of nitrogen such as dinitrogen trioxide, N_2O_3, or nitrogen dioxide, NO_2, have been used for centuries as catalysts.

The so-called chamber process for making sulfuric acid dates from 1746 and uses a mixture of nitric oxide, NO, and NO_2 to accomplish the oxidation in a reaction generally represented by:

$$SO_2 + N_2O_3 + nH_2O \longrightarrow H_2SO_4 + (n-1)H_2O + 2NO.$$

sulfur dioxide; dinitrogen trioxide; water; sulfuric acid; water; nitric oxide

The nitric oxide (NO) is reconverted to N_2O_3 by air oxidation:

$$2NO + \tfrac{1}{2}O_2 \longrightarrow N_2O_3.$$

nitric oxide; oxygen; dinitrogen trioxide

As mentioned in the section describing general properties, the heavier chalcogens resist oxidation to the highest state (VI), but once this state is formed it reverts readily to the lower (IV) state or to the elemental (0) state. Thus, Te(VI) and Se(VI) compounds are strong oxidizing agents. Se(VI) oxide, SeO_3, is difficult to form in the pure state but may be formed in a glow discharge with oxygen and SeO_2. Although the acid H_2SeO_4 is preparable, dehydration to SeO_3, its anhydride, does not occur. Conversely, orthotelluric acid, H_6TeO_6, does exist and is dehydrated to TeO_3 at temperatures above 300° C.

Sulfur trioxide exists in three allotropic forms. Liquid sulfur trioxide can be stabilized by adding a small proportion of boric acid to prevent the asbestos-like or icelike modifications from forming.

Reactions with the halogens. The chalcogens react directly with all the halogens to form a wide variety of compounds. The fluorides are the most abundant and in general the most stable. Table 3 summarizes some of the data on these compounds.

Table 3: The Chalcogen Halides

compound	preparation and some properties
S_2F_2, disulfur difluoride	from sulfur and AgF, mp = −120.5° C, bp = −38.4° C
SF_4, sulfur tetrafluoride	direct combination of SCl_2 and NaF in CH_3CN; very reactive, used in synthesis of fluorides; mp = −124° C, bp = −40° C
SF_6, sulfur hexafluoride	kinetically inert to alkalies and oxygen but reacts with sodium; mp = −50.5° C and bp = 63.8° C
S_2F_{10}, disulfur decafluoride	from SF_5Cl and H_2 or direct combination; reactive and poisonous; mp = −92° C, bp = 29° C
SF_5Cl, sulfur pentafluoride chloride	from SF_4, Cl_2 and CsF; a partially ionic compound, $SF_5^-Cl^+$
SeF_4, selenium tetrafluoride	direct combination or from $SeCl_4$ and AgF; very reactive and thermally unstable
TeF_4, tellurium tetrafluoride	direct combination or from TeF_6 and Te, very reactive and thermally unstable; mp = 130° C
$TeCl_4$, tellurium tetrachloride	direct combination; mp = 224° C, bp = 380° C
TeI_4, tellurium tetraiodide	direct combination in sealed tube; decomposes 100° C
Te_2F_{10}, ditellurium decafluoride	direct combination; mp = −34° C, bp = 53° C
SeF_6 and TeF_6, selenium and tellurium hexafluorides	direct combination; stable but more reactive than SF_6; bp and mp of SeF_6, −34.5° C and −39° C respectively; bp and mp TeF_6, 35.5° C and −36° C respectively
S_2Cl_2, disulfur dichloride	direct combination in excess of S; good solvent for S_8; hydrolyzes; orange-yellow; mp = −80° C, bp = 135.6° C
S_2Br_2, disulfur dibromide	direct combination in excess S; deep red; mp = −40° C; only S–Br compound well known
SCl_2, sulfur dichloride	direct combination in excess Cl_2 or S_2Cl_2 with Cl_2; red liquid; bp = 59° C
SCl_4, sulfur tetrachloride	from SCl_2 and Cl_2; stable only as solid, probably $SCl_3^+Cl^-$; pale yellow; mp = −30° C
Se_2X_2, diselenium dihalides	all from Se and SeX_4, quite unstable
$SeCl_2$, selenium dichloride	from decomposition of $SeCl_4$
$SeBr_2$, selenium dibromide	from Se_2Br_2 decomposition
$TeCl_2$, tellurium dichloride	direct combination or reaction of Te and $TeCl_4$; very reactive, in water forms Te and H_2TeO_3; mp = 209° C, bp = 327° C
$SeBr_4$ and $TeBr_4$, selenium and tellurium tetrabromides	direct combination; thermally unstable; form ionic salts such as $K^+[SeBr_6^-]$; decompose to Br_2 and $SeBr_2$ or $TeBr_2$; relative hydrolysis rates: $SX_4 < SeX_4 < TeX_4$

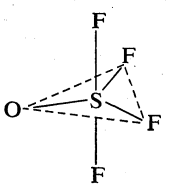
Polonium forms a volatile fluoride; a red dichloride, $PoCl_2$; a red tetrabromide, $PoBr_4$; and a black tetraiodide, PoI_4. The general chemistry of polonium resembles that of tellurium and bismuth.

Chalcogen oxyhalides. There are a large number of compounds in which a chalcogen atom is bonded to both oxygen and halogen atoms. In naming these compounds, the term thionyl is used to designate the unit $>SO$ and sulfuryl to designate the unit $>SO_2$; examples are thionyl chloride, $SOCl_2$, and sulfuryl bromide, SO_2Br_2. Table 4 summarizes some of the properties of the best known oxyhalides.

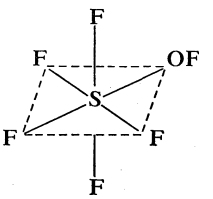
Table 4: Oxyhalides of the Chalcogens

compound	preparation and properties
Thionyl halides or chalcogen(IV) oxyhalides	
SOF_2, thionyl fluoride	from $SOCl_2$ and SbF_3; more stable than $SOCl_2$; hydrolyzes slowly; mp = −110.5° C, bp = −43.8° C
$SOCl_2$, thionyl chloride	from PCl_5 and SO_2; hydrolyzes readily to H_2SO_3; bp = 78.8° C
SOFCl, thionyl chloride fluoride	from $SOCl_2$ and SbF_3
$SOBr_2$, thionyl bromide	from $SOCl_2$ and HBr; red-yellow liquid, unstable when heated
$SeOF_2$, selenyl fluoride	from $SeOCl_2$ and AgF; very reactive, good fluorinating agent
$SeOCl_2$, selenyl chloride	distillation of $SeO_2 \cdot 2HCl$ in presence of P_4O_{10}
$SeOBr_2$, selenyl bromide	action of bromine on Se or SeO_2
Sulfuryl or chalcogen(VI) oxyhalides	
SO_2F_2, sulfuryl fluoride	combination of SO_2 and fluorine or by thermal decomposition of $Ba(SO_3F)_2$; mp = −136.7° C, bp = −55.4° C; thermally stable and chemically inert
SO_2Cl_2, sulfuryl chloride	direct union of Cl_2 and SO_2 in presence of camphor; more reactive than SO_2F_2; bp = 69.1° C
SO_2ClF, sulfuryl chloride fluoride	from SO_2Cl_2 and SbF_3
SeO_2Cl_2, selenium(VI) oxychloride	from $BaSeO_4$ and HSO_3Cl

A number of other oxyhalides are known, not in either of the thionyl or sulfuryl general classes. Two fluorides (formulated SOF_4 and SOF_6) have the structures shown below:



SOF_4



SOF_6

Both compounds result from the fluorination of thionyl fluoride (SOF_2); the presence of cesium fluoride favours formation of SOF_6. The ion SF_5O^- exists as a salt $SF_5O^-Cs^+$; by analogy SF_5OF might be one of the few compounds with a partially positive fluorine atom.

Other complex oxy-species are the highly poisonous $S_2O_5F_2$ (which melts at −48° C and boils at 51° C), $S_2O_5Cl_2$, and the mixed halide, S_2O_5ClF. The fluoride and chloride form from the reaction of sulfur trioxide and the appropriate halogen. The mixed chloride fluoride is formed from $S_2O_5Cl_2$ and antimony trifluoride, SbF_3. They all have an S—O—S structural unit. A trisulfoxy compound, $S_3O_8F_2$, is formed from the action of boron trifluoride (BF_3) and sulfur trioxide. Two peroxo halides are reported: $S_2O_6F_2$ and SO_3F.

Nitrogen compounds of the chalcogens. The molecular structures of sulfur—and to some extent those of selenium and tellurium—in the eight-membered ring systems have a number of analogues in which the sulfur atoms are replaced by one or more nitrogen atoms. These compounds are thought of as nitrides rather than sulfides (or selenides) because, on hydrolysis by acids, ammonium salts are formed rather than hydrogen sulfide.

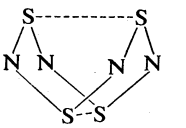
Cyclotetrasulfur tetranitride. The compound cyclotetrasulfur tetranitride, S_4N_4, is preparable using the following procedures:

1. The action of sulfur dichloride or disulfur dichloride on ammonia in solvents such as carbon disulfide, benzene, chloroform, or dimethylformamide. The chloride S_4N_3Cl, a canary-yellow solid, is also formed.
2. The direct action of dry ammonia on sulfur with removal of the by-product, hydrogen sulfide.

$$10S + 4NH_3 \longrightarrow 6H_2S + S_4N_4$$

sulfur — ammonia — hydrogen sulfide — cyclotetrasulfur tetranitride

The structure of the molecule is best described by a plane of nitrogen atoms with a pair of sulfur atoms above and another pair below this plane:

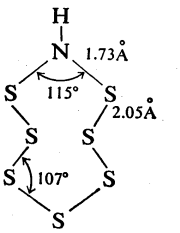


All the S—N distances are equal, with lengths between those expected for single and double bonds. The short S—S distance of 2.58 angstroms suggests that there are weak bonds between these pairs of atoms. If such is the case, the sulfur atoms may be tetravalent and the nitrogen atoms trivalent.

Cyclotetrasulfur tetranitride forms orange crystals, insoluble in water but soluble in benzene, carbon disulfide, or liquid ammonia. Its melting point is 187° C. It is a convenient starting material for other sulfur–nitrogen compounds. It is said to facilitate the starting action of diesel fuel and to serve as a repellent for birds when placed on grain before planting.

Treatment of the compound with tin and hydrochloric acid reduces the nitrogen atoms to imido groups ($>NH$), forming the compound cyclotetrasulfur tetraimide, $S_4N_4H_4$.

Heptasulfurimide. Heptasulfurimide (S_7NH) usually is formed along with S_4N_4 in reaction of disulfur dichloride (S_2Cl_2) and ammonia. If the reaction temperature is lowered to −10° C and dimethylformamide is used as solvent, the yield of S_7NH increases. When highly purified, the material is nearly pure white and has a melting point of 113.5° C, very near that of orthorhombic sulfur. The molecule contains an eight-membered ring closely resembling in its dimensions the S_8 ring with an imide group ($>NH$) replacing one sulfur atom:



cycloheptasulfurimide

The hydrogen atom is reactive and salts may be prepared. Alkalies convert S_7NH to intensely blue species thought to contain free radicals and the ion S_7N^-. Variations of these ring compounds are possible, with smaller rings being formed as well as those with more than one NH group present in either a six- or eight-membered ring.

The only analogous compounds of nitrogen with selenium or tellurium are cyclotetraselenium tetranitride, Se_4N_4, and cyclotetratellurium tetranitride, Te_4N_4. They are formed from ammonia and the elemental chalcogens or their so-called monohalides. They are unstable and easily revert to the free chalcogen.

The chalcogen oxyacids. The number of oxychalcogen acids, especially those of sulfur, is large, as is their importance to the chemical industry, especially sulfuric acid, H_2SO_4. Of the some 16 oxyacids of sulfur (including only the most common polythionic acids) reported, no

Table 5: Oxyacids of the Chalcogens

	apparent oxidation state	apparent anhydride	name		related compounds
			acid	salt	
H_2SO_2	II	SO	sulfoxylic	sulfoxylate	SCl_2, $S(CN)_2$
H_2SO_3	IV	SO_2	sulfurous	sulfite	$SOCl_2$, SF_4
H_2SeO_3	IV	SeO_2	selenious	selenite	$SeOCl_2$, SeF_4
H_2TeO_3	IV	TeO_2	tellurous	tellurite	$TeOCl_2$, $TeCl_4$
H_2SO_4*	VI	SO_3	sulfuric	sulfate	SO_2Cl_2, SF_6
H_2SeO_4	VI	SeO_3	selenic	selenate	SeO_2Cl_2, SeF_6
H_6TeO_6*	VI	TeO_3	telluric (ortho)	tellurate	$K_2^+ [TeO_2(OH)_4]^{2-}$, $TeCl_6$
H_2SO_5*	VI	SO_4 and S_2O_7	peroxymonosulfuric (Caro's)	peroxymonosulfate	—
$H_2S_2O_2$	0 and II	S_2O	thiosulfurous	thiosulfite	S_2Cl_2
$H_2S_2O_3$	0 and IV	—	thiosulfuric	thiosulfate	—
$H_2S_nO_3$†	II and IV	—	sulfanesulfonic	sulfanesulfonate	—
$H_2S_2O_4$	III	S_2O_3	dithionous, hydrosulfurous, or hyposulfurous	dithionite, hydrosulfite, or hyposulfite	—
$H_2S_2O_5$	IV	SO_2	pyrosulfurous	pyrosulfite	$SOCl_2$, HOSOF
$H_2S_2O_6$	V	S_2O_5	dithionic	dithionate	—
$H_2S_nO_6$‡	II and IV	—	trithionic	trithionate	—
$H_2S_2O_7$*	VI	SO_3	pyrosulfuric	pyrosulfate	$S_2O_5F_2$
$H_2S_2O_8$*	VI	SO_4	peroxydisulfuric	peroxydisulfate	$S_2O_6F_2$

*Pure compounds. †$n \geq 2$. ‡$n \geq 3$.

more than four or five have been prepared in the pure state. The others are known only in dilute aqueous acid form or as salts or esters. Table 5 includes the most important of the oxyacids, including those of selenium and tellurium. No completely satisfactory system of categorization exists. In this table they are arbitrarily listed in order of increasing apparent oxidation number of the chalcogen followed by increasing chalcogen complexity. In the polychalcogen acids all the chalcogen atoms need not be, and in many cases certainly are not, in the same oxidation state. There are no examples of polyselenium or polytellurium acids, in conformance with the decreasing tendency of catenation in the Group VIa family with increasing atomic number. Examples of polythio acids are numerous, as expected from the strong catenation tendency of sulfur.

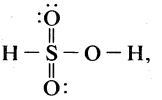
Sulfoxylic acid. This acid does not exist in the free state, nor can it be prepared from its hypothetical anhydride, SO. Its zinc salt results from the reaction of sulfuryl chloride and zinc:

$$SO_2Cl_2 + 2Zn \longrightarrow ZnSO_2 + ZnCl_2$$

sulfuryl chloride; zinc; zinc sulfoxylate; zinc chloride

Thiosulfurous acid. The prefix "thio" in the name of this compound implies that an oxygen atom in sulfurous acid, H_2SO_3, has been replaced by a sulfur atom. Only esters of this acid (compounds in which the hydrogen atoms have been replaced by atomic groups derived from organic substances) are known, having structures $(RO)_2S_2$. Hydrolysis of such esters gives a product that behaves as a very weak acid, as expected of a molecule in which a sulfur atom has a zero formal charge and hydrogen atoms are attached to the oxygen atoms: H—O—S—S—O—H.

Sulfurous acid, selenious acid, and tellurous acid. When sulfur dioxide is added to water, sulfurous acid, H_2SO_3, is but one of several species present. Much of the sulfur dioxide remains in the molecular form with water molecules attached essentially by dipole–dipole interaction. Some of the acid is in a form



in which one of the hydrogen atoms is bonded to the sulfur. In addition to $(HO)_2S{=}O$, there are also ionic species from partially and completely dissociated H_2SO_3:

$$H_2SO_3 \rightleftharpoons H^+ + HSO_3^-$$

sulfurous acid; hydrogen ion; hydrogen sulfite ion

$$HSO_3^- \rightleftharpoons H^+ + SO_3^{2-}$$

hydrogen sulfite ion; hydrogen ion; sulfite ion

Sulfurous acid is considered to be a moderately weak acid. The free acid has never been prepared. On concentration of the aqueous solution, sulfur dioxide is liberated.

Sulfites and bisulfites

Both normal and acid salts of sulfurous acid are preparable either by neutralization of the aqueous solution of the acid and the normal salts or by direct action of sulfur dioxide on metal oxides. The formation of potassium salts is representative.

$$H_2SO_3 + KOH \rightleftharpoons KHSO_3 + H_2O$$

sulfurous acid; potassium hydroxide; potassium hydrogen sulfite; water

$$KHSO_3 + KOH \rightleftharpoons K_2SO_3 + H_2O$$

potassium hydrogen sulfite; potassium hydroxide; potassium sulfite; water

$$K_2O + SO_2 \rightleftharpoons K_2SO_3$$

potassium oxide; sulfur dioxide; potassium sulfite

Acidification of sulfites, especially the acid sulfites (*e.g.*, $KHSO_3$), liberates sulfur dioxide after first forming sulfurous acid:

$$KHSO_3 + HCl \rightleftharpoons H_2SO_3 + KCl$$

potassium hydrogen sulfite; hydrochloric acid; sulfurous acid; potassium chloride

$$H_2SO_3 \rightleftharpoons SO_2 + H_2O.$$

sulfurous acid; sulfur dioxide; water

Thermal treatment of alkali metal acid sulfites forms pyrosulfites:

$$2KHSO_3 \xrightarrow{\text{heat}} K_2S_2O_5 + H_2O.$$

Metal sulfites are used in the paper and pulp industries in the formation of a fibre free of lignin. The disposal of the sulfite wastes presents a major environmental problem for the paper industry.

Selenium dioxide, SeO_2, reacts with water to form selenious acid, H_2SeO_3, a white solid. There does not appear to be a pure tellurous acid, H_2TeO_3. Selenious and

tellurous acids are both somewhat weaker than sulfurous acid.

Sulfites are easily converted to sulfates, thus sulfurous acid is a better reducing agent than oxidizing agent. Only with strong reducing agents (*e.g.*, hydrogen iodide) does it serve in the latter capacity:

$$6HI + H_2SO_3 \longrightarrow H_2S + 3I_2 + 3H_2O$$

hydrogen iodide; sulfurous acid; hydrogen sulfide; iodine; water

On the other hand, H_2SeO_3 and H_2TeO_3 in solution are more likely to function as oxidizing than reducing agents, themselves being reduced to the free chalcogen as illustrated by H_2SeO_3:

$$H_2O + H_2SeO_3 + 2SO_2 \longrightarrow 2H_2SO_4 + Se$$

water; selenious acid; sulfur dioxide; sulfuric acid; selenium

Sulfuric acid, selenic acid, and telluric acid. Sulfuric acid is the most abundantly produced chemical in the world. It is made industrially from its anhydride, sulfur trioxide, SO_3.

The process of dissolving sulfur trioxide in water to form sulfuric acid is not a trivial one. The reaction is very exothermic, with one water molecule interacting strongly with SO_3, to form a monohydrate $O_3S \cdot OH_2$, which is relatively inert, resisting solution. Introduction of gaseous SO_3 into a moist atmosphere produces a dense fog composed of this material. Dihydrates and tetrahydrates are also stable. Sulfur trioxide dissolves readily in concentrated sulfuric acid to form an acid represented as $H_2SO_4 \cdot xSO_3$. When $x = 1$ the compound is called pyrosulfuric acid, $H_2S_2O_7$. A product containing 30 percent or more SO_3 is marketed as oleum. Dilution of concentrated sulfuric acid may generate enough heat to cause vigorous boiling of the water, resulting in the expulsion of steam and spattering. Any dilution process of a strong acid is carried out by adding acid to water (rather than the reverse) with rapid stirring.

Sulfuric acid dissociates in aqueous solution in two steps:

$$H_2SO_4 \rightleftharpoons H^+ + HSO_4^- \text{ and } HSO_4^- \rightleftharpoons H^+ + SO_4^{2-}.$$

Acid and normal salts as well as the pyrosalts (see above) are easily formed:

$$H_2SO_4 + KOH \longrightarrow KHSO_4 + H_2O$$

sulfuric acid; potassium hydroxide; potassium hydrogen sulfate; water

$$KHSO_4 + KOH \longrightarrow K_2SO_4 + H_2O$$

potassium hydrogen sulfate; potassium hydroxide; potassium sulfate; water

$$2KHSO_4 \xrightarrow{\text{heat}} K_2S_2O_7 + H_2O$$

potassium hydrogen sulfate; potassium pyrosulfate; water

The acid and pyrosalts have use in preparing intractable metal oxide ores for solution metallurgy.

Production of sulfuric acid

Two processes have been responsible for the major output of commercial sulfuric acid over the past century and a half: (1) the chamber process and (2) the contact process. The two processes are discussed only briefly here.

In the mid-18th century the importance of oxidizing sulfur dioxide catalytically to sulfur trioxide by oxides of nitrogen or by nitric acid (and thence the formation of sulfuric acid) was recognized. All of the reactions in the oxidation of SO_2 to H_2SO_4 are not, even today, fully understood, but the formation of nitrosylsulfuric acid (lead chamber crystals) is of primary importance in the chamber process. The overall process may be described by the following reactions:

$$SO_2 + H_2O \rightleftharpoons H_2SO_3$$

sulfur dioxide; water; sulfurous acid

$$H_2SO_3 + NO_2 \longrightarrow H_2SNO_5$$

nitrogen dioxide; nitrosylsulfuric acid

$$H_2SNO_5 \longrightarrow H_2SO_4 + NO$$

sulfuric acid; nitric oxide

$$NO + \tfrac{1}{2}O_2 \longrightarrow NO_2$$

oxygen

The use of platinum metal catalysts to convert sulfur dioxide to sulfur trioxide was known as early as 1831. This metal is sensitive to such materials as arsenic and hydrocarbons, which poison the catalytic action. Vanadium pentoxide, V_2O_5, is less sensitive than platinum metals to poisons and is presently the most commonly used catalyst in industrial production of sulfuric acid. The essential reactions are much simpler than those noted earlier for the chamber process.

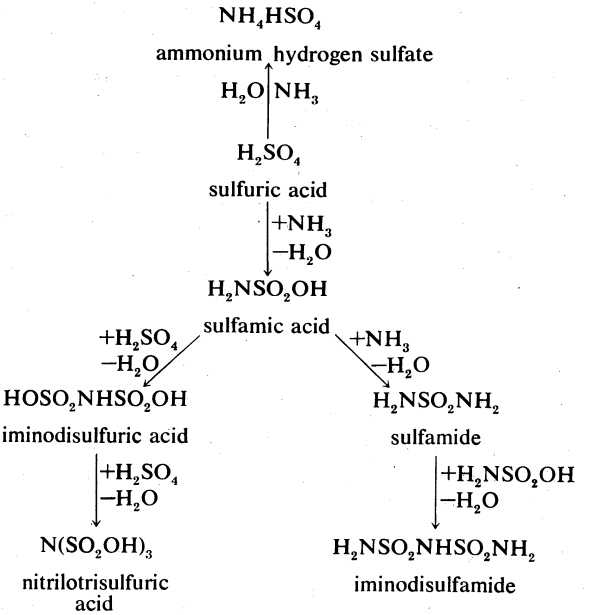
$$S + O_2 \longrightarrow SO_2 \qquad (1)$$

sulfur; oxygen; sulfur dioxide

$$SO_2 + \tfrac{1}{2}O_2 \xrightleftharpoons[\text{(vanadium pentoxide catalyst)}]{V_2O_5} SO_3 \qquad (2)$$

sulfur trioxide

Optimum conditions of temperature, contact time of the gases on the catalyst surface, ratio of the reacting and product gases (SO_2, SO_3, O_2 and N_2 from air), pressure as well as purity (freedom from poisons) must be experimentally determined. Too high a temperature (above about 425° C) is unfavourable for the formation of SO_3, since reaction (2) above is reversed. Too low a temperature (below about 400° C) does not allow reactions (1) and (2) to proceed rapidly enough even in the presence of the catalyst. Multiple chambers for catalytic oxidation may be used for a more efficient conversion of SO_2 to SO_3. The sulfur trioxide is absorbed in concentrated sulfuric acid. In addition to its strongly acidic nature, sulfuric acid is an oxidizing agent, especially when concen-



Some compounds derived from ammonia and sulfuric acid.

trated and warmed. Iodide and bromide ions are oxidized to the elemental halogens while reducing sulfuric acid to hydrogen sulfide, H_2S. This property prevents H_2SO_4 from being used in generating hydriodic or hydrobromic acids, HI and HBr, from their salts, NaI and NaBr. Certain metals (such as copper and silver) will reduce sulfuric acid to sulfur dioxide. The pure acid, 100 percent H_2SO_4, serves as a solvent for some organic reactions.

A number of compounds derived from both ammonia and sulfuric acid are known (see scheme, above).

Sulfamic acid

One of the most important of these compounds in sulfamic acid, NH_2SO_2OH, a versatile chemical produced in very large quantities. It is more conveniently and economically produced by the interaction of sulfur trioxide (or oleum) with urea, NH_2CONH_2. It is a white, solid, nonhygroscopic, strong acid useful as a primary standard for bases. It is also a reagent for the analysis of the nitrite ion, NO_2^- because it reacts rapidly and quantitatively with it to give nitrogen gas:

$$\underset{\text{sulfamic acid}}{H_2NSO_2OH} + \underset{\text{nitrite ion}}{NO_2^-} \longrightarrow \underset{\text{nitrogen}}{N_2} + \underset{\text{hydrogen sulfate ion}}{HSO_4^-} + \underset{\text{water}}{H_2O}.$$

Salts of sulfamic acid are used as herbicides and as flame retardants. The dairy industry uses large amounts of the acid in removing "milkstone" formed in pasteurization and evaporation of milk.

Selenic acid, H_2SeO_4, is formed as an oxidation product of selenium dioxide with halogens such as chlorine and bromine. Heavy metal selenites such as silver selenite, Ag_2SeO_3, are good sources of selenic acid because upon oxidation by a halogen the resulting halide ion is precipitated leaving a relatively pure form of the acid:

$$\underset{\text{silver selenite}}{Ag_2SeO_3} + \underset{\text{bromine}}{Br_2} + \underset{\text{water}}{H_2O} \longrightarrow \underset{\text{selenic acid}}{H_2SeO_4} + \underset{\text{silver bromide}}{2AgBr}$$

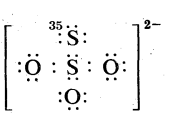
The anhydrous acid can be isolated but is less stable than sulfuric acid at elevated temperatures. Oxygen is evolved from selenic acid at about 210° C. The acid strength of H_2SeO_4 is comparable to that of sulfuric acid.

There are marked differences between tellurium and its two predecessors in Group VIa in the formation of oxyacids. The large size of the atom, even in a high oxidation state (VI), and its resistance to forming multiple bonds favour a highly hydroxylated acid of the form $(OH)_6Te$. When strong oxidizing agents, such as hydrogen peroxide or chromic acid, CrO_3, in nitric acid react with tellurium dioxide this hexahydroxytelluric(VI) acid is formed. Although it is a very weak acid, in a strongly basic environment two protons can be removed, forming the ions $[OTe(OH)_5]^-$ and $[O_2Te(OH)_4]^{2-}$. Telluric acid is an oxidizing agent, but its reactions usually are slow.

Thiosulfuric acid. This acid is known as its salts and in aqueous solution but not in the pure state. Sulfur dissolved in alkali sulfites forms the thiosulfate ion:

$$S_8 + 8Na_2SO_3 \longrightarrow 8Na_2S_2O_3.$$

The structure of the thiosulfate ion suggests a replacement of an oxygen atom in the sulfate ion by a sulfur atom. When radioactive sulfur (^{35}S) is used in the above preparation, it can be proved that the added sulfur has no oxygen attached to it:



since acidification of the thiosulfuric acid regenerates the radioactive sulfur:

$$\underset{\substack{\text{thiosulfate ion made from sulfur-35 and sulfite}}}{S_2O_3^{2-}} + \underset{\text{hydrogen ion}}{2H^+} \longrightarrow \underset{\text{sulfur-35}}{^{35}S} + \underset{\text{sulfurous acid}}{H_2SO_3}.$$

A mechanism has been postulated for the formation of thiosulfurous acid by attack of sulfurous acid (or sulfites) on the S_8 ring:

$$\underset{\text{sulfur}}{S_8} + \underset{\text{sulfurous acid}}{H_2SO_3} \longrightarrow \underset{\text{a sulfanesulfonic acid}}{HOSO_2S{-}S_6{-}SH}$$

$$HOSO_2S{-}S_6{-}SH + H_2SO_3 \longrightarrow HOSO_2S{-}S_5{-}SH + H_2S_2O_3$$

$$HOSO_2S{-}S_5{-}SH + H_2SO_3 \longrightarrow HOSO_2S{-}S_4{-}SH + H_2S_2O_3$$

(etc.)

Eventually, eight molecules of thiosulfurous acid form.

Alkali metal thiosulfates are important industrially as solubilizing agents for silver halides in the photographic industry. The portion of silver bromide, AgBr, unaffected by light on exposure of the emulsion in the camera is dissolved by the thiosulfate ion:

$$AgBr + S_2O_3^{2-} \longrightarrow AgS_2O_3^- + Br^-.$$

This is called the fixing process and the sodium salt, $Na_2S_2O_3$, is called hypo from an improper but often used name, sodium "hyposulfate."

Dithionous, hyposulfurous, or hydrosulfurous acid. All these names are or have been used to describe the acid that results from the reduction of sulfur dioxide by such agents as sodium amalgam (a solution of sodium in mercury) or zinc metal. The free acid is unknown, but the salts are stable enough to be used as reducing agents. A unique feature of the $S_2O_4^{2-}$ ion is the long S—S bond of 2.39 Å, which suggests a ready dissociation of the ion to an active radical:

$$\underset{\text{dithionite ion}}{S_2O_4^{2-}} \rightleftharpoons \underset{\text{a radical anion}}{2 \cdot SO_2^-}.$$

Dithionic acid. Pure dithionic acid, $H_2S_2O_6$, is too unstable to be isolated. When sulfur dioxide reacts with MnO_2, manganese dioxide, the manganous salt of dithionic acid, MnS_2O_6, is formed. Treatment with barium carbonate or barium hydroxide forms the barium salt, which in turn can be converted to sodium dithionate, $Na_2S_2O_6$. Solutions of the acid $H_2S_2O_6$ are strong proton providers. The geometry shows a tetrahedral orientation of oxygen atoms about each sulfur atom and a normal S—S distance of about 2.0 Å. There are no analogous compounds of tellurium or selenium.

The polythionic acids. The polythionic acids $H_2S_nO_6$ ($n = 3$ to 6) are unique in that they contain S—S—S units in which there is at least one sulfur atom between two tetrahedra of $-SO_3H$. Polythionic acids may be thought of as sulfanedisulfonic acids. No acid salts of the polythionates are known (*e.g.*, KHS_4O_6 does not exist), and the acids are unstable.

A general method of preparation of these acids is the action of sulfites, sulfurous acid, or thiosulfates on sulfur chlorides or iodine, as illustrated in the following equations:

$$\underset{\substack{\text{potassium hydrogen sulfite}}}{2KHSO_3} + \underset{\substack{\text{sulfur dichloride}}}{SCl_2} \longrightarrow \underset{\text{hydrochloric acid}}{2HCl} + \underset{\text{potassium trithionate}}{K_2S_3O_6}$$

$$\underset{\text{sulfurous acid}}{2H_2SO_3} + \underset{\text{disulfur dichloride}}{S_2Cl_2} \longrightarrow \underset{\text{hydrochloric acid}}{2HCl} + \underset{\text{tetrathionic acid}}{H_2S_4O_6}$$

$$\underset{\text{thiosulfate ion}}{S_2O_3^{2-}} + \underset{\text{iodine}}{I_2} \longrightarrow \underset{\text{iodide ion}}{2I^-} + \underset{\text{tetrathionate ion}}{S_4O_6^{2-}}$$

$$\underset{\text{thiosulfate ion}}{2S_2O_3^{2-}} + \underset{\text{sulfur dichloride}}{SCl_2} \longrightarrow \underset{\text{chloride ion}}{2Cl^-} + \underset{\text{pentathionate ion}}{S_5O_6^{2-}}$$

$$2S_2O_3^{2-} + S_2Cl_2 \longrightarrow 2Cl^- + S_6O_6^{2-}$$

thiosulfate ion; disulfur dichloride; chloride ion; hexathionate ion

A solution of sulfur dioxide saturated with hydrogen sulfide (Wackenroder's solution) contains a mixture of polythionic acids, especially $H_2S_4O_6$ and $H_2S_6O_6$.

Sulfur–nitrogen oxides. A few S—N—O compounds are known, with chain or zigzag structures. Ammonia and thionyl chloride, $SOCl_2$, react to give one of these compounds, $S_3N_2O_2$, a bright yellow substance also formed from ammonium chloride and either thionyl chloride or disulfur dichloride, S_2Cl_2. Sulfur trioxide reacts with S_4N_4 to give this same oxide and also another with the formula $S_3N_2O_5$. $S_3N_2O_2$ is stable in the absence of moisture and soluble in benzene. Two of the sulfur atoms in this compound appear to be in the S(VI) state and one in the S(IV) state.

Analysis of polythionates may be accomplished by the action of cyanide ion, CN^-, on the —S—S—S— chain, giving the thiocyanate ion, SCN^-, in a quantity of three moles less than the total S-atom number in the polythionate compound. Thus, in the analysis of $S_5O_6^{2-}$ two moles of SCN^-, one mole of $S_2O_3^{2-}$, and one mole of SO_4^{2-} are formed. The thiocyanate ion and the thiosulfate ion are easily determined by standard procedures.

Total sulfur in any sulfur-containing species is usually determined by oxidation to sulfate ion, precipitation as barium sulfate, and weighing. Peroxides and perchloric acid are used in the oxidation as well as for the destruction of any organic material present.

The thiosulfate ion is readily estimated by titration with a standard solution of iodine.

BIBLIOGRAPHY. Among texts written at the general chemistry and secondary level covering the Group VIa elements, C.W. KEENAN and J.H. WOOD, *General College Chemistry*, 4th ed. (1971); and J.A. CAMPBELL *et al.*, *Chemical Systems* (1970), are representative of sources of basic information; at a higher level are W.L. JOLLY, *The Synthesis and Characterization of Inorganic Compounds* (1970); R.T. SANDERSON, *Inorganic Chemistry* (1967); HEINRICH REMY, *Lehrbuch der anorganischen Chemie*, 4th ed. (1943; Eng. trans., *Treatise on Inorganic Chemistry*, vol. 1, 1956); C.S.G. PHILLIPS and R.J.P. WILLIAMS, *Inorganic Chemistry* (1965); JACOB KLEINBERG, W.J. ARGERSINGER, JR., and ERNEST GRISWOLD, *Inorganic Chemistry* (1960); B.E. DOUGLAS and D.H. MCDANIEL, *Concepts and Models of Inorganic Chemistry* (1965); and E.S. GOULD, *Inorganic Reactions and Structure*, rev. ed. (1962). The biosphere and ecosystems related to sulfur and selenium are discussed in A.B. ROY and P.A. TRUDINGER, *The Biochemistry of Inorganic Compounds of Sulphur* (1970); and in W.W. KELLOGG, "The Sulphur Cycle," *Science*, 175:587–596 (1972); allotropy and structure are presented in L.B. GITTINGER, "Sulphur, 1970," *Engng. Min. J.*, 172:133–136 (1971); K.W. BAGNALL, *The Chemistry of Selenium, Tellurium, and Polonium* (1966); D.M. CHIZHIKOV and V.P. SHCHASTLIVYI, *Selenium and Selenides* (1968; orig. pub. in Russian, 1964); F. TUINSTRA, *Structural Aspects of the Allotropy of Sulfur and the Other Divalent Elements* (1967); A.V. TOBOLSKY and W.J. MACKNIGHT, *Polymeric Sulfur and Related Polymers* (1965); and BEAT MYER (ed.), *Elemental Sulfur* (1965). Compilations of physical and numerical data include W.M. LATIMER, *The Oxidation States of the Elements and Their Potentials in Aqueous Solutions*, 2nd ed. (1952); ROBERT C. BRASTED, *Comprehensive Inorganic Chemistry*, vol. 8 (1961); F.A. COTTON and GEOFFREY WILKINSON, *Advanced Inorganic Chemistry*, 2nd rev. ed. (1966); and L.H. AHRENS, *Distribution of the Elements in Our Planet* (1965).

(R.C.Br.)

Pacific Coast Ranges

The Pacific Coast mountain ranges run parallel to the coastline of the states of California, Washington, and Oregon, and they extend into Canada and Mexico. In their fastness stand the giant redwood forests and the magnificent coastal scenery of California and Oregon. Also lying within the ranges are several sprawling urban regions, the largest of which are those centred on San Francisco and on the megalopolis of Los Angeles. The effort to preserve the beauty of the coastline, together with the ever-present threat of earthquakes and forest fires and the growth of the region's cities, has brought the ranges into national focus in the 1970s, because of the continuing debate over the nature and quality of the United States environment.

History of scientific study. The first generation of students to make thorough investigations into the complex geology of the Coast Ranges came from local universities in the late 19th century. Their studies were continued by an active second generation working for state and federal geological surveys and for petroleum-exploration companies. The great San Francisco earthquake of 1906 precipitated a thorough investigation of the earthquake problem, especially of epicentres (focuses of seismic activity) in the Californian Coast Ranges; once it was recognized that earthquakes start from movements along faults, or fractures, the fault system of the ranges was mapped with care, and an intensive seismological program developed in major Pacific coast universities. With the great influx of population into the region from the mid-20th century onward, investigations produced internationally significant work on problems ranging from highway construction to various manifestations of environmental pollution (see below). The ranges are also of considerable interest to geologists exploring new concepts relating ocean-floor spreading to the evolution of mountains at the continental margins.

Seismographic investigation

Major relief divisions. The Coast Ranges are separated from the lofty Sierra Nevada and Cascade mountains, to the east, by broad depressions known in Oregon and Washington as the Willamette Valley–Puget Sound region and in California as the Great Valley. On the west, there is virtually no coastal plain, for deep water sets in within 25 miles of the coast.

The ranges themselves have six major divisions. The unit stretching from San Francisco north to the Klamath Mountains of the Oregon–California boundary is known as the northern Coast Ranges; the coastal ranges of Oregon and Washington have similar geological features and form a separate division north of the Klamath Mountains. The Klamath Mountains, however, although fronting on the Pacific and topographically part of the marginal Coast Ranges, are made up of older rock structures resembling those of the Sierra Nevada (*q.v.*). Hence, they are not usually included as a subdivision. The coast ranges of British Columbia are also of the Sierra Nevada type and decidedly different from those of Oregon or Washington. South of San Francisco, the central Coast Ranges extend to the vicinity of Santa Maria, at latitude 35° N; beyond, they are succeeded by a sector that includes the Santa Barbara, Ventura, and Los Angeles areas. The sector's name—the Transverse Ranges—derives from a marked east–west topographic and geologic trend. The ocean bed south of the Transverse Ranges consists of basins and platforms belonging to the continental margin, rather than to the Pacific deep, and it can be counted as an extension of the Coast Ranges. Finally, the southern tip of the great formation is formed by the Peninsular Ranges extending down the fingerlike protrusion of Baja California.

The Coast Ranges for the most part exhibit a subdued and soft relief of below about 3,300 feet, but some peaks and ridges stand above 6,600 feet or even 11,500 feet (3,506 metres) at San Gorgonio in the eastern Transverse Ranges.

Geological history. *The basic rock structures.* Although the Coast Ranges were molded into their present shape in the geologically recent Cenozoic Era (beginning about 65,000,000 years ago), an important element in their composition dates from as far back as the Jurassic period (190,000,000 years ago). The rocks concerned in this underlying, or basement, complex have been called the Franciscan Group. These sedimentary and volcanic rocks were subsequently changed by great pressures, sheared, intricately folded, and invaded by granite, then molten. Many of the rock layers making up the rest of the Coast Ranges are composed of sediment washed out by erosion from the then-existing margin of the North American continent to the east. These deposits accumulated in local structural basins and along the continental shelf and its slope.

The Franciscan rock basement

The emergence of the Coast Ranges. The subsidence of the sediment-filled basins and related processes of structural uplift and subsequent erosion provide the key to the complex evolution of the Coast Ranges. The trend of the basins and the associated uplifts generally paralleled the coast, but in the Transverse Range sector the basins took an east–west trend, subsided rapidly and deeply, and received very thick layers of sediment that eventually formed the rich petroleum source beds of southern California. The uplifts, on the other hand, often brought older Franciscan and Upper Cretaceous rocks to a new surface exposure.

This process of mountain building took place in two clear-cut phases. In the first, the Early Cenozoic rock layers were folded and thrust up as a result of earth movements in mid-Miocene time (about 20,000,000 years ago), and this process of deformation spread all the way from the Peninsular Ranges in the south to the Coast Ranges of Oregon and Washington in the north. The upfolded layers of the central and Transverse ranges were subsequently eroded almost to a plain and were then covered by additional layers of younger sedimentary rocks washed down from the east. This latter development set the stage for a second mountain-building movement, which began in Quaternary time, possibly 1,000,000 years ago, and has continued to the present, for the central and Transverse ranges are still being built up, deformed, and eroded.

Farther north, however, the Coast Ranges of Oregon and Washington participated only in the first of the two mountain-building movements, when their Early Cenozoic rock layers were most affected by folding. Sediments and lava flows accumulated to great thicknesses in a large arc-shaped basin lying between the Klamath Mountains on the south and the British Columbian Coast Ranges on the north. The first mountain-building movement caused a notable structural uplift in the north in the form of a dome, the site of the present Olympic Mountains of Washington. Since this region escaped a second mountain-building phase, the forces of erosion were able to cut away much of the uplifted rock, revealing the great upward folds, or anticlines, and their downfolded equivalents, the synclines. The summits in the Olympic Mountains exceed 6,500 feet, and the relief has been scoured and molded by the action of extensive Ice Age glaciers, portions of which remain to enhance the scenic attractions of the area.

The San Andreas Fault system. A major feature of the Coast Range system is the San Andreas Fault, a line of structural weakness extending from the Gulf of California northwestward to the Transverse Range and then westward to the central Coast Ranges. It has been traced northward to the San Francisco peninsula and beyond, to Cape Mendocino, where it appears to join the structurally weak submarine Mendocino Ridge. It sends out a number of associated faults, notably in southern California and the San Francisco area. The great fault marks the splitting of a huge sliver of the North American continental margin. The land west of the fault line is being forced, or, in geological terms, translated, to the northwest. Over the past 15,000,000 years, it has travelled some 200 miles. The translation has occurred in short, violent movements best illustrated by the 1906 San Francisco earthquake, when the land west of the fault moved about 21 feet horizontally northwestward along the fault line. An intensive program of seismic investigation is underway, aimed at predicting the almost inevitable future dislocations along the fault, which runs through areas of dense population.

Movement of the San Andreas Fault

The continental shelf. The underwater fringe of the Coast Ranges is also of exceptional geological interest. The bold, sea-pounded coast has few natural harbours, but San Francisco Bay provides an exception: through the narrow 360-foot-deep trench spanned by the soaring Golden Gate Bridge, it opens into one of the most magnificent harbours in the world.

Beyond the coastline, from Santa Barbara to Vancouver Island, the underwater shelf juts out above the Pacific abyss. It has a width of 15–30 miles and a maximum depth of about 650 feet, at which point it begins to slope down to its base. The shelf and slope are fairly regular in submarine relief, except where indented by such fissures as the Monterey Canyon. Where the sliver of continental crust is being pushed into the Pacific Basin, an underwater headland or projection of the shelf and slope is prominent, such as Cape Mendocino, latitude 40° 30′ N.

South of the Transverse Ranges the underwater topography changes from a narrow shelf to a wider basin and platform region about 150 miles wide. From measurements made of crustal thickness, this is believed to have once been a portion of the North American continent that subsided in separate blocks between a series of faults.

The Coast Ranges and sea-floor-spreading theory. In the 1960s and 1970s the new and popular theory of sea-floor spreading provided a framework into which the origin of the San Andreas Fault and the entire evolution of the Coast Ranges could be quite plausibly fitted.

The ocean bed is now held to be cracking into great trenches. These fissures are constantly healed by eruptions of new lava from the earth's core that form mid-oceanic ridges. The East Pacific Ridge is of great significance in the evolution of the Coast Ranges: it has been charted northeastward along the Pacific floor as far as the Gulf of California, where it is deflected westward a number of times by a series of faults. The San Andreas Fault further deflects the ridge as far as the ocean floor off Oregon. This slow process of sea-floor spreading is responsible for the northwesterly movements associated with the San Andreas Fault. The rate of movement is about one inch per year.

Climate and vegetation. The northern sections of the Coast Ranges experience a climate known as humid mesothermal; *i.e.*, with cool summers and mild winters. South of San Francisco the climate becomes mesothermal subtropical, with dryer summers. Water is plentiful to the north but scarce—as needs increase—further south.

The rain forests of the northern Coast Ranges are among the finest in the world. Coniferous stands make up all but 10 percent of the wooded area, and there are no fewer than 50 species of trees, including the famed giant redwoods, which are found mostly on the westward facing slopes.

Economic development. Nature has provided the Coast Ranges, particularly the California section, with diversified climate, scenery, and economic potential. Resources include gold, petroleum, vineyards, and a supply of fruits, nuts, and vegetables sufficient for half the American nation. These resources help to explain the westward population movement that has made California the most populous state in the Union. Settlement there is concentrated in the Coast Ranges and the adjacent Great Valley.

Apart from the social problems of the great urban complexes centred on Los Angeles and San Francisco, this expansion of settlement has had a deleterious effect on the natural environment of the Coast Ranges. Landslides, caused by the removal of vegetation, and barren and eroding scars have become an eyesore in many places, while atmospheric pollution, brought about chiefly by the automobile, adds a host of problems. Even the offshore areas have been affected: by the early 1970s the region adjoining Santa Barbara was the scene of much oil-leasing activity, and, as drilling got underway in 1969, a serious oil leak occurred at one offshore site. All these factors helped to make the Coast Ranges area a focus of great ecological and conservationist discussion in the 1970s.

Pollution problems

BIBLIOGRAPHY. M.N. CHRISTENSEN, "Late Cenozoic Deformation in the Central Coast Ranges of California," *Bull. Geol. Soc. Am.*, 26:1105–1123 (1965), an up-to-date rendition of the nature of deformation in the Central Coast Ranges; A.J. EARDLEY, *Structural Geology of North America*, 2nd ed., ch. 29 (1962), a general résumé (to 1960) of the geology of the Coast Ranges from Canada to Mexico; T.W. DIBBLEE, JR., "Evidence for Cumulative Offset on the San Andreas Fault in Central and Northern California," *Bull. Calif. Div. Mines Geol.*, 190:375–384 (1966), an authoritative article widely referred to for the amount of displacement on the fault; W.G. ERNST, "Tectonic Contact Between the Franciscan Melange and the Great Valley Sequence," *J. Geophys. Res.*, 75:886–901 (1970), an article in the forefront of

modern thinking on the origin of the Coast Ranges in California, representing the adaptation of the postulates of seafloor spreading to earth-building movements of the Pacific Coast; J.C. CROWELL, "The California Coast Ranges," *UMR Journal*, Series 1, pp. 133–156 (1968), a modern résumé of Coast Range geology with emphasis on the relation of the San Andreas Fault to the other range structures.

(A.J.E.)

Pacific Islands

The Pacific Islands is an expression commonly accepted as including all of those islands in the Pacific Ocean that are collectively referred to as Melanesia, Micronesia, and Polynesia. This usage rules out the Australian island continent, the Asia-related Indonesian, Philippine, and Japanese archipelagoes, and the Ryukyuan, Bonin-Volcano, and Kuril island arcs that project seaward from Japan. Neither does the term encompass the Aleutian chain connecting Kamchatka and Alaska nor such isolated islands as Cocos, Galápagos, and Juan Fernandez off the coasts of Central and South America.

Although the Pacific Ocean makes up nearly one-third of the Earth's surface, the Pacific Islands discussed in this article add up to a little less than 500,000 square miles (1,300,000 square kilometres) of territory. New Guinea, the largest island in the world after Greenland, represents 60 percent of this total, and New Zealand accounts for 20 percent. The remaining one-fifth of the land area in the Pacific is divided among more than 10,000 islands. The Pacific Islands exist mainly in the area bounded by latitudes 12° north and 25° south and longitudes 130° east and 130° west. Exceptions to this are the Mariana Islands, which extend northward toward the Bonin-Volcano chain; the Hawaiian Islands, which straddle the Tropic of Cancer; New Zealand, which lies in the southern temperate zone; and Easter Island, which stands in isolation almost halfway to South America. (For a detailed discussion of the Pacific Ocean see the article PACIFIC OCEAN; for a detailed consideration of individual islands, see the articles BIKINI; EASTER ISLAND; FIJI; FRENCH POLYNESIA; HAWAII; NEW ZEALAND; and SAMOA; as well as the articles AUSTRALIAN EXTERNAL TERRITORIES; and UNITED STATES OUTLYING TERRITORIES. For historical aspects, see the articles NEW ZEALAND, HISTORY OF; and OCEANIA, HISTORY OF.)

Melanesia, Micronesia, and Polynesia

For convenient reference, the Pacific Islands are customarily divided into three ethnogeographic groupings. The great arc of islands located north and east of Australia and south of the Equator is called Melanesia (from the Greek words *melas*, "black," and *nēsos*, "island") after the predominantly dark-skinned peoples of New Guinea, the Bismarcks, the Solomons, the New Hebrides, New Caledonia, and Fiji. North of the Equator and east of the Philippines is another island arc that ranges from Palau and the Marianas in the west through the Caroline and Marshall clusters all the way to the Gilbert Islands. This is Micronesia, named in recognition of the smaller size of these islands and atolls. In the central Pacific, and largely enclosed within a huge triangle formed by Hawaii in the north, New Zealand to the south, and Easter Island far to the east, are the "many" (poly) islands of Polynesia. Other components of this widely scattered collection are Samoa, Tonga, and the Society, Tuamotu, Marquesas, and Cook Islands. In this, the last section of the Pacific Ocean to be occupied by man, the islanders share a cultural tradition that relates them closely to many Fijians. Fiji, indeed, is actually a transitional territory between Melanesia and Polynesia.

From the 16th century onward the Western world has shown an interest in the Pacific Islands that has been expressed in the activities of explorers, scientists, poets, missionaries, commercial entrepreneurs, and imperialistic statesmen. The endless variety of the Pacific's environments, both physical and biotic, has been, and still is, a laboratory for experimenting in social and cultural adaptation. Though insularity has often dominated this process, the effect of this has been offset by the opportunities for human contact and exchange in many directions across the ocean's expanse. In the 20th century the islands and their inhabitants continue to attract international interest, although for new reasons, such as their strategic significance in the balancing of power between East and West, and the problems created by nature's limitation of land and resources in the face of expanding populations and accelerating living standards.

THE ISLAND ECOSYSTEM

To know what it is like to live on a Pacific island, the intermixture of physical and biological characteristics of the particular island must be considered. Each of the myriad ecological systems in the Pacific is a unique complex of living organisms and their nonliving environment. Each is a functional system of interacting components that tends toward an equilibrium that is never quite achieved. The limited size of most Pacific islands makes almost any change, whether by the hand of man or by some natural agency, capable of repercussions elsewhere within the ecosystem. The landform, climate, soils, vegetation, and animal life—all are elements to which people who live on an island must relate themselves and their behaviour, for they, too, occupy a niche in the total ecological scheme.

Island types. The islands may be classified as either continental or oceanic. The former are associated with the ancient continental platforms of Asia and Australia, now partially submerged. Oceanic islands, located eastward in the deeper Pacific basin, are differentiated as high volcanic or low coral islands. A coral island may be single, or two or more coral islets may form an atoll if connected by a reef ringing a lagoon. The "high–low" distinction is misleading as the two types occur in many combinations, and some coral islands have been elevated considerably by changes in the ocean level.

Continental and oceanic islands

Continental islands. Faulted and folded in mountainous arcs that parallel deep ocean trenches, the islands of the broad western Pacific margin are formed mainly of metamorphosed sediments and of intrusive volcanic materials with a high content of silica and alumina. These islands are separated from the deeper Pacific basin—along the eastern borders of Kamchatka, Japan, the Marianas, New Guinea, the Solomons, Fiji, and New Zealand—by a geological demarcation known as the andesite or sial line. (These names refer to distinguishing features: the volcanic rock andesite is distinguished from basalt, the principal lava rock within the basin; the sial rock layer—composed principally of silica and alumina—underlies all continents.) This line marks the eastern limits of the earth's greatest volcanic activity and earthquake disturbance. Continental islands are generally larger, and more numerous, and have richer mineral-bearing soils that support almost every kind of vegetation. New Guinea, 1,300 miles long and with a maximum width of nearly 500 miles, is a good example. Its snow-capped mountains rise to 16,400 feet, its interior is dissected by high plateaus and extensive river systems, and its slopes and coastal margins are covered with dense forests and vast swamps.

High volcanic islands. Extensive mountain ranges in the Pacific basin proper rise abruptly from the ocean deep, their cores of dense black basalt built up from lava flows from the fractured sea floor. Where summits stand above sea level, they represent most of Polynesia and Micronesia. Small to intermediate in land area, nowhere do they match the extent of continental islands. Hawaii can boast a snow-topped peak 13,796 feet high, though most oceanic islands have peaks of somewhat less than 5,000 feet. Topography is extremely rugged, with sharp peaks and ridges, deep canyons, cliffs, and waterfalls abounding. Human communities occupy the more congenial lower slopes, floodplains, and wide strands. The islands, rich in iron and magnesium oxides, are densely forested but lack the metallic mineral wealth of continental islands.

Coral islands. Most Pacific islands are coral formations, although all have volcanic or other cores. In the shallow waters of the tropics, both continental and oceanic islands attract coral growth in the form of *fringing reefs*, partially submerged platforms of consolidated limestone, with coral organisms at the ocean edge feeding

on materials carried in by sea waves and currents. Coral-building polyps and algae secrete calcium carbonate (lime) from seawater, forming skeletal frameworks that adhere to land surfaces, or the rock remains of coralline ancestors. The gradual submergence of some islands, and a continuing coral growth outward, produce a *barrier reef* formation farther from shore and separated from it by a narrow lagoon.

Coral atolls

A coral *atoll* results when further subsidence reduces an island to a submarine condition. The usually irregular reef continues to flourish in the warm shallows: it encircles a clear-surfaced lagoon of moderate depth, and in time supports a number of islets, known as motus, built up from reef debris to 20 or 30 feet above sea level. Atolls of all shapes and sizes may be formed. Kwajalein Atoll in the Marshalls is the world's largest, being about 90 miles long and 20 miles wide, with a lagoon area of 839 square miles, and 38 islets totalling just over 6 square miles of land. Openings, or passes, which commonly occur on the leeward side of atoll reefs, permit access to the lagoon from the sea. The only source of fresh water is rain.

Successive elevations of an island above sea level create a variety of "raised" coral formations. Raised fringing reefs on Guam, for example, have become coastal limestone cliffs that rise to nearly 600 feet. Nauru and Ocean Island, as raised atolls, stand at an elevation of about 200 feet, and possess deeper soil and a more adequate water supply than other types of atoll, as well as refined guano deposits from which phosphate is mined commercially.

Climate. To describe the climate as tropical and oceanic is to stress the influence of the lower latitudes and the tremendous expanse of the sea. Humidity and temperature tend to be high and are generally uniform throughout the year. Regional differentiation in climate is linked principally to rainfall patterns. Here, factors of altitude and longitude, as well as latitude, come into play as they relate to the various systems of air circulation that prevail in the Pacific. The climate of New Zealand, which lies well outside the tropics, constitutes an exception.

Across the eastern and central Pacific, air currents flow south and north from the midlatitudes toward the Equator and, trending westward, form the northeast and southeast trade winds. These brisk winds bring light to moderate showers interspersed with brief rainy spells or clear skies. The windward sides of high islands are cloudy and wet, in contrast to the relatively dry leeward coasts. Short seasonal shifts in wind direction often presage stormy weather. Where the trade winds meet near the Equator lie the doldrums, a region of little or no wind, considerable cloud, and high humidity. The trade winds merge or give way to the monsoon winds in the far western Pacific, where the alternate cooling and heating of continental Asia produces a seasonal reversal of winds. From about November to March, the northwest monsoon from Asia brings rain to the northerly slopes of the western Carolines, New Guinea, and the Solomons. In summer the southeast monsoon reverses the process.

Typhoons, or hurricanes, occur frequently in western Micronesia from July to November, and are active south of the Equator from Australia to the Society Islands four to six months later. These winds of gale force are accompanied by torrential rains and tidal waves, and cause extensive damage to crops and buildings, especially on low-lying coral islands.

Soils. Pacific island soils develop through the action of temperature, rainfall, and organic matter on the original rock materials. This process is further influenced by factors of time and land relief. Coral island soils are the least mature, and are deficient in organic materials, low in fertility, and too porous to ensure proper soil moisture. The mineral-bearing continental islands are geologically more complex and, being also favoured by a longer period of weathering, possess richer and more varied soils than the volcanic islands. The most productive soils on high islands occur in the lower slopes and valleys, in the form of alluvial floodplains and deltas, and of recent volcanic ash deposits. Tropical temperatures and rainfall have produced a series of soils (laterites) from which nutrients have been leached. These soils, of only moderate fertility, decline rapidly after two or three years of crop use; fertilizers must then be added, or else the soil must be abandoned, and fresh land cleared and planted.

Vegetation. Most island vegetations reveal Asian ancestries stemming from Indonesia and New Guinea. Generic variety declines eastward across the Pacific, providing evidence that transient seeds and fruits carried by ocean currents, birds, winds, and early man, encountered mounting obstacles to acceptance. New Zealand was host to plant dispersal movements from South America via Antarctica. Plant adaptation to local differences in moisture, soil, salinity, and temperature has resulted in countless new, endemic species. Plant introductions from other world sources during the past century have, however, markedly altered island vegetations.

Seacoast vegetation

The seacoast, or strand, vegetation is the most widespread of Pacific zonal types. Depending on the availability of moisture, this relatively unfavourable setting supports shrubs, herbs, and woody vines, as well as such trees as coconut, casuarina (which takes its name from the resemblance of its twigs to the feathers of the cassowary bird), pandanus (screw pine), and hibiscus. Mangrove thickets proliferate in brackish swamps. Breadfruit, banana, papaya, and tubers resembling the taro (a plant with an edible starchy rootstock) are cultivated either in moist locations or in pits enriched with plant debris. On high islands, primary forests still survive in valley bottoms, as well as on lowland plains and on intermediate slopes. The prevalent rain forest constitutes a community of huge trees that overlook smaller trees and a ground growth of shade-tolerant ferns, vines, and shrubs. Species differentiate enormously. Yam, cassava, taro, and sweet potato are important staples in high island economies. Secondary forests and savannas (grasslands) have replaced virgin forests destroyed by fire, or by shifting cultivation practices. Grasslands are also associated with areas of poor soil and little rain, as on leeward slopes.

Animal life. Bats, rats, and, in New Guinea, marsupials such as tree kangaroos, flying opossums, and bandicoots were the only mammals to precede man into the Pacific, after which pigs, dogs, poultry, goats, deer, and cattle were introduced. Most islands have some snakes and lizards, but crocodiles are restricted to the west. Sea birds, such as terns, frigates, albatrosses, petrels, and boobies, supplemented by migratory ducks, plovers, and curlews, are found almost everywhere. While oceanic islands support only a few land birds, New Guinea has exotic species such as the cockatoo, hornbill, bird of paradise, and cassowary, while New Zealand is the home of the kiwi.

The abundant marine life is infinitely valuable for human subsistence. Reefs and lagoons provide lobsters, shrimps, clams, crabs, oysters, snails, eels, octopus, turtles, and innumerable fish species. In deeper waters are tuna, bonito, swordfish, and many other sport fishes. Sharks are prominent predators, and whales have been commercially fished. Insects are unquestionably the most numerous of island fauna. The more pestiferous include centipedes, cockroaches, lice, houseflies, and the malaria-carrying *Anopheles* mosquito, which exists in Melanesia from coastal New Guinea to the New Hebrides. Scorpions are also found.

HUMAN OCCUPANCY

Man's adaptation to island environments

Occupying a niche in the food chain, man in the Pacific has had to adapt to island environments just as other species have done. The technologies and organizational systems introduced into the Pacific by migrants had earlier been established in environments that varied from receptive to hostile. The earliest arrivals, as food-gathering peoples, probably provoked little disruption of the ecological balance (*i.e.*, the balance between organisms and their environment). Their aboriginal successors, practiced in horticulture and skilled in transporting themselves across the sea, were able in limited degree to fashion and control physical habitats after their own custom. Later, Western man with his advanced technology and civilization often threatened the balance of the inherently

unstable ecosystems (complexes of ecological community and environment forming a functioning whole in nature).

Aboriginal groupings. Natives of the Pacific tend to identify themselves by their home island or mother tongue, saying, for example, that they are from Nauru, or that they speak Fijian. Occasionally, however, they may invoke another, and larger, identity, claiming to be Polynesian, Micronesian, or Melanesian. As a geographic designation this practice has value, but as a mark of racial, linguistic, or cultural affiliation it is misleading. While Melanesians appear to be more Negroid and Micronesians more Mongoloid, a great deal of racial mixture throughout the Pacific has taken place since the first immigrants arrived in the southwest. The linguistic pattern is also complex. Some valid generalizations about traditional practices and institutions in the three areas may nevertheless be made, although it must be remembered that overlap between the areas is common, and that there are exceptions to every statement.

Polynesia. Polynesians are strikingly homogeneous in speech, custom, and physical appearance, although western Polynesians are moderately distinct from the rest. Accomplished as gardeners and fishermen, they have directed their principal energies to nonmaterial pursuits. Epic mythology, copious genealogies, sophisticated social etiquettes, hereditary aristocracies, and elaborated religious formalism, characterized society in pre-European Polynesia, receiving varying degrees of emphasis. A kinship system that accepted the worth of both maternal and paternal family ties supported group solidarity in community enterprises. Secular leaders, as lineal descendants of deified ancestors, served gods and humans alike. The social-religious-political hierarchies encouraged and rewarded aesthetic creativity in wood and stone sculpture, featherwork, tapa (barkcloth), and tattooing, according privileges to the artist, commensurate with those accorded the warrior, navigator, herbalist, and seer.

Micronesia. Eight or ten cultural-linguistic areas in Micronesia attest to its greater heterogeneity. The Palau Islands, the Yap Islands, and the Marianas suggest affinities with Melanesia and Indonesia. The habitual placing of island farmsteads and hamlets near the shore reflects the prevailing interest in fishing, canoeing, and overseas trade. Except in the Gilbert Islands, matrilineal clans and lineages characterized by marriage outside each group influence property inheritance, succession to traditional titles, and intracommunity competition. Local political autonomy was formerly overshadowed by loose confederations and tribute arrangements in western Micronesia, as well as in the area of Ponape Island, Kusaie Island, and the Marshall Islands, where class stratification still exists today. Indigenous religions lack formality, and are largely of personal or family concern. Art is mainly decorative, and is manifest in matwork, shell ornament, loom weaving, tattooing, and functionally crafted wood and stone artifacts.

Melanesia. This is a region of unending contrast. "Beach" natives, who have acquired advantages from coastal trading and cultural exchanges, may be compared with more isolated and traditional "bush" (unwesternized) populations inland. Polynesian influence touches Fiji and outlying islands to the northwest. The massive extent of New Guinea, with its thousands of indigenous groups, requires separate consideration. Melanesians are gardeners, with a penchant for pig raising, who promote social mobility by accumulating wealth through trading and competitive feasting; *i.e.*, local notables acquire more influence by giving more feasts than their rivals. Unilateral descent groups (*i.e.*, those that trace descent through either the maternal or paternal line only), usually patrilineal, are the essence of community organization. Leadership depends on the local "big man," whose personally acquired power gives him authority in his own village and influence in nearby villages. Headhunting and raiding of neighbouring tribes continue in remote parts of New Guinea. The animistic religion, a mixture of magic, sorcery, totemism, and ancestor worship, is dominated by men's clubhouses, secret societies, and elaborate initiations. Art forms associated with these activities include dance masks, sculptured figures, body scarification, and carved mortuary standards.

The indigenous heritage. The earliest humans entered Oceania after 20,000 BC, drifting into New Guinea from the Indies as small groups of migratory hunter-gatherers. Representing a genetic intermixture of Negrito and ancient Caucasian, they spoke languages ancestral to those nowadays termed "Papuan." Between about 3000 and 2000 BC, other migrants from Indonesia had begun to move into Melanesia in sailing canoes. These were traders, skilled in root and tree crop cultivation. Of Mongoloid provenance, and speaking Austronesian (Malayo-Polynesian) languages, they introduced their genes and customs to the simpler habitants of New Guinea and in time occupied all of Melanesia. Similar peoples entered western Micronesia from the west between about 2500 and 2000 BC. Later, in a continuing pattern of accidental voyaging, some natives of the New Hebrides region sailed northward to the Carolines, while others sailed eastward to Tonga and Samoa. This movement out of eastern Melanesia, between about 1500 and 1000 BC, was linked with the diversification of the Malayo-Polynesian languages, the largest group in the Austronesian phylum (a category of language stocks considered likely to have a common origin). All of Polynesia was settled by AD 1000, following migrations from the Marquesas and Society islands. Some Polynesians travelled westward and established outlying settlements in Melanesia and Micronesia. Polynesians were also in contact with South America, but American Indian influence was only nominal.

Early migratory movements

The alien heritage. More than five centuries passed before Balboa viewed the Pacific Ocean in 1513, and Magellan sailed across it in 1521. Other Europeans who pioneered in rediscovering the islands included the 16th-century Spanish explorer Álvaro de Mendaña de Neira, the 17th-century Dutch explorer Abel Janszoon Tasman, Louis Antoine de Bougainville, the French navigator, and the English explorer Captain James Cook, whose 1778 landfall in Hawaii ended the era of exploration. During the next seven decades, American and British whalers often wintered in the tropical isles, while shipwrecked sailors and fortune hunters settled in the region, as also did traders, who imported Western manufactures in exchange for sandalwood, coconut oil, and pearls. In 1797 the London Missionary Society became active in southern Polynesia. Other Protestant groups also entered the region, the Boston Mission working in Hawaii and eastern Micronesia, and the Anglican Mission in Melanesia. Though Spanish priests had built a mission in the Marianas in 1668, Catholicism was established on other Pacific islands only after French missionaries based on Tahiti became active in 1836.

Between 1850 and World War I, political and economic imperial interests were furthered when the governments of Germany, the United Kingdom, France, Spain, and the United States claimed island groups as colonies and protectorates. Corporate businesses, such as those of Godeffroy of Hamburg and Burns Philp of Sydney, extended their jurisdictions over shipping, trade, and plantation development. Large-scale production of sugar and copra encouraged the illegal recruitment of local labourers, although more satisfactory contract systems later brought shiploads of Indian workers to Fiji; Javanese and Tonkinese workers to New Caledonia; and Chinese, Japanese, and Filipino labourers to Hawaii. Many of these workers later settled, establishing their own commercial enterprises. Australia and New Zealand, areas of British colonialization, gained dominion status after 1900, subsequently strengthening their own spheres of influence by occupying offshore territories. Hawaii, annexed by the United States in 1898, like the two dominions began to play a role as a self-conscious agent of westernization in Oceania. Spain's withdrawal from the Pacific followed the Spanish-American War of 1898, and the onset of World War I ended Germany's regime as a colonial power.

Battle operations in World War I barely affected the Pacific. Japan, Australia, and New Zealand were mandated responsibility for the former German colonies by the

League of Nations. In violation of the mandate, Japan in the 1930s annexed Micronesian territories. During World War II administering authorities in the South Pacific were taken unaware when Japanese forces crossed the Equator in 1941. In the next three years Western superiority faltered but then was recovered as the tide of war turned.

From trusteeship to independence

Following Japan's surrender, territorial administrations accorded greater recognition to the islanders' welfare in the spirit of trusteeship pledged to the United Nations. By 1970 some island groups had won political independence. Meanwhile, tourists have been visiting the South Sea islands in increasing numbers.

Interaction and reaction. During the past four and a half centuries of culture contact there has been a constant flux between change and disorganization on the one hand, and stability and reintegration on the other. Relatively balanced ecosystems of prehistory were disrupted when islanders, reacting to the novelty and authority of Western civilization, discarded traditional crafts, directed their skills to serve new economies, and, under pressure from missions and alien governments, at least nominally adopted practices and beliefs that were strange to them. From this initial confusion there emerged reasonably stabilized island societies that, while preserving the native ethos, reflected a fusion of elements from both cultures. European, American, and Asian residents continued to conserve their own identities in cultural enclaves. Meanwhile, increasing influence was exercised by a population of mixed blood, formed of culturally marginal individuals.

The direct impact of World War II on island communities stimulated many islanders to seek a greater participation in Western society. Their efforts were expressed in two ways. The first was in a proliferation of mystical cults. Occurring principally in Melanesia, these cults blended traditional and Western practices but provided solutions only in fantasy. The second, the way of economic and political nationalism, was supported mainly by native and part-native residents in urbanized port towns and island capitals who had benefitted by travel and higher education. Opposed to assimilation, the latter held pride in their island origins and negotiated for more equal treatment within the framework of Western civilization.

Population dynamics. Almost 7,500,000 people are estimated to inhabit the Pacific islands. Continental islands are thinly settled with five to 17 persons per square mile. Many smaller islands, in Micronesia and Polynesia, have high population densities. On Nauru, for example, the density is 863 per square mile. New Zealand excluded, the tropical Pacific population is 81 percent native, 13 percent Asian (mainly Indian, Japanese, Filipino, and Chinese), and 6 percent Caucasian of whom three-fourths live in Hawaii. Most of New Zealand's 2,900,000 people are white, only 8 percent claiming to be half or more of Maori descent. Other anomalies are the Indian preponderance over native Fijians in Fiji, and Hawaii's mixture of Japanese, Caucasian, Filipino, and Chinese.

Killing epidemics of Western-introduced disease contributed to population decline until around 1900 when they were checked by modern medicine, health education, and improved diet, which allowed a high native fertility to re-assert itself. Many island populations almost trebled in the next five decades, with crude birth rates for 1956 being estimated at 36 to 45 per 1,000 population, as opposed to recorded death rates of 7 to 16. Family planning and economic pressures have not helped to retard the fast rate of population growth. The resettlement of individuals and of whole communities from coral atolls, as in the Gilberts and Carolines, has only momentarily relieved the crowded condition.

Emigration to port towns

Voluntary emigrants to such port towns as Papeete in Tahiti, Agaña in Guam, Suva in the Fiji Islands, and Port Moresby in Papua have responded in part to overcrowding at home but more often seek employment, entertainment, or further education. As such towns are built around trade, mission, hospital, school, and administration facilities, they service hinterland populations and link them to the outside world by sea, air, and tourism networks. Almost universally these heterogeneous, multiracial communities suffer from poor housing, underemployment, and juvenile delinquency; gambling and drinking are widespread. Native and part-native residents, growing in numbers and more vociferous in their demands, still do not fulfill their aspirations to participate more fully in these artificial societies.

POLITICAL AND ECONOMIC SYSTEMS

During the 20th century's third quarter, overpopulation in a region of fragmented land areas, widely scattered populations, poor communications, inadequate resources, and rising costs of living posed a fundamental dilemma. Political responsibility for the situation rested largely with the five metropolitan nations administering the possessions, protectorates, and trusteeships that perpetuated a no longer popular colonial heritage. Amelioration of social and economic conditions seemed to await changes in the political environment, while the question of the five nations' willingness to share their territorial commitments through international cooperation and with the emerging native elites remained undecided.

Administrative groupings. Two United Nations trust territories have achieved independence—Western Samoa in 1962, and Nauru in 1968. The first continued to depend on New Zealand in foreign affairs, while Nauru ended its ties with Australia, New Zealand, and the United Kingdom. When British Fiji and Tonga became independent in 1970, both elected to remain within the Commonwealth of Nations.

New Zealand, a Commonwealth member with a Maori Polynesian population surviving on its territory, maintains close relations with Polynesian territories to the northeast—the Cook Islands, Niue, and the Tokelaus. Cook Islanders in 1965 became self-governing in a free association with New Zealand, from which they required only support in external affairs, defense, and finances. The people of Niue and the Tokelaus are administered more directly from the metropolitan nation. Like the Cook Islanders, they have migrated in great numbers to New Zealand. Australia, a Commonwealth country with exclusively Melanesian interests, created the Territory of Papua and New Guinea in 1949. It combines its administration of Papua, which was acquired from the British in 1906, with its United Nations trusteeship administration of northeast New Guinea, which it was granted in 1946. British territories include the Solomon Islands protectorate, the Gilbert and Ellice Islands colony, and the British portion of the New Hebrides condominium. These territories are scattered throughout the South Pacific, and are all administered from Honiara, on the island of Guadalcanal in the Solomons, by the British High Commission for the Western Pacific.

The United States dominates the island region north of the Equator. Hawaii, formerly a territory, became the 50th U.S. state in 1959. Other territories, under civilian control after 1950–1951, are American (eastern) Samoa; the Trust Territory of the Pacific Islands, a strategic area trusteeship under the United Nations, formerly a Japanese mandate, in Micronesia; and Guam, a U.S. outlying territory whose people are U.S. citizens. France is represented in the Pacific by three overseas territories—French Polynesia, New Caledonia, and Wallis and Futuna Islands—as well as in the New Hebrides condominium shared with the United Kingdom. The New Caledonian governor, as High Commissioner, is responsible for French interests in the condominium. Each French overseas territory has a degree of local autonomy and is represented in the French parliament by an elected deputy and senator.

Two territories are politically and geographically peripheral to contemporary Oceania. Western New Guinea, a former Dutch colony, became a province of the Republic of Indonesia in 1963, and was renamed West Irian (Irian Barat). Easter Island was acquired by Chile in 1888, and is leased to a Chilean stock-raising company.

The international coordination of health, social, economic, and educational development is organized by the South Pacific Commission, established in 1947 to advise island governments. A research council advises the commission in its work. Commissioners meet annually at

Administrative Groupings of Pacific Islands

territory	area (sq mi)	population (1970 estimate)	capital	political status and remarks
American (eastern) Samoa	76	29,000	Pago Pago	U.S. unincorporated territory
British Solomon Islands	11,500	163,000	Honiara	U.K. protectorate
Cook Islands	93	24,000	Avarua, Rarotonga	N.Z. self-governing territory
Easter Island	45	1,200	—	Chilean dependency
Fiji	7,055	531,000	Suva	independent parliamentary state (1970), former U.K. colony; includes Rotuma Island
French Polynesia	1,544	109,000	Papeete	French overseas territory; includes Clipperton Island, Marquesas Islands, Society Islands, Tuamotu and Gambier Islands, and Tubuai (or Austral) Islands
Gilbert and Ellice Islands	376	56,000	Bairiki, Tarawa	U.K. colony; includes Central and Southern Line Islands (also claimed by U.S.), Northern Line Islands (except Jarvis, Palmyra, and Kingman Reef), Ocean Island, and Phoenix Islands
Guam	206	100,000	Agana	U.S. organized unincorporated territory
Hawaii	6,424	769,000	Honolulu	U.S. state
Nauru	9	7,000	Domaneab	independent republic (1968)
New Caledonia	7,082	109,000	Nouméa	French overseas territory; includes Bélep, Chesterfield, Huon, Loyalty, and Walpole Islands, and Isle of Pines
New Guinea, Northeastern	93,000	1,752,000	Port Moresby	U.N. trust territory, administered by Australia as part of Territory of Papua and New Guinea; includes Bismarck Archipelago, northeastern New Guinea, and Bougainville, Buka, and adjacent islands
New Guinea, Western (Irian Barat)	160,000	907,000	Djajapura	province of Indonesia
New Hebrides	5,700	86,000	Vila	Anglo-French condominium
New Zealand	103,736	2,853,000	Wellington	parliamentary state
Niue Island	100	5,000	Alofi	N.Z. territory
Papua	90,540	669,000	Port Moresby	Australian territory, administered as part of Territory of Papua and New Guinea; includes D'Entrecasteaux, Louisiade, Trobriand, and Woodlark islands, and the southeastern part of the island of New Guinea
Pitcairn Island	2	82	Adamstown (defacto)	U.K. colony
Tokelau Islands	4	2,000	—	N.Z. territory
Tonga	259	86,000	Nukualofa	independent state (1970)
Trust Territory of the Pacific Islands (American Micronesia)	706	107,000	Saipan	U.N. trust territory, administered by U.S.; includes Caroline, Marshall, and Mariana islands (except Guam)
Wallis and Futuna	60	9,000	Mata Utu	French overseas territory
Western Samoa	1,140	143,000	Apia	independent state (1962)

Source: United Nations; official government figures.

International coordination of development

headquarters in Nouméa, following the South Pacific Conference attended by island delegates, and which has been held in Nouméa in New Caledonia, Suva in Fiji, Lae in New Guinea, Rabaul on the Australian-administered island of New Britain, and Pago Pago in American Samoa. Proposals to add a political dimension to the commission's work have not been accepted. Commission membership includes Australia, France, New Zealand, Nauru, the United Kingdom, the United States, and Western Samoa.

Communication networks. Most islanders in the 1970s, by resorting to some combination of track, road, canoe, interisland freighter, local air service, or international airline, can travel within a week or two to such distant cities as Sydney, Auckland, San Francisco, Vancouver, Tokyo, or Singapore. Inhabitants of remote islands may need two months more to make such a journey, owing to less frequent or irregular travel schedules. More than a score of shipping and airline companies ensure the movement of inbound and outbound cargo and passengers through such island centres as Honolulu, Agana, Port Moresby, Honiara, Vila (in the New Hebrides), Nouméa, Suva, Pago Pago, and Papeete, thus bringing hinterland populations into closer touch with countries on the Pacific rim. This readier access, in both directions, has developed partly because World War II created new traffic patterns and facilities, partly because islanders have demanded better transport, trade, and mail services, and partly because of the growth of the tourist industry.

Communication by radio, radiotelephone, cable, and news services has also improved. Two-way radio transmission can inform territorial service centres of emergencies and needs in even the most distant locations. Local broadcasts and news releases in English, French, pidgin English (in western Melanesia), and indigenous tongues, inform the island public about world events.

Economic development. Pacific islanders, as producers of agricultural, mineral, and marine commodities, face problems of market demand, labour supply, development planning, and transportation that relegate them to an insignificant role in world trade. Their small and scattered populations do not constitute an attractive consumer market. Present aims are to achieve more self-sufficiency, and to export selectively so as to cover the cost of importing essential machinery, petroleum products, motor vehicles, textiles, and other manufactures.

Economic handicaps

Coconut products, principally copra from which the oil is extracted overseas, form the principal export in nearly every territory. Production, which depends as much on native family enterprise as it does, for example, on the plantation system that predominates in Melanesia, is likely to continue as the mainstay of most island economies. Sugar, exported mainly from Hawaii and Fiji, requires careful management, costly machinery, and specialized labour for its production. Perishable fruits, such as pineapples and bananas, demand markets close at hand unless locally processed. Experiments in growing coffee, cacao, and other cash crops have been undertaken with a view to stimulating economic diversification, so as to avoid the hazards of a one-crop economy. Grassland areas support

the production of beef cattle and sheep as subsidiary exports. Timber is processed commercially on the larger Melanesian islands.

New Caledonia is rich in nickel, chrome, and other metallic ores, New Guinea and Fiji have profited from gold discoveries, and oil dominates West Irian's export production. Reserves of natural phosphates on Nauru and Ocean Island are expected to last for 30 years before they become depleted. Marine resources, although almost unlimited, require skilled labour and capital facilities for commercial exploitation. Deep-sea fishing for tuna and bonito is practiced, mostly by Japanese crews. Canneries in Hawaii, American Samoa, Fiji, and the Carolines are in operation. Local island cooperatives have also successfully marketed fresh fish.

About half of the Pacific Islands' exports are sent to Australia and New Zealand, Japan, Canada, and the United States; most of the rest go to northwestern Europe. Traders based in Australia control most of the import commerce south of the Equator. Government grants and training programs, many promoted by the South Pacific Commission, have developed native experience in labour, marketing, and management. To some extent this has discouraged foreign, profit-seeking enterprise, although outside capital is needed. In the 1970s tourism opened up new income opportunities for islanders. Tahiti, American Samoa, and Fiji followed Hawaii's earlier lead in developing the tourist industry. Other territories are also attracting visitors by improving transport and accommodation, while at the same time endeavouring to safeguard local custom against excessive commercialization.

BIBLIOGRAPHY. K.B. CUMBERLAND, *Southwest Pacific: A Geography of Australia, New Zealand, and Their Pacific Island Neighbors*, rev. ed. (1968), an authoritative account of the political, social, and economic problems south of the Equator; O.W. FREEMAN (ed.), *Geography of the Pacific* (1951), a college text on physical, human, economic, and political geography by regional specialists; GREAT BRITAIN, NAVAL INTELLIGENCE DIVISION, *Pacific Islands*, 4 vol. (1943–45), the most complete geographic account available with regional chapters on island groups, prepared for intelligence use in World War II; N.V. HARRIS, *The Tropical Pacific* (1966), an authoritative high school geography textbook on the Pacific, excluding New Zealand; F.M. KEESING, *The South Seas in the Modern World*, rev. ed. (1945), the best analysis of the colonial impact on island peoples and cultures to 1939; N. MCARTHUR, *Island Populations of the Pacific* (1968), a demographic analysis of Fiji, Tonga, Samoa, the Cook Islands, and French Polynesia to 1956; D.L. OLIVER, *The Pacific Islands*, rev. ed. (1961), the best review to the mid-20th century of the historical consequences of westernization; C.R.H. TAYLOR, *A Pacific Bibliography: Printed Matter Relating to the Native Peoples of Polynesia, Melanesia, and Micronesia*, 2nd ed. (1965), an extensive Pacific Islands bibliography to 1960, indispensable for the specialist; A.P. VAYDA (ed.), *Peoples and Cultures of the Pacific: An Anthropological Reader* (1968), a representative collection of articles by specialists on Polynesia, Micronesia, and Melanesia. Current information may be found in the journal *South Pacific Bulletin* (quarterly), technical articles on economic and social developments written by territorial specialists; and in the *Pacific Islands Year Book and Who's Who*, 10th ed. (1968), the most detailed reference work on contemporary island conditions and events.

(L.E.M.)

Pacific Islands, Trust Territory of the

The Trust Territory of the Pacific Islands, a United Nations strategic-area trusteeship administered by the United States, consists of more than 2,000 islands, 96 of them populated, scattered over about 3,000,000 square miles (7,770,000 square kilometres) of the tropical western Pacific Ocean. The area lies north of the Equator between latitudes 1° and 22° N and longitudes 130° and 172° E. The total land area of the islands is about 700 square miles. The population of the group numbered 107,054 in 1971.

The territory generally covers the ethnic area known as Micronesia ("tiny islands") and is composed of three major island groups—the Marianas, Carolines, and Marshalls. The Micronesian Gilbert Islands to the southeast and Guam in the Marianas are politically excluded, while the islands of Kapingamarangi and Nukuoro, which are Polynesian instead of Micronesian in language and culture, are included. The political definition of the area was inherited from the Japanese, who governed it before World War II. (For articles on the region, see PACIFIC ISLANDS; PACIFIC OCEAN; and BIKINI. For historical aspects, see OCEANIA, HISTORY OF.)

Three types of islands

The natural environment. Three basic types of islands exist in the trust territory—low coral islands, or atolls; high islands of volcanic origin; and islands that represent a combination of coral limestone and volcanic uplift. The Marshalls and the eastern Carolines are generally of the coral variety, with elevations seldom exceeding 30 feet above sea level. In the western Carolines and Marianas are volcanic "high" islands, built upon a range of submarine mountains. The largest, Babelthuap in the Palau group, has 143 square miles of land area. Ponape, the second largest, has approximately 129, while Saipan, site of the administrative capital, has 47. The highest elevation, 3,166 feet, occurs on Agrihan, in the northern Marianas.

Coral platforms built on submerged peaks form an atoll's base. On these platforms, shell, broken coral, and floating debris build up narrow islands with the open ocean on one side and the more or less sheltered waters of a lagoon on the inside. Around this lagoon, which may cover more than a hundred square miles, as in Ulithi, many small islands dot the reef, with few or no deep passages connecting open sea and lagoon.

Climate, soil, and vegetation. The climate of the entire area is generally warm and humid. Daily or seasonal changes in temperature are slight, ranging usually from about 75° to 85° F (24° to 29° C). Rainfall on larger volcanic islands may reach 300 to 400 inches a year, but on atolls in the northern Marshalls annual rainfall may be as low as 20 to 30 inches, with prolonged periods of drought.

The soil of coral islands is sandy and lacking in important minerals, so that only limited and specialized vegetation can grow. Wind and sea erosion are a constant threat to atoll soils, and major typhoons have been known to completely wash over even sizable islands of this type, extinguishing both plant and human life. Smaller islets have sometimes been simply washed away.

Along the coral-sand beaches, coconut palms are the most visible and important form of vegetation, although pandanus (the screw pine) is also found. On less favoured islands these trees, together with a few ground creepers, may be the only land vegetation. On larger islands breadfruit and even bananas grow, as well as taro, a starchy root with a buried stem; the marsh taro (*Cyrtosperma chamissonis*), an edible member of the arum family, is grown in swampy areas or man-made compost pits.

High islands frequently have steep, heavily forested areas in the interior, as in Ponape or Babelthuap. Heavy rainfall leaches minerals from the volcanic soil and causes severe erosion when the vegetative cover is destroyed. Excessive timber removal and slash-and-burn farming have left badly eroded areas. Soils are generally shallow and low in fertility.

In the northern Marianas, where rainfall is less abundant, there occur level or gently sloping areas suitable for grazing. Areas resembling savanna (grassy parklands) also appear, intermingled with dense forests where local differences in soil or rainfall occur. Comparatively little use is made of the interiors of the steep volcanic islands, and settlement and economic activity cling close to the shore. Along the ocean fringe, steep, rocky cliffs, sandy beaches, or swampy areas overgrown with mangrove thickets may occur. Here, with ample rainfall, coconut, breadfruit, and pandanus thrive, together with a wider variety of native and introduced food crops than is found on the atolls.

Animal life. The original animal life on the islands was usually extremely limited, probably being confined to one or two species of bats. Pigs, small dogs, and a species of rat were introduced before the European era. Begin-

ning with the Spaniards in the 17th century, other varieties of rats, some fowl, and the water buffalo were introduced, followed by horses, cattle, goats, and cats. Deer were introduced into the Marianas by the Germans in the early 20th century and were later carried to Palau and Ponape. The larger animals were able to thrive only on the larger, high islands.

Marine and shore birds are more common throughout the trust territory than land birds. There are local variations in prevailing species. The albatross, frigate bird, shearwater, tern, and other birds of passage are widespread. Parrots, cockatoos, owls, white-eyes, finches, and other land birds are limited chiefly to the larger volcanic islands in the west. Two species of crocodile and two species of poisonous sea snakes are found in the Palau group. Other harmless snakes and lizards are fairly widespread. More than 7,000 species of insects are known.

Life in the lagoons

In contrast to the relatively few kinds of land animals, life on the shores and in the lagoons is varied and rich. Many types of reef fish, shellfish, and edible algae are found in the shallow waters of lagoons or fringing reefs. Deeper waters between islands or atoll groups offer tuna, barracuda, sharks, and other large species.

The landscape under human settlement. Variations in life-style make it impossible to speak of a typical Micronesian village or pattern of life. The simple, scattered, generally rectangular shelters of remote atolls still house Marshallese and eastern and central Carolinians as they have for centuries. Yet each area has distinctive patterns of design. Roofs and walls of traditional structures are thatched or plaited with pandanus, coconut-palm, or nipa-palm leaves, with supporting timbers of whatever wood is available. Yet, even in remote villages, modern additions in the form of boards, corrugated sheet metal, or concrete blocks have begun to come into use. Houses are built on the ground or occasionally are elevated on stone platforms, pillars, or wooden posts. The elevated houses have bamboo or board floors.

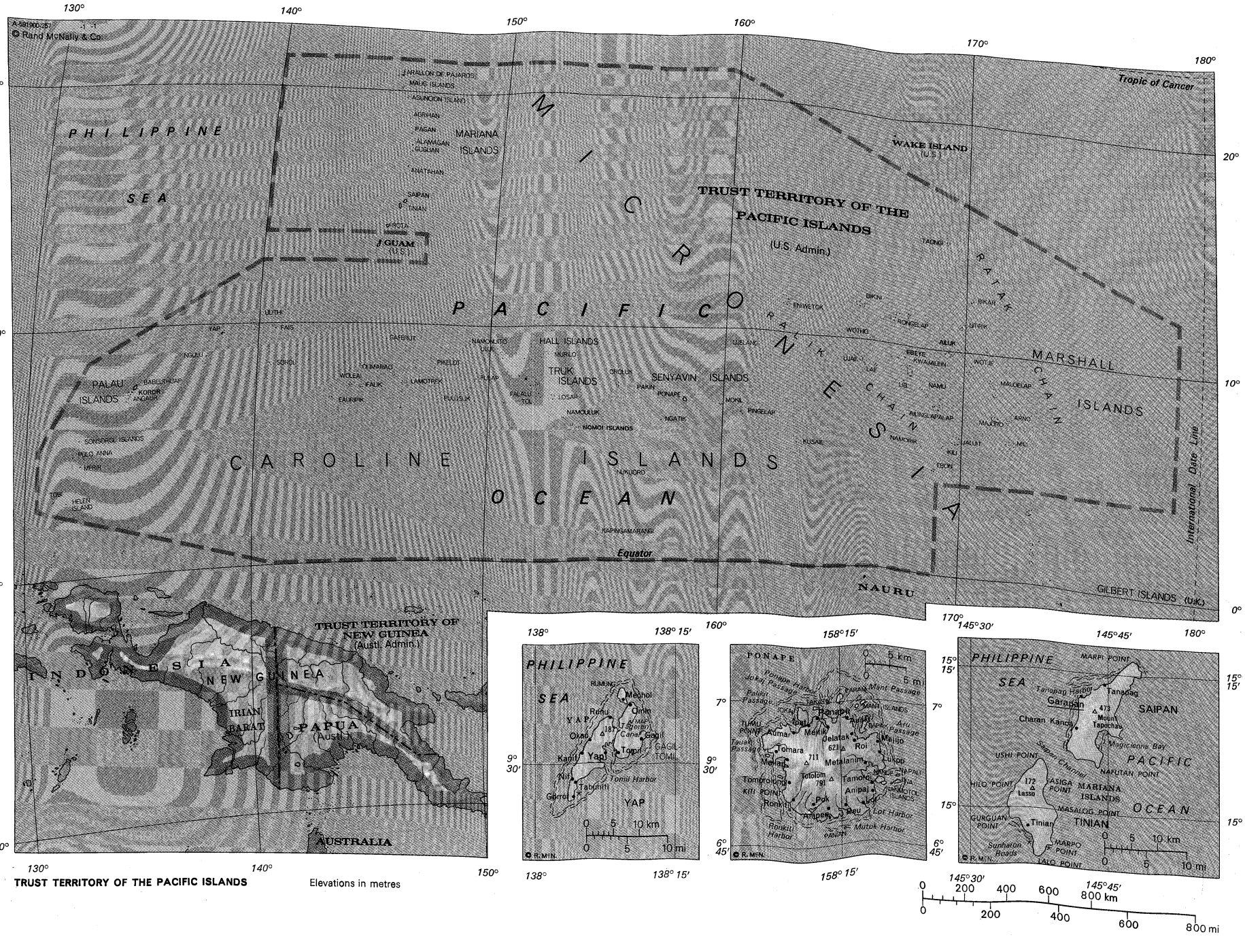
Individual dwellings are scattered among the trees near the shores of the islands, joined together by footpaths made of sand or smooth stones. When the buildings themselves are not unsightly shacks of rusting, salvaged materials, the villages thus loosely clustered make an attractive sight. In high islands where extensive mangrove swamps occur along the shore, as in Ponape, villages are usually farther back from the water's edge, with access to the sea being gained by means of water passages at the mouths of streams. In the small communities most influenced by outside contact and commerce, more tightly clustered groups of buildings along a more definitely marked main street or road have appeared.

In most villages, often in a central location, is a larger structure, which is the dwelling of the chief, a communal meetinghouse, or men's clubhouse. Schoolhouses, clinics, administrative housing, or community centres of modern construction may constitute the largest, if not the most beautiful, buildings. Substantial wooden or concrete buildings of Japanese construction may still be found, particularly in the Marianas and western Carolines. Concrete construction has become increasingly common during the recent period of American rule, particularly where exposure to devastating typhoons has made lighter construction appear impractical.

Major settlements

The principal towns of the trust territory are the district centres, which form the main focusses of foreign influence. Some of them, notably Koror in the Palau group and Truk in the Carolines, were much larger communities in the Japanese period and still have an air of decline, despite recent building.

The district-centre towns are usually built of wood or of corrugated iron. Power and water supplies are found only in the district centres and, even there, are often unreliable. Capitol Hill on Saipan, the administrative centre of the territory, is exceptionally efficient in this respect. Among the largest of the district centres are Majuro, an unplanned and struggling town on a long coral island in the Marshalls, and Koror, on a green volcanic island in the western Carolines. Their populations each numbered about 6,000 in 1970. Residents have been attracted to



TRUST TERRITORY OF THE PACIFIC ISLANDS Elevations in metres

these communities by the concentration of administrative- and commercial-employment opportunities there.

A special type of community has developed at Kwajalein, the American missile-testing base, where a replica of a typical town in the southwestern United States has been built for American personnel. Another type of town has been allowed to develop on the tiny nearby island of Ebeye, where Micronesians who also work on Kwajalein are huddled into crowded quarters quite unlike anything to be found in their native islands.

Life in the more remote villages continues to revolve about subsistence economic activities: the cultivating of small garden plots, the gathering of seafood on the reef, or fishing there or in the sea beyond. The traditional economy is supplemented by the gathering and preparing of copra for sale through a government-controlled system of cooperatives. In the larger towns and district centres, the administration itself is a major contributor to economic life; there, however, modern education and the availability of comparatively highly paid positions place new strains upon traditional patterns of authority.

People and population. *Ethnic groups.* The inhabitants of the trust territory belong to an ethnic stock generally known as Micronesian, with the exception of the Polynesians of Kapingamarangi and Nukuoro. Micronesians are usually described as of Malayan origin and are characterized by medium stature, brown skin, and straight or wavy hair. Mongoloid features are more common in the western part of the territory, suggesting a degree of mixture with other people to the west.

Perhaps the most distinctive ethnic group among the Micronesians consists of the Chamorro of the Marianas. They represent a mixture of the aboriginal inhabitants and a variety of European and Asian groups with whom they have intermarried over a period of some four centuries. The term Micronesian, however, refers to a variety of differing cultures, separately influenced by geographic isolation and—in recent times—by contact with foreigners.

Linguistic groups. Nine major languages exist among the people of the territory, though all of them belong to the Malayo-Polynesian family. Within the individual language groups there are variations in dialects that may result in mutual unintelligibility at first contact. Two of the languages, Chamorro and Palauan, have been described as Indonesian in type. Yapese, Ulithi, Trukese, Ponapean, Kusaiean, and Marshallese are classed as Micronesian languages. A type of Polynesian is spoken by the inhabitants of Kapingamarangi and Nukuoro.

Religious groups. The original religions of Micronesia were not marked by a highly developed set of institutions or by priestcraft. Religion was a pervasive part of the way of life, however, and religious and political power were and still are often ceremonially combined in a chief or shaman.

Christianity was introduced early to the Marianas by the Spanish. Widespread mission work began only in the second half of the 19th century, however, when Spanish Catholics vied with American and, later, German Protestants in an effort that eventually influenced most of the people of the territory. Today, the missions play a vital role in the educational system. Christian services are held regularly in all the major island groups.

Oriental farmers

Demography. During the 1920s thousands of Japanese and Korean workers were brought into the area to provide labour for intensive agriculture, as well as for phosphate mining on the island of Angaur and for commercial fishing. The Marianas and western Carolines were the islands mainly affected. There the Micronesians were often displaced from their lands and resettled by order or by choice on the fringes of the Japanese communities. After World War II nearly all the Japanese and Koreans were repatriated, creating serious economic and social dislocations among Micronesians who had come to rely on the Japanese-built economy and who had intermarried with Japanese. A strong trace of Japanese influence on life patterns remains.

Though the total population was drastically reduced by repatriation, the number of Micronesians more than doubled between 1945 and 1970, rising from an estimated 48,000 to about 100,000. This pattern of growth continues. The 1967 United Nations Visiting Mission reported the population density of the trust territory as 130 persons per square mile, and by 1970 this figure had risen to 140. Yet there are places where the land available is sparsely settled and little cultivated, while district centres and the atolls of the Truk group in general face problems of population pressure upon the resources available.

Trust Territory of the Pacific Islands (Carolines, Marshalls, Marianas), Area and Population

	area		population	
	sq mi	sq km	1967 census	1971 estimate
Districts				
Mariana	184	475	11,000	13,000
Marshall	69	181	19,000	23,000
Palau	179	465	11,000	13,000
Ponape	176	455	18,000	21,000
Truk	46	118	25,000	29,000
Yap	46	119	7,000	7,000
Total Trust Territory of the Pacific Islands	700	1,813	91,000	107,000*

*Figures do not add to total given because of rounding.
Source: Official government figures.

One reason for the uneven distribution of population lies in the greater economic opportunities available at the administrative centres. Another may be found in the unevenness of communications and transportation. Many outlying islands have been visited only rarely by supply ships bringing medical aid. Schools, particularly at the secondary level, also draw residents to the islands on which they are located, sometimes from hundreds of miles away. The outlook for a redistribution of population that would be in closer harmony with the subsistence base is clouded by these factors as well as by landholding traditions and strong ties of association with home islands.

In most of the islands, the introduction of modern medicine has been accompanied by a rising rate of population growth. While this trend points to possible future overcrowding on islands with limited soil and resources, the danger point does not seem to lie in the immediate future, except in some islands in the Truk and Marshall districts. Pre-European population levels in most of the islands were greater than they are today. Ponape, for example, with 20,093 inhabitants in 1969, had less than half the population estimated to have existed two centuries ago. The northern Marianas, the large Palauan island of Babelthuap, and Yap have abundant evidence of settlements in areas now left untilled.

Given the greater productive power resulting from modern scientific methods, the islands in general can support a growing population of indigenous peoples for some time, though at a limited standard of living.

The economy. The traditional subsistence economy of the Micronesians was varied by interisland trading traditions. Production of copra for sale to Western traders was introduced in the late 19th century and continues to be a mainstay of commerce. From the proceeds of copra sales, islanders in the remotest atolls have purchased supplies of Western cloth, kerosene, metal implements, and fishhooks and even tinned food to supplement the regular diet of fish and island vegetable products.

The copra trade

Attempts are being made to diversify the economy and to reduce the dependence upon copra. Cacao and pepper have been introduced experimentally in Ponape with some success. Grazing has been expanded and improved in the northern Marianas, but the number of livestock is still below that kept by the Japanese in the 1930s. The high cost of imported feed and fertilizer forms a barrier to development of dairying, large-scale poultry production, or diversified commercial farming. Sugar production was practiced by the Japanese on lands now largely untilled, but imported labour was required.

If good agricultural soil and minerals are scarce in the trust territory, there are both abundance and variety of

marine resources. Fishing is still chiefly a subsistence activity for the Micronesians, and they lack the technological equipment for a modern commercial fishing industry. Young men are being trained in Japanese and Hawaiian methods, and boats are being built; but fuel will still have to be imported for modern vessels and for machinery to be used in canning or refrigerating plants. Japanese competition and American tariff policies offer additional obstacles in this field.

Some income has been secured from sales of Micronesian handicrafts. Here again, potential expansion is hampered by the amount of time and hand labour required. Another supplement to the economy has been the salvaging of scrap metal left from World War II activities, carried on chiefly by Japanese teams under contract; by the 1970s, however, the richest sources had become exhausted.

One of the most promising economic possibilities for Micronesia appears to lie in tourism. The beauty of the islands, the year-round mildness of the climate, and the exotic nature of island cultures all offer attractions to visitors. Improved transportation and accommodations are being developed in parts of the territory. In parts of the Marshalls and in conservative Yap, however, resistance to the growth of tourism and its inevitable intrusion into traditional ways of life appears likely to limit the industry.

By far the largest economic enterprise in the trust territory is the administration itself. Salaries paid to Micronesians and Americans in the employ of the government compose a major part of the money economy.

Taxation

Levels of taxation are generally low, being hampered by the difficulty of levying property taxes where land titles are still so uncertain. Absence of adequate tax revenues is viewed by administration officials as an obstacle to early self-government. For the immediate future it appears clear that economic support from outside will be needed to maintain the economy. Some young Micronesian leaders, however, feel that most of the money now used to support the administration goes to a bureaucracy of American and Micronesians that would not be needed under self-government. They foresee a much simpler political structure that Micronesians could support with only minimal foreign aid.

Transportation. The vast open ocean spaces between islands make transportation and communication a major problem. Canoes are inadequate to meet the freight needs of today, and the Micronesians generally lack the capital to purchase modern motor ships or the skills required to operate them. Fuel, also, must be imported.

The administration operates motor ships in each district to provide interisland freight and passenger service. Two additional ships are operated on a territory-wide basis to meet educational, medical, and community needs. Private local concerns handle lading and ship-agency work at district ports. On outlying islands the whole population may turn out to help bring imports in over the reef or through difficult waters.

In 1968, Micronesian Interocean Line, Inc., began operating between the trust territory and ports in the United States and the Far East, on a ten-year contract with the government. It is hoped that Micronesian labour and capital will play an increasingly important role in this enterprise and in intraterritory shipping.

Air Micronesia, a combination of the United Micronesia Development Association (which owns a 40 percent interest) and two American airlines, began regular air-transport service in 1968. Airfields at the six district centres can handle jet aircraft, and plans have been made to build hotels nearby for American and Japanese tourists. Charter airlines carry some local freight and passenger traffic among Guam, Saipan, and major centres in the Carolines.

Land transportation throughout the territory is limited in extent and is hampered by poor roads outside the main population centres. Heavy rainfall and sea erosion make road maintenance difficult and costly, and it is seldom possible to maintain roads circling larger islands to connect their shoreline settlements.

Bus services exist in the largest towns, extending to a few outlying communities. By the 1970s, other vehicles, including trucks and motor scooters, numbered about 7,000. Radiotelephone and teletype stations connect the district centres, and some form of radio contact is maintained with the outlying populated islands. A telephone link from Saipan to Guam and thence to the outside world was completed in 1970. Ship-to-shore communications and navigational aids are maintained at the district centres.

Administration and social conditions. In accordance with the provisions of the 1947 United Nations Trusteeship Agreement, the legal basis for government within the area was laid down in the Code of the Trust Territory, enacted December 22, 1952. Later amended, the code defined citizenship, provided a formal law code, and created six administrative districts, centred at Saipan (northern Marianas), Moen (Truk), Colonia (Yap), Koror (western Carolines), Kolonia (on Ponape in the eastern Carolines), and Majuro (Marshalls). The Code recognizes local customary law, but in many cases, especially on land tenure, formal codification of customary law has not been completed. The administrative officials of the trust territory are appointed, not elected. A high commissioner is appointed by the president of the United States, and he, in turn, appoints the six district administrators, their deputies, and the heads of administrative departments at Saipan, under authority from the secretary of the interior. The judicial system consists of a High Court, six district courts, and community courts throughout the islands. Only the High Court justices, appointed by the secretary of the interior, are Americans. The law applied by the courts is a mixture of United States and Trust Territory codes, together with local customary law.

During the 1950s criticism of administration policies in the trust territory was expressed by Micronesians, in the Trusteeship Council of the United Nations, and within the United States. Specific complaints concerned inadequate budgets for education and health services, the slow pace of economic development, and an apparent absence of American plans for future self-government. Following this criticism, trust-territory budgets were markedly increased, particularly after 1960.

Early in the 1960s, an advisory Council of Micronesia was created as a step toward representation of the inhabitants of the trust territory, and in 1965 this body was succeeded by an elected Congress of Micronesia with legislative powers. These powers are subject to approval by the high commissioner or, on appeal, by the secretary of the interior of the United States. The 12 senators and 24 representatives of the Congress are elected in even-numbered years. The former serve four-year terms and the latter, two. There are two senators for each district, while the representatives are apportioned on the basis of population. Six district legislatures are also elected, each under its own charter. Agitation for increased self-government grew among Micronesian leaders and subsequently found support both in the United Nations and in the United States.

Medical care in the trust territory is entirely a public service. There are hospitals at all six district centres, three sub-district hospitals, and rural dispensaries. Doctors and nurses visit the remoter islands on field trip ships, and Peace Corps physicians and medical aides supplement their work. Cases requiring special treatment are still referred to Guam or Honolulu hospitals, though a larger, new hospital with modern equipment opened at Truk in 1971. Micronesians have been trained as nurses, technicians, and medical practitioners, but not yet as qualified physicians. Micronesian students are sent outside, mostly in Guam, Hawaii, and the mainland United States. By 1971 there were public high schools in all district centres and five additional ones on outer islands. At Ponape a community college offers training, primarily to teachers.

Prospects for the future. In 1968 Pres. Lyndon B. Johnson suggested 1972 as a possible date for a plebiscite on the future political status of the trust territory. In 1970 the Congress of Micronesia sent a committee to the United States and to various self-governing island groups to

Movements toward self-government

examine other systems. A resulting report led the Congress to vote in favour of self-government in free association with the United States. When the U.S. Department of the Interior offered the Micronesians a commonwealth status under United States supervision, members of their Congress voted to reject the offer and to seek either "free association" status or complete independence from the United States. The question remains unsettled, subject to further negotiations.

In contemplating the future, U.S. administrators see in the existence of different traditional systems an obstacle to unified government. Micronesian leaders, on the other hand, tend to minimize this difficulty and to place a higher value on maintenance of local custom than on foreign tutelary dominance. The problem of balancing the need for unity in administration against cultural diversity is also marked in the field of language. Micronesian sentiment appears divided between the need for a modern common language of administration on the one hand and pride in local dialects on the other. Leaders with modern education appear more attuned to change than are tradition-oriented Micronesians, and they are also more determined to secure self-government on a unified basis.

An independent Micronesia would require economic support from abroad until a viable economy combining the traditional and modern sectors could be developed. Military protection from without also appears likely to be needed, for the trust territory still occupies an area that is of potential strategic importance in the relations between major powers. If neutralization through international agreement cannot be arranged, then some form of major-power protectorate appears certain.

BIBLIOGRAPHY. ELY J. KAHN, JR., *A Reporter in Micronesia* (1966), popular account with keen insights on policies and problems; LEONARD MASON, "The Ethnology of Micronesia," in ANDREW P. VAYDA (ed.), *Peoples and Cultures of the Pacific* (1968), a sound introductory survey; NORMAN MELLER, *The Congress of Micronesia* (1969), a scholarly treatment of the subject by a first-hand observer; *Official Visitor's Guidebook to the Trust Territory of the Pacific Islands* (1969), well-written tourist guide; UNITED NATIONS, TRUSTEESHIP COUNCIL, VISITING MISSION TO THE TRUST TERRITORY OF THE PACIFIC ISLANDS, *Report* (1950–), detailed reports submitted every two or three years; U.S. DEPARTMENT OF THE INTERIOR, *Trust Territory of the Pacific Islands* (annual), official reports to the United Nations on the administration of the Trust Territory of the Pacific Islands; CONGRESS OF MICRONESIA, *Report of the Political Status Delegation of the Congress of Micronesia* (1970), important statement of views on future political status.

(D.D.J.)

Pacific Ocean

Of the three oceans that extend northward from the Antarctic continent, the Pacific, occupying about a third of the surface of the terrestrial globe, is by far the largest. Its area, excluding adjacent seas, encompasses about 64,000,000 square miles (166,000,000 square kilometres). It has double the area and more than double the water volume of the Atlantic—the next largest division of the hydrosphere. Its area exceeds that of the whole land surface of the globe, Antarctica included, with Africa counted twice. The Pacific stretches from the shores of Antarctica to the Bering Strait through 135° of latitude, or for 8,350 nautical miles (9,600 miles or 15,500 kilometres). Its greatest longitudinal extent measures 11,500 nautical miles along the parallel of 5° N, between the coasts of Colombia in South America and the Malay Peninsula in Asia. The mean depth of the Pacific (excluding adjacent seas) is 14,050 feet (4,280 metres). Its greatest known depth is 36,198 feet or 11,033 metres (in the Mariana Trench).

Extent of the Pacific

The Pacific and Arctic systems mingle their waters in the Northern Hemisphere at the shallow Bering Strait—whose width is a mere 55 nautical miles—and in the Southern Hemisphere the Pacific and Atlantic mix in the relatively narrow Drake Passage between Tierra del Fuego in South America and Graham Land in Antarctica. The separation between the Pacific and Indian oceans is less distinct; in this article, it is considered to lie along the line of islands extending eastward from Sumatra, through Java to Timor, thence across the Timor Sea to Cape Londonderry in Australia. To the south of Australia the boundary extends across the Bass Strait and thence from Tasmania to Antarctica.

Because of the pattern of major mountain systems of the globe, a relatively small proportion (one-seventh) of the total continental drainage enters the Pacific—*i.e.*, a total drainage area of not more than about three times that of Australia. Of the rivers that drain into the Pacific, those of China and Southeast Asia are of the greatest importance; the basins of these rivers support more than one-quarter of the world's population.

The eastern boundary of the Pacific is associated with the American Cordilleran mountain system, which stretches from Alaska in the north to Tierra del Fuego in the south. Except for its extreme northern and southern sections, which are characterized by fjords (narrow arms of the sea bordered by steep cliffs) and their numerous off-lying islands, and except for the deeply indented Gulf of California, the coastal boundary is relatively regular and the continental shelf narrow. The western, or Asiatic, coastal boundary, in contrast, is irregular. Although the mountain systems there lie roughly parallel to the coast, as they do on the eastern Pacific coastlands, the western Pacific is noted for its many peripheral seas. From north to south they include the Bering Sea, Sea of Okhotsk, the Sea of Japan, the Yellow Sea, the East China Sea, and the South China Sea. Their eastern boundaries are formed by southward-jutting peninsulas or by island arcs or both. It is of oceanographic significance that the great rivers of eastern Asia—including the Amur, the Yellow River, the Yangtze, the Hsi Chiang, and the Mekong—enter the Pacific indirectly by way of peripheral seas. (For associated physical features, see BERING SEA AND STRAIT; BIKINI; CHINA SEA; DRAKE PASSAGE; EASTER ISLAND; GALAPAGOS ISLANDS; JAPAN, SEA OF; PACIFIC ISLANDS; and YELLOW SEA; for historical background see OCEANIA, HISTORY OF; see also CONTINENTAL DRIFT; CONTINENTAL SHELF AND SLOPE; OCEAN BASINS; OCEAN CURRENTS; OCEANIC RIDGES; OCEANS AND SEAS; and OCEANS, DEVELOPMENT OF.)

THE NATURAL ENVIRONMENT

Relief. The Pacific Basin may conveniently be divided into three major physiographic regions, the eastern, western, and central Pacific regions.

Eastern region. The eastern region, which extends southward from Alaska to Tierra del Fuego, is relatively narrow and is associated with the American Cordilleran system of almost unbroken mountain chains, the coastal ranges of which rise steeply from the western American shores. The continental shelf, which runs parallel to it, is steep and comparatively narrow. Significant oceanic trenches in this region are the Acapulco and Guatemala trenches in the North Pacific and the Peru–Chile Trench in the South Pacific.

Oceanic trenches

Western region. The second physiographic province is the western region, the seaward boundary of which is marked by a broken line of oceanic trenches, extending from the Aleutian Trench in the north through the Kuril and Japan trenches and southward to the Tonga and Kermadec trenches, terminating close to the southeast of North Island, New Zealand. The western region has a structure more complex than that of the eastern region. Characteristically associated with the ocean trenches of the western region are festoons either of peninsulas or islands or both. The islands, which include those of Japan as well as numerous small islands, represent the upper parts of mountain systems that rise abruptly from the deep ocean floor. The island clusters of the western Pacific form the boundaries of the several wide and deep continental seas of the region.

Central Pacific region. The third province is the central Pacific region lying between the boundaries of the eastern and western regions. The largest and the most geologically stable of the structural provinces of the earth's crust, it is characterized by expansive areas of low

relief, lying at a general depth of about 15,000 feet below the surface.

Principal ridges and basins. To the east of the meridian of 150° W the relief of the ocean floor is considerably less pronounced than it is to the west. In the eastern Pacific the Cocos Ridge extends southwestward from the Central American isthmus to the Galápagos Islands. To the southwest of the Galápagos lies the Southeast Pacific Basin, which is separated by the extensive Southeastern Pacific Plateau from the Pacific-Antarctic Basin, which, in turn, is separated from the Southwest Pacific Basin by the indeterminate Pacific-Antarctic Ridge, which runs from the Southeastern Pacific Plateau to Antarctica in the vicinity of 150° W.

Extending southward from the Tasman Basin (between New Zealand and East Australia) is the Macquarie Ridge, which forms a significant boundary between the deep waters of the Pacific and Indian oceans. The Hawaiian Ridge extends westward from Hawaii to the meridian of 180°.

The submerged parts of the series of ridges that are capped by the island archipelagoes of the western Pacific are continuous and are to be found at depths of less than about 2,000 feet. These ridges include the Aleutian ridge in the northwest Pacific; the series of ridges extending southward through the Kurils, Bonin, Marianas, Yap, and Patou; those extending eastward from Patou including the Bismarck, Solomons, and Santa Cruz ridges; and finally the ridges extending southward from which rise the Samoan islands, Tonga, Kermadec, Chatham, and Macquarie.

Islands. The islands of the western region—including the Aleutians, Kurils, Ryukyus, Taiwan, the Philippines, Indonesia, New Guinea, and New Zealand—are continental in character. Geologically they consist, in part, of sedimentary rocks, and their structures are not dissimilar to those of the coastal mountain ranges of the adjacent continent. Most are sufficiently large for oceanic climatic influences not to be pervasive.

"High" islands and "low" islands

A geologically important boundary between the continental, or "high" islands, and the numerous truly oceanic, or "low" islands, of the Pacific is the Andesite Line. (Andesite is a fine-grained rock typical of the geologic boundary line separating the basic igneous rock of the ocean basins from the acid igneous rock characteristic of continental areas.) In the north and west Pacific the Andesite Line follows close to seaward the trend of the island arcs from the Aleutians southward to the Yap and Patou arcs, thence eastward through the Bismarcks, Solomons, and Santa Cruz, and thence southward through Samoa, Tonga, Chatham, and Macquarie to Antarctica. Within that boundary, islands are essentially of basalt, a basic igneous rock characteristic of ocean basins; basalt contrasts with the more acid igneous rocks associated with volcanic activity in continental areas.

The numerous "oceanic" islands of the Pacific are unevenly distributed. They lie, in the main, between the Tropics of Cancer and of Capricorn and occur in great numbers in the western Pacific. The northernmost chain of oceanic islands is associated with the Hawaiian Ridge. The Hawaiian archipelago consists of about 2,000 islands, although the term Hawaiian Islands is usually applied to the small group that lies at the eastern end of the archipelago.

The numerous small islands of Micronesia lie mainly north of the Equator and to the west of the 180th meridian. Nearly all are coralline; the principal groups are the Marianas, the Marshalls, the Carolines, and the Gilbert and Ellice Islands.

To the south of Micronesia lie the islands of Melanesia, most of which are small coral islands. The region's physiography is dominated by a group of large continental islands, however, including New Guinea. The principal groups of the Melanesian islands are the Bismarck Archipelago, the Solomons, the New Hebrides, New Caledonia, and Fiji.

The immense triangular area, with Hawaii in the north, Easter Island in the southeast, and New Zealand in the southwest is peppered with the multitude of islands known as Polynesia. The main Polynesian groups are the Phoenix Islands, Samoa, Tonga, the Cook Islands, Society Islands, Tuamotu, and the Marquesas.

Geology. Corroboratory evidence drawn from various geophysical fields—seismology, vulcanology, gravimetry (the measurement of weight or density), and paleomagnetism (the study of the traces of magnetic polarization of rocks from previous eras)—points to the general validity of the theory of continental drift (see CONTINENTAL DRIFT).

Origin of island arcs

The Pacific island arcs, or epicontinental (*i.e.*, near-continental) islands as they are sometimes called, are believed to have originated in the lateral thrusting of the Asian continental crust toward and over the Pacific floor, the shapes of the island arcs corresponding to the intersections of the thrust planes with the spherical earth's surface. The intense folding and faulting along the volcanic zone of the western Pacific provide unmistakable evidence of orogenic (mountain-building) forces at work. The deep basins that lie between the Asian continent and the island arcs were evidently caused by the local downfolding of the earth's crust; the island arcs themselves, along the line of which volcanic and seismic activity are pronounced, are the result of weakening of crustal strata by strong upfolding. On the oceanic side of the island arcs, intense downfolding has resulted in the formation of the deep trenches that fringe the line of the island arcs.

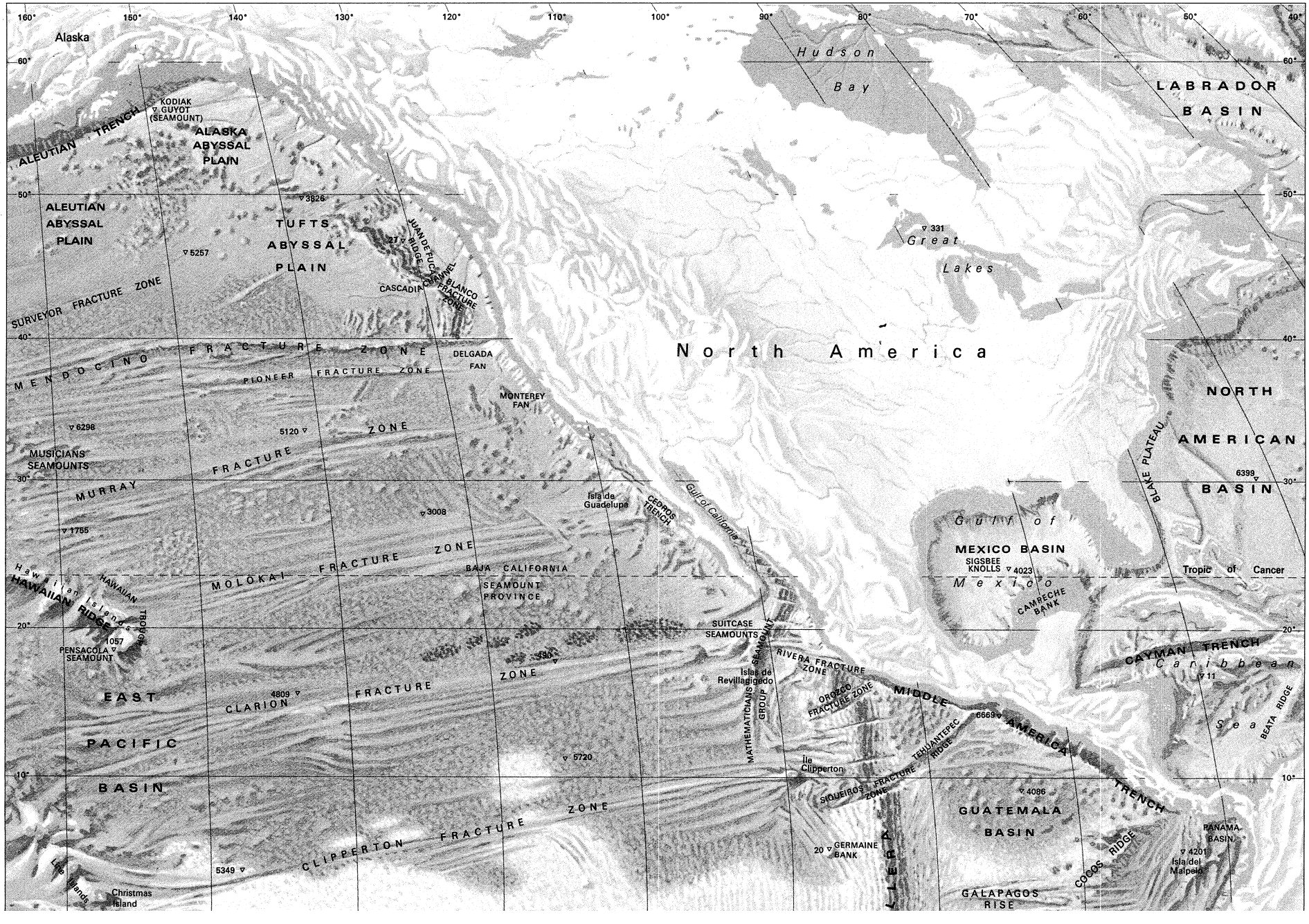
The character of the eastern marginal physiographic region of the Pacific suggests a lateral movement of the North American and South American continents westward over the Pacific floor. The floor of the northeastern Pacific is remarkable for its several major fracture zones, which extend east and west and which, in some instances, are identifiable over distances of thousands of miles.

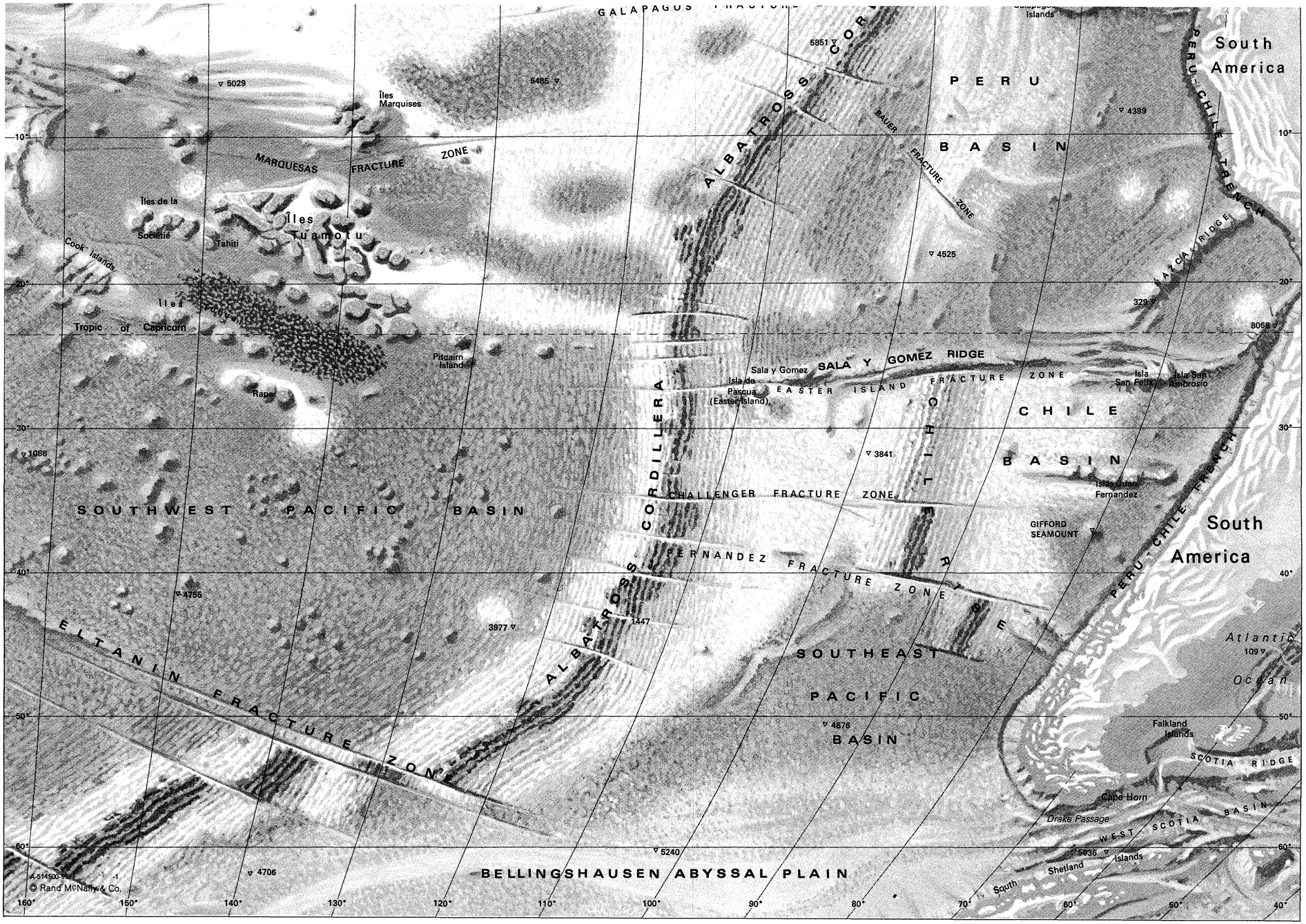
Of great geological interest are the seamounts (called guyots) and the oceanic islands of the Pacific. The numerous tropical islands of the Pacific are mainly coralline. The three principal types of coral reef—fringing, barrier, and atoll, as well as the flat-topped guyots, which are abundant on the ocean floor in extratropical latitudes within the Pacific—are explained by the slow subsidence theory advanced by the English naturalist Charles Darwin during the 19th century. (According to this theory, the ocean floor in the vicinity of a submarine volcano or volcanic island sinks slowly in response to gravitational and other forces; coral reefs are assumed to grow upward and outward on tropical islands that are submerging, taking the form first of a fringing reef, then of a barrier reef, then of an atoll. Outside the tropics, where reef-building corals are not found, the tops of oceanic volcanic islands would be expected ultimately to be truncated by the erosive action of sea waves. The submergence of the island would then give rise to a guyot.)

Climate. The wind and pressure systems of the Pacific conform closely with the so-called planetary system—*i.e.*, the pattern of air pressure and the consequent wind pattern that develops when the earth's surface (land or sea) is smooth. This conformity is partly a consequence of the great extent of the Pacific and the comparative uniformity of its surface. Climatic conditions in the South and East Pacific, where the steadiness of the trade winds and the westerlies is remarkable, are the most uniform on the globe. In the North Pacific, however, conditions are not so uniform, particularly the considerable climatic differences between the eastern and western regions in the same latitude. The rigour of the winters off the east coast of the Soviet Union, for instance, contrasts sharply with the relative mildness of winters in the region of British Columbia.

Uniformity of climate

The trade winds. The trade winds of the Pacific represent the eastern and equatorial parts of the air-circulation system; they move around the subtropical high-pressure zones centred, respectively, over the northeast and southeast Pacific between the 30th and 40th parallels N and S. The obliquity of the ecliptic (the term used to denote the angle, approximately 23½°, contained between the planes of the earth's rotation on its axis and its revolution around the sun) limits the seasonal shifting of


Alaska
Hudson Bay
LABRADOR BASIN
KODIAK GUYOT (SEAMOUNT)
ALEUTIAN TRENCH
ALASKA ABYSSAL PLAIN
ALEUTIAN ABYSSAL PLAIN
TUFTS ABYSSAL PLAIN
3826
5257
331
Great Lakes
JUAN DE FUCA RIDGE
CASCADIA
CHANNEL
BLANCO FRACTURE ZONE
SURVEYOR FRACTURE ZONE
MENDOCINO FRACTURE ZONE
PIONEER FRACTURE ZONE
DELGADA FAN
MONTEREY FAN
North America
NORTH AMERICAN BASIN
BLAKE PLATEAU
6298
5120
MUSICIANS SEAMOUNTS
MURRAY FRACTURE ZONE
6399
3008
1755
Isla de Guadelupe
CEDROS TRENCH
Gulf of California
Gulf of Mexico
MEXICO BASIN
SIGSBEE KNOLLS
4023
Tropic of Cancer
CAMPECHE BANK
MOLOKAI FRACTURE ZONE
BAJA CALIFORNIA SEAMOUNT PROVINCE
Hawaiian Islands
HAWAIIAN RIDGE
Tropic
SUITCASE SEAMOUNTS
1057
PENSACOLA SEAMOUNT
930
CAYMAN TRENCH
Caribbean Sea
711
Islas de Revillagigedo
RIVERA FRACTURE ZONE
MATHEMATICIANS SEAMOUNT GROUP
OROZCO FRACTURE ZONE
MIDDLE AMERICA TRENCH
6669
TEHUANTEPEC RIDGE
BEATA RIDGE
EAST PACIFIC RISE
4809
CLARION FRACTURE ZONE
5720
Ile Clipperton
SIQUEIROS FRACTURE ZONE
4086
GUATEMALA BASIN
PANAMA BASIN
20
GERMAINE BANK
4201
Isla del Malpelo
COCOS RIDGE
5349
CLIPPERTON FRACTURE ZONE
Line Islands
Christmas Island
GALAPAGOS RISE
160°
150°
140°
130°
120°
110°
100°
90°
80°
70°
60°
50°
40°
60°
50°
40°
30°
20°
10°



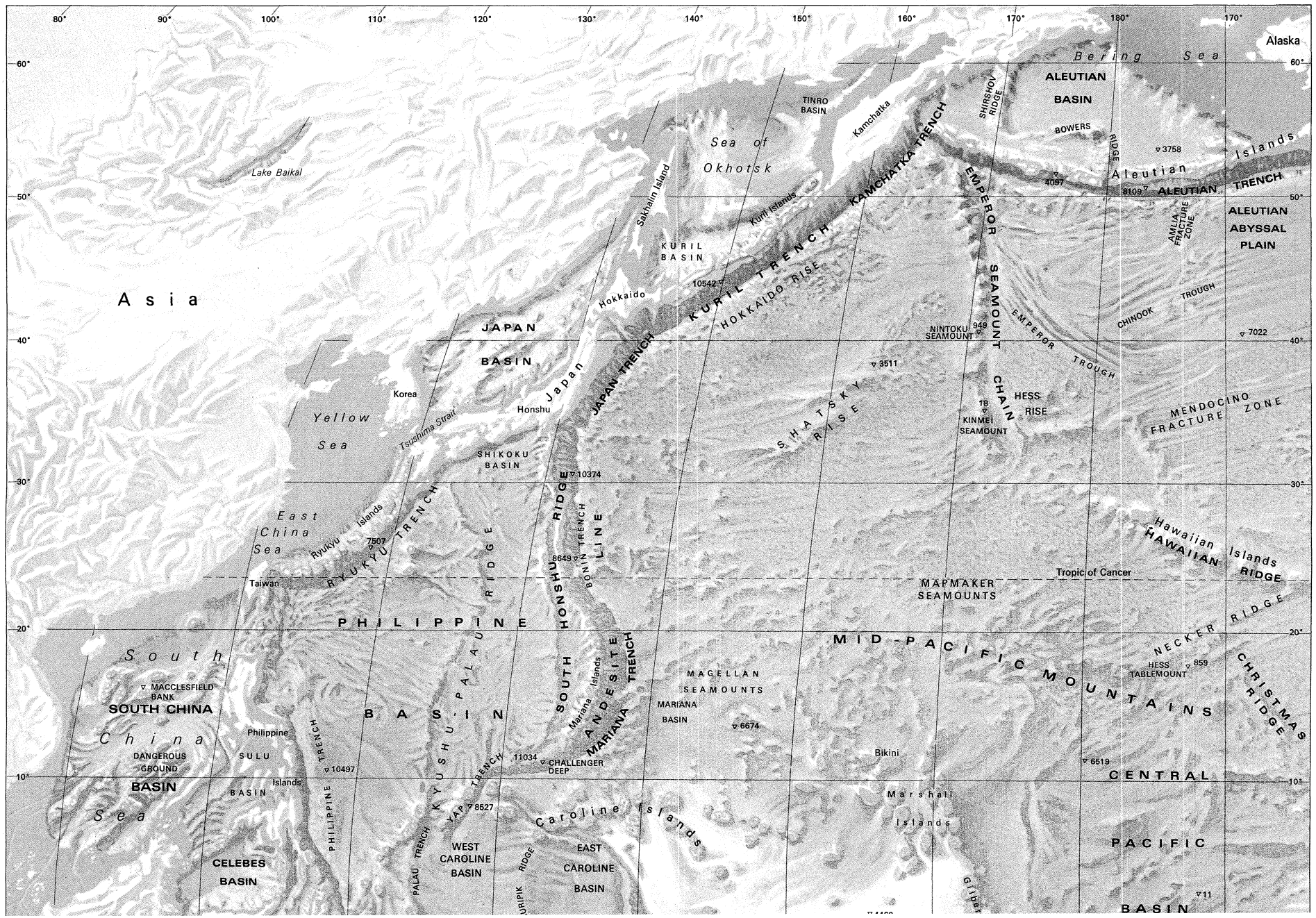
EASTERN PACIFIC OCEAN

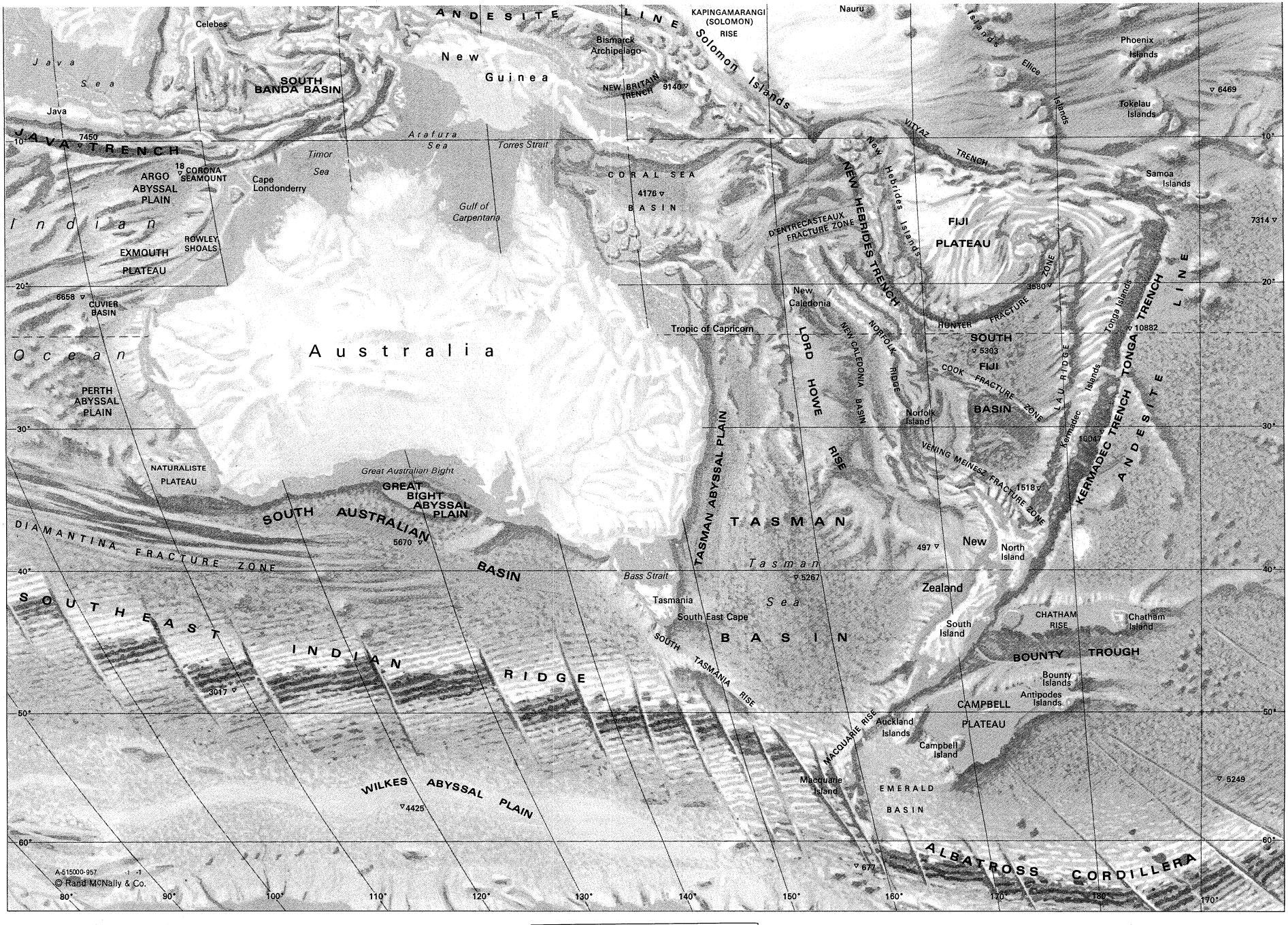
Depths in metres

Colours used are thought to be those of the various rocks and sediments on the sea floors. Differences in relief are shown by relief shading.

0 400 800 1200 km

0 400 800 1200 mi


80°
90°
100°
110°
120°
130°
140°
150°
160°
170°
180°
170°
60°
50°
40°
30°
20°
10°
Alaska
Bering Sea
ALEUTIAN BASIN
TINRO BASIN
Kamchatka
SHIRSHOV RIDGE
BOWERS RIDGE
Sea of Okhotsk
Lake Baikal
Sakhalin Island
Kuril Islands
KAMCHATKA TRENCH
EMPEROR SEAMOUNT CHAIN
Aleutian Islands
ALEUTIAN TRENCH
ALEUTIAN ABYSSAL PLAIN
AMLIA FRACTURE ZONE
KURIL BASIN
KURIL TRENCH
HOKKAIDO RISE
Asia
Hokkaido
JAPAN BASIN
JAPAN TRENCH
NINTOKU SEAMOUNT
CHINOOK TROUGH
EMPEROR TROUGH
SHATSKY RISE
HESS RISE
KINMEI SEAMOUNT
MENDOCINO FRACTURE ZONE
Korea
Honshu
Japan
Yellow Sea
Tsushima Strait
SHIKOKU BASIN
HONSHU RIDGE
BONIN TRENCH
BONIN LINE
East China Sea
Ryukyu Islands
RYUKYU TRENCH
Hawaiian Islands
HAWAIIAN RIDGE
Taiwan
Tropic of Cancer
MAPMAKER SEAMOUNTS
PHILIPPINE BASIN
KYUSHU-PALAU RIDGE
South China Sea
MACCLESFIELD BANK
SOUTH CHINA BASIN
Philippine Islands
DANGEROUS GROUND
SULU BASIN
PHILIPPINE TRENCH
Mariana Islands
ANDESITE LINE
MARIANA TRENCH
MAGELLAN SEAMOUNTS
MARIANA BASIN
CHALLENGER DEEP
MID-PACIFIC MOUNTAINS
NECKER RIDGE
HESS TABLEMOUNT
CHRISTMAS RIDGE
Bikini
Marshall Islands
CENTRAL PACIFIC BASIN
YAP TRENCH
PALAU TRENCH
WEST CAROLINE BASIN
EAST CAROLINE BASIN
Caroline Islands
PALAU RIDGE
CELEBES BASIN
3758
4097
8109
7022
3511
949
18
859
6519
11
10542
10374
7507
8649
11034
10497
8527
6674



WESTERN PACIFIC OCEAN

Depths in metres

Colours used are thought to be those of the various rocks and sediments on the sea floors. Differences in relief are shown by relief shading.

0 400 800 1200 km

0 400 800 1200 mi

the Pacific trade-wind belts to about 5° of latitude. The easterly winds in between the two subtropical zones form the intertropical air flow, which is most pronounced in the eastern Pacific.

The trade winds, especially in the eastern Pacific, convey relatively cool air toward the Equator; in moving, the air comes in contact with the sea and thus becomes increasingly humid and warm, and high lapse rates (the term used to denote the rate of change of air temperature with increasing height above sea or land surface) result. The average wind speed of the Pacific trade winds is about 13 knots (nautical miles per hour). The weather in the trade-wind belts is normally fine, with relatively little cloud; such cloud as there is characteristically takes the form of broken cumulus (small piles of cloud with flat bases) at about 2,000 feet above sea level. Precipitation, usually in the form of light showers, is slight; visibility is generally excellent.

Off the west coasts of the American continent in the trade-wind belts, upwelling of cold, subsurface water results in the overlying air being cooled below its dew point (the air temperature below which vapour condenses as dew), with the consequent formation of widespread and low, thick clouds. Fog in these regions is not uncommon.

The equatorial region, in which the trade winds of the Northern and Southern hemispheres converge, is the region of calms or light variable breezes and is known as the doldrums.

Tropical storms. Although, in general, the climatic conditions of the trade-wind belts are characteristically regular and uniform, storms of great violence do originate there. In such storms, winds of exceptionally great force blow around a centre of exceedingly low sea-level air pressure. These are the tropical storms known in the western Pacific as typhoons. The mechanism that operates to trigger a tropical storm is not understood completely, but the energy required to engender and sustain a storm is undoubtedly related to the enormous quantity of latent heat associated with tropical maritime air.

Typhoons of the western Pacific

Ideal conditions for the development of tropical storms occur in the western Pacific between the parallels of 5° and 25° N during late summer and early autumn. The regions to the east of the Philippines and in the South and East China seas are notorious for these storms, which imperil shipping and often cause severe coastal flooding, accompanied by loss of life and property.

The westerlies. Within the belts of the westerly winds, cold easterly winds from polar regions meet the warm westerly winds of the middle latitudes, causing the formation of the travelling depressions characteristic of middle latitudes. The zone of convergence, or polar front, is most strongly developed in winter when the contrast in temperature and humidity of the air between the converging flows is greatest.

The westerlies in the Southern Hemisphere are steady and strong, and "brave west winds" is an apt name. The gales that accompany the depressions have led to the term roaring forties named for the latitudinal zone in which the storm winds are of greatest frequency.

The monsoon regime. The western Pacific is subject to a seasonal climatic regime, which replaces the planetary system. This is the regime of the monsoon (rain-bearing winds), which is associated with the heating of the Asian landmass in summer and its intense cooling in winter. The heating of tropical Asia in summer initiates a low-pressure system, which becomes the focal point of the trade winds of both hemispheres. The doldrums, therefore, do not exist in the western Pacific during northern summer because of the large-scale flow of maritime air into the Asian low-pressure zone. The cooling of the continent in winter results in the development of the Asian high-pressure system, which leads to a strengthening of the trade winds of the Northern Hemisphere.

As a result of seasonal changes in pressure and wind circulation, marked seasonal contrast between continental and maritime influences—the first associated with drought and cold and the second with moisture and heat—is to be found in the whole of the western Pacific from the Sea of Japan southward.

Temperature and salinity. Temperature and salinity of sea water are the principal determinants of density; and density differences play an important role in relation to the circulation of the ocean.

Density and water circulation

Temperature. The oceans tend to be stratified, with the densest water lying at the bottom and the least dense at the surface. The principal factor in establishing this pattern is temperature; the bottom waters of the deep parts of the ocean are intensely cold, with temperatures only slightly above freezing.

Ocean temperatures in the North Pacific tend to be higher than those experienced in the South Pacific because the ratio of land to sea areas is larger in the Northern Hemisphere than it is in the Southern Hemisphere and because the ice continent of Antarctica also influences water temperature.

The mean position of the thermal Equator (the line on the earth on which the highest average air temperatures are found; the line migrates latitudinally with the changing angular distance from the equator of the sun) in the Pacific, although it lies in the Northern Hemisphere, is nearer to the geographical Equator than in the Atlantic and Indian oceans.

Salinity. The waters within the belt of calms and variable winds near the Equator have lower salinities than those in the trade-wind belts. In the equatorial belt, relatively large amounts of rain fall and little evaporation occurs both because of low windspeeds and because of the generally cloudy skies; salinity in the equatorial belt runs as low as 34 parts per thousand.

The highest surface salinities in the open Pacific occur in the southeastern area, where they reach 37 parts per thousand; in the corresponding trade-wind belt in the North Pacific the maximum salinity seldom reaches 36 parts per thousand. Pacific waters near Antarctica have salinities of less than about 34 parts; the lowest salinities—less than about 32 parts—occur in the extreme northern zone of the Pacific.

The heavy rainfall of the western Pacific, associated with the monsoons of the region, gives rise to relatively low salinities. Seasonal variations there, as well as in the eastern Pacific where seasonal changes in surface currents occur, are significant.

Hydrology. *Surface currents.* Pacific trade winds drive surface waters toward the west to form the North and South Equatorial currents, whose axes coincide, respectively, with the parallel of 15° N and the Equator. Squeezed between the Equatorial currents is a well-defined countercurrent, the axis of which is always north of the Equator and which extends from the Philippines to the shores of Ecuador.

The major part of the North Equatorial Current swings northward in the vicinity of the Philippines to form the warm Kuroshio Current. To the east of Japan the Kuroshio swings eastward to form the Kuroshio Extension. The branching of this current in the region of 160° E results in the movement known as the North Pacific Drift. A branch of the warm Kuroshio Current passes through the Tsushima Strait to form the Tsushima Current, which flows in a northward direction on the eastern side of the Sea of Japan.

The North Equatorial Current

The surface waters of the Bering Sea circulate in an anticlockwise direction. The southward extension of the Kamchatka Current forms the cold Oya-shio, which flows to the east of Honshu to meet the warm Kuroshio waters in the vicinity of 36° N.

The California Current forms the eastern branch of the main circulation system of the North Pacific. Having a great east–west extent, its surface-water movement is very slow.

The South Equatorial Current, on reaching the Solomon Islands, swings southward to form the East Australian Current, which links with the southern line of the main South Pacific circulation system, which, in turn, reaches the South American coast in the region of 45° S. One branch flows northward to form the Peru (Humboldt) Current, and a second branch flows southward to pass through the Drake Passage.

Between January and February—but sometimes as late

as March, or even April—the axis of the Equatorial Countercurrent is displaced southward of its mean position, north of the Equator, so that warm saline waters reaching the coast of Ecuador swing southward to merge with the relatively cold northward-flowing waters of the Peru Current south of the Equator. The warm south-flowing current is called El Niño ("The Child")—a reference to the infant Christ. The strong development of El Niño, which may extend to 14° S, causes destruction (due to changes in temperature and salinity) of vast numbers of marine planktonic life and of fish and other higher marine life-forms that feed on the plankton. Mass starvation of birds that rely on the fish of these waters for food follows.

Effects of El Niño Current

Deepwater circulation. Observations of temperature and salinity at different levels in the ocean reveal well-defined layers, each forming a water mass distinguished by its own temperature and salinity characteristics.

It appears that the most important influence on the vertical circulation of the Pacific is the cold water generated around the Antarctic continent. This dense circumpolar water sinks and then spreads northward to form the bottom layer of the greater part of the Pacific. It has been suggested that cold, deep water flows northward in the western Pacific in a relatively well-defined current from the vicinity of Antarctica to Japan. Branches from this deep main stream convey cold water eastward and then poleward in both hemispheres.

Deepwater circulation is influenced by the descent of surface water at zones of convergence of neighbouring water flows. In the zone known as the Pacific Tropical Convergence, which coincides with the Equatorial Countercurrent, water sinks to a depth of about 300 feet before it spreads laterally. The Pacific Sub-Tropical Convergences are located between the parallels of 35° and 40° N and S. Water that sinks at the convergences spreads laterally at increasing depths as the distance from the Equator increases. The most important convergence in the Pacific is the Antarctic Convergence, which lies in the zone of the brave west winds. A corresponding Arctic Convergence is prominent in the northeastern Pacific.

The Antarctic Convergence

To compensate for downward-moving water, some water rises at zones of divergence, particularly along the so-called cold water coasts of both North and South America, where upwelling of cold water is a well-marked phenomenon.

Tides. In contrast to the tides of the Atlantic, those of the Pacific include many examples of diurnal and mixed tides. In the diurnal type of tidal oscillation, only a single high water and a single low water occur each tidal day (which lasts for about 24 hours and 50 minutes). Tides of this type occur in the Gulf of Tonkin in Southeast Asia, Vancouver Island (Canada), and the Torres Strait (between Australia and the island of New Guinea). Mixed tides, in which both diurnal and semidiurnal oscillations appear, are characterized by large inequalities in successive high (or low) water heights. This type of tide is prevalent along the Pacific coast of the United States. At certain places in the South Pacific the natural period of oscillation of the sea accentuates the solar tidal oscillation. At these places the time of the AM (or PM) high (or low) water, instead of getting later each day by about 50 minutes (as is generally the case), occurs at about the same time for several days in succession. The tide at Tahiti, for example, follows the sun and not the moon—the time of high water occurring, day after day, at about midnight and noon, and that of low water at about 6 AM and 6 PM.

In general, tidal ranges within the Pacific are small. That at Tahiti is about one foot; at Honolulu it is about two feet; at Yokohama, the port for Tokyo, it seldom exceeds five feet; and at Cape Horn it is never more than about five feet. In the upper reaches of the Gulf of California and the Gulf of Korea, however, large tidal ranges of over 40 feet are common.

Marine life. The most abundant plant forms of the ocean are the microscopic phytoplankton that favour regions where upwelling or mixing of waters brings plant nutrients from deep levels to the euphotic zone (the upper layers of water where light penetrates). Diatoms (unicellular algae) occur in vast numbers in Arctic and Antarctic waters and along the cold-water coasts of both North and South America. Other important phytoplankton include numerous species of flagellates (species of algae having antennae). The phytoplankton provide for the needs of zooplankton (animal forms of plankton), important among which are the radiolarians (microscopic unicellular marine animals, most species of which secrete shells of silica), which are abundant in tropical waters of the Pacific. Foraminifera of the genus *Globigerina* (a minute form of shell-building acellular animal life) are widespread in the South Pacific but are conspicuously absent in the North Pacific.

An important marine animal in the Pacific is the reef-building coral. Some of these animals have the ability to secrete calcareous (chalky) material and construct compound limestone skeletons, which form the major elements of coral reefs. Reef-building corals are sensitive to changes in temperature and salinity. They flourish on firm foundations in clear and shallow water in the Central Pacific between about 30° N and 30° S. The numerous coral atolls of this region are characteristic of this type of formation.

Bottom deposits. The surface waters of the sea form one wide nursery for forms of life. The sea bed, correspondingly, forms one vast burial ground. Apart from the narrow coastal zone of the eastern region and the broad continental seas of the western region, the Pacific is floored with pelagic (*i.e.*, oceanic) material derived from the remains of marine organisms that once inhabited the waters lying above. Red or brown radiolarian ooze is found along the zone of the North Equatorial Current, east of the meridian of 170° W, and on the floors of some deep Indonesian basins. A belt of diatom ooze occurs between the parallels of 45° and 60° S and across the northern Pacific, between Japan and Alaska. *Globigerina* ooze occurs in the shallower parts of the South Pacific, the dissolving power of the sea water at great depths being sufficient to dissolve calcareous material to such an extent that calcareous oozes are not generally found at depths in excess of about 15,000 feet. Silica-containing material, such as radiolarian and diatom ooze, is found at greater depths, but even these siliceous remains are dissolved at very great depths, where the characteristic deposit is red clay. Red clay, which covers no less than half the Pacific floor, is believed to be formed of colloidal (very finely divided) clays derived essentially from the land.

Occurrences of ocean-floor ooze

Among the many different forms of land-derived muds (formed by the erosive action of rivers, tides, and currents) that floor the continental shelves and slopes of the Pacific, the yellow mud of the Yellow Sea is of particular interest. The mud is conveyed to the sea bed by the Huang Ho that drains a vast area of North China blanketed with loess, a fine-grained soil.

THE IMPRINT OF MAN

Economic resources. *Fisheries.* The resources of the Pacific include large quantities of fish, mammals, crustaceans, and mollusks. The northeast Pacific is noted for its salmon fisheries based on the Skeena, Fraser, and Columbia rivers. The fisheries of the northwest Pacific, especially in the seas of Japan and Okhotsk, harvest great quantities of herring, cod, tunny, bonita, crab, lobster, shrimp, and prawn. More fishermen find employment in these waters than in any other fishing ground, and the annual catch of Japan exceeds in value that of any other nation. The waters off the South American coasts are rich in marine life; the growth of the Peruvian anchoveta fishery—anchoveta are a species similar to anchovy—has been phenomenally rapid since 1957. In the early 1970s the quantity—but not the value—of Peru's annual catch was the world's largest. The greater proportion of Peru's catch is converted into fishmeal for cattle and other domestic animals.

Minerals. The mineral resources of the waters of the ocean are vast and virtually inexhaustible. But apart from the extraction of sea salt, magnesium, and bromine, the

Pacific's mineral resources for economic reasons remain untapped. The mineral deposits of the sea beds, especially on the continental shelves, however, are being increasingly exploited. Sea-bed oil prospecting started as far back as 1891 when a well was drilled beneath the shallow waters off the southern California coast. Sea-bed surveys in the Yellow Sea and East China Sea in the 1960s indicated that the continental shelf there may contain one of the most extensive and richest petroleum reserves in the world. Petroleum and natural gas exist in continental shelves adjacent to Alaska, Washington, Oregon, California, Central America, Ecuador, and Peru, as well as in the western Pacific off Japan, Indonesia, Southeast Australia, and New Zealand.

Minerals of the ocean floor

During the scientific voyage of the British naval vessel HMS "Challenger" in the 1870s, nodules or concretions of manganese compounds and minerals were discovered on the floors of the ocean basins. Considerable economic interest was aroused during the International Geophysical Year (1957 to 1958) when scientists discovered huge quantities of nodules on the Pacific floor, off the coast of California as well as in other localities. Research into economical methods of dredging and extracting the valuable metallic nodules continues.

Resource exploitation. In exploiting the resources of the sea, men require knowledge as well as tools and machines. Prospects for making optimum use of the resources of the sea, therefore, depend upon improved technology. No Pacific nation is more aware of this need than Japan, which dominates the commercial fishing of the Pacific.

The sophisticated oceangoing factories of the Japanese fishing fleet are capable of remaining at sea for long periods of time and thus are taking an increasingly important role in Pacific fishing. Fish canneries have been established in American Samoa, the New Hebrides, and Fiji; and these, together with older established factories in Hawaii and the Galápagos Islands, are supplied with tuna and bonita caught by Japanese fishermen.

The formerly prosperous whaling and pearling industries of the Pacific have dwindled to negligible proportions. Pearling for pearls and shell used to be important in many parts of Micronesia and Polynesia, but overgathering has reduced production to insignificance. Fishpond culture and oyster farming are practiced, techniques that may grow in economic importance.

It is worthy of note that the sea and lagoons play a relatively important role in Pacific island life; Polynesian food gatherers learned long ago the art of tending reef "gardens" in order to obtain food from the sea.

History. *Exploration and mapping.* Although the peoples of the western Pacific coastlands formed flourishing civilizations in very early times, they appear to have made no attempt to explore the Pacific beyond their coastal waters. Direct contact between Europe and the Pacific dates from the 16th century, when the exploration of the Pacific began in earnest. The initial European explorations of the Pacific were designed to open up oceanic trade routes to link Europe with the spiceries of the Far East. Later, attention was given to the discovery of Terra Australis. These explorations brought the peoples of western Europe into direct contact with the world of the Pacific islands.

Eight years after the Spanish conquistador Vasco Núñez de Balboa had sighted the "peaceful sea" from a peak in Darien in the Isthmus of Panama in 1513, the ship of Ferdinand Magellan, the Portuguese navigator, voyaged across the Pacific. Subsequent Spanish voyages led to the European discovery of the Moluccas, the Carolines, Papua, the Hawaiian Islands, and the Solomons. From the beginning of the 17th century, Pacific voyages of discovery were dominated by the Dutch, and credit is due to the Dutch navigator Abel Tasman for his significant discoveries of Tasmania, New Zealand, and Fiji.

Dutch voyages of discovery

A later major period of Pacific exploration is marked by the 18th-century voyages of the English navigator James Cook, the French navigator Louis-Antoine de Bougainville, and others who made possible the accurate mapping of the whole of the Pacific.

In the 19th century, the English naturalist Charles Darwin sailed on the British naval vessel HMS "Beagle," which circumnavigated the globe between 1831 and 1836. Growing interest, at that time, in the physical and biological problems of the ocean paved the way for the many oceanographic voyages made later in that century. These were initiated by the previously mentioned "Challenger" expedition.

The voyage of USS "Tuscarora," from 1874 to 1875, contributed significantly to oceanographic discovery in the northern Pacific; scientists aboard the German research vessel "Gazelle" in its 1874–76 voyage added considerably to knowledge of oceanic physics. During the International Geophysical Year the Soviet vessel "Vityaz" carried out detailed soundings of the deep trenches of the western Pacific. In the same year the United States bathysphere "Trieste" descended to the bottom of the Mariana Trench, reaching the greatest known depth of 36,198 feet.

Eras of naval predominance. The European intrusion into the Pacific resulted in the establishment of the first Spanish colony in the Philippines, in 1564. The 17th century witnessed the rise of Dutch naval power in the East Indies. From the beginning of the 19th century onward, the Pacific became the scene of a struggle for colonies and coaling stations by the maritime powers of western Europe.

The history of modern naval strategy in the Pacific dates from 1894, when Japan eliminated the Chinese navy. Ten years later Japan destroyed the Russian Far East fleet and became ruler of the western Pacific, from its own shores to the Philippines. To safeguard British interests in the Pacific an alliance between Great Britain and Japan was concluded in 1902.

The rise of Japan's naval dominance

After World War I Japan acquired former German Pacific islands, thus strengthening its naval influence. Although the opening of the Panama Canal (*q.v.*) in 1914 allowed the United States to operate its fleet in the Pacific, Japanese control of the western Pacific remained unchallenged. World War II committed the major part of the British fleet to the Atlantic, and in 1941 significant U.S. naval forces were transferred to the Atlantic from the Pacific. Nevertheless, Japan, despite its military successes in Asia, failed to establish control of the eastern Pacific. By degrees the United States wrested command of the western Pacific from Japan and, since World War II, has been master of the whole of the Pacific. The South Pacific remains a sphere of influence in which Great Britain, Australia, and New Zealand are important naval powers.

Trade and communications. Although Thor Heyerdahl, the Norwegian author and anthropologist, has amply demonstrated the possibility that at least some Pacific islands may first have been inhabited by peoples from South America, the weight of anthropological evidence suggests that the indigenous peoples of the Pacific islands originated in Asia and that the peninsulas of Southeast Asia were the principal centres from which Polynesian, Micronesian, and Melanesian settlement spread out. The European impact on the Pacific, which began in the early 16th century when the Spanish entered Central America, did not become significant until about 1800, when local Pacific economies were severely disrupted by European exploitation of the resources of the Pacific islands. This epoch was marked by the arrival of numerous planters, traders, and missionaries, and the introduction into the Pacific islands of relatively large numbers of Indian, Chinese, and Japanese labourers. The native populations decreased sharply as a result of diseases introduced by Europeans, the introduction of slavery, and the more lethal local warfare caused by the availability of guns.

The development of the North American west coast during the early 19th century, and the concomitant extension of United States trade into the Pacific, ushered in a period of profitable trading in which manufactures were exported from North America in exchange ultimately for silks, spices, and other products of the Far East and Oceania.

Since the beginning of the 20th century, Japan's influ-

ence in the Pacific has increased. After World War II, when Japan's economic power was temporarily eclipsed, Australia, New Zealand, France, The Netherlands, the United States, and the United Kingdom were instrumental in establishing the South Pacific Commission, with headquarters in Australia, which aims to encourage and coordinate research into the economic and social problems of the region.

The South Pacific Commission

The major trading patterns of the Pacific are essentially determined by political factors. Strong trading links have long existed between Hawaii (which achieved statehood in 1959) and the United States; between Great Britain and France and their respective scattered colonial Pacific islands; and between Australia and New Zealand and their dependent territories.

Communications. Communications across the Pacific are better than they are between island communities within the Pacific. Regular ocean-freight and passenger services connect North America with Japan and the Philippines; and Honolulu, the capital of Hawaii, serves as an important port of call. The Panama Canal provides a sea-trade route linking Atlantic and Pacific ports. The principal focal point of air and shipping services in the Far East is Japan; from there, routes lead southward to Australasia and, by way of the Strait of Malacca, to the Indian and Atlantic oceans. Important airports providing calling points in an increasingly complex network of air services include Honolulu (Hawaii), Papeete (Tahiti), Pago Pago (American Samoa), and Nandi (Fiji).

Prospects for the future. The peoples of the colonial island territories of Melanesia, Micronesia, and Polynesia are becoming increasingly responsible for their own political affairs, and self-government appears to be inevitable for them all. The rising birthrate in the Pacific islands, especially in the low-lying islands where land resources are severely restricted, is resulting in overpopulation. Mixed populations, in which the indigenous peoples in many instances have been, or are being, rapidly outnumbered by people of Indian, Chinese, Japanese, and European origin, give rise to complex social problems.

Population problems of the islands

Jet airplanes and modern airports have marked the beginning of a new epoch for many Pacific islands. Tourism, doubtless, will boom in numerous hitherto remote Pacific areas.

It appears that the comparatively new states of Indonesia and the Philippines, as well as China and Japan, will become increasingly important in the western Pacific, politically as well as economically. The vast populations of the countries of eastern Asia represent an enormous demand for food and manufactures. Japan, the United States, and the Soviet Union, with their advanced technologies, are capable of producing enough manufactured goods to meet this demand. The mineral wealth of the Pacific rimlands is, moreover, substantial. Pacific sea trade may consequently be expected to increase.

There can be no doubt that the recent devastating wars in Korea and the Indochina peninsula have retarded economic and social progress and have had far-reaching effects in the whole of the Pacific region. Were it not for these hostilities, Theodore Roosevelt's prophecy of 1905 that "The 20th century will be the century of the Pacific," might well already have come to pass.

BIBLIOGRAPHY. ADMIRALTY (BRITISH), NAVAL STAFF, NAVAL INTELLIGENCE DIVISION, *Pacific Islands*, 4 vol. (1943–45), an invaluable compilation of physical and sociological material; O.W. FREEMAN (ed.), *Geography of the Pacific* (1951), a standard geographical text; M.N. HILL (ed.), *The Sea*, 3 vol. (1962–63), oceanographical articles for the specialist by experts in their respective fields; H.W. MENARD, *Marine Geology of the Pacific* (1964), a general and very readable account; F. OSBORN (ed.), *The Pacific World* (1944), an excellent collection of articles concerning the peoples and lands of the Pacific; H.U. SVERDRUP, M.W. JOHNSON, and R.H. FLEMING, *The Oceans* (1942), a general text embracing physical, chemical, and biological oceanography; C.R.H. TAYLOR, *A Pacific Bibliography*, 2nd ed. (1965), a valuable work relating to the native peoples of Polynesia, Melanesia, and Micronesia.

(C.H.C.)

Pacifism and Nonviolent Movements

Pacifism is taken in this article to embrace the sum of all endeavours and programs for the realization of a lasting or, if possible, a perpetual peace between peoples in the belief that this goal is of positive value and can be attained within the historically foreseeable future. This is a broader definition, of course, than that of common usage, which refers to movements for the total abolition of war. The term nonviolent movements refers to what have been, as a rule, sects or movements of small minorities that share a fundamental conviction that it is possible to transmit the individual ethic of "negative actions" (nonresistance, sufferance, and nonviolence) to the collective arena.

NATURE AND RELATIONS OF THE MOVEMENTS

In its general form as just defined, pacifism can be found to exist in all higher cultures and in all historical times, as a more or less distinct and vivid idea, often supported by (1) religious and philosophical or ethical demands for the abandonment of violence, (2) the postulate of tolerance, and (3) programs aimed at the improvement of relations between nations, the limitation of armaments, the moderation and rational discussion of conflicts, and the institution of neutral courts of arbitration. As a rule, the basis for such programs lies in the conception of an organic social ethic and a harmonious human society. Yet, there are different degrees of rigour among pacifist conceptions and ambitions. Whereas there is an integral pacifism that eschews, under all circumstances, the violent settlement of conflicts and rejects war unconditionally, there is also a less severe, or compromising, semipacifism that permits wars under certain conditions, as, for instance, when they are "just" or decidedly wars of "defense" or wars against "unbelievers" or "rebels." Naturally, the conception of pacifism becomes problematic in this case. Furthermore, there exists a conception of a positive peace, which connects the abandonment of armed force with religious or cryptoreligious sanctions on the one hand and demands of social justice and inner peace on the other. There can, in addition, be distinguished a conception of a negative peace, which considers that the mere renunciation of armed force is the most that can be attained through intelligent politics.

Types of pacifism and peace

Social and religious movements with the renunciation of violence on their banners have often been the germ cells of pacifistic challenges, for which, on the basis of religious doctrines, they could provide a more or less precise ideological foundation. Among groups exemplifying nonviolent movements (as earlier defined) are those of Chinese Taoism, of Buddhism, of the Shīʿite branch of Islām, of medieval Christianity, and of certain Protestant sects. In its methodical application, however, the principle of negative actions espoused by these groups always turns out to be a procedure tenable only for minorities or elite cadres and this often causes the nonviolent movements to separate themselves from pacifism. The nonviolent movement of Mahatma Gandhi, for instance, is to be explicitly understood as a method of constitutional opposition within an existing governmental system aimed at relieving its own tensions and not at the solution of international conflicts; thus, Gandhi was not an integral pacifist. Nevertheless, besides certain emotional suppositions, the nonviolent movements generally share with pacifism an emphasis on the peaceful solution of conflicts through the renunciation of violence—though the nonviolent movements and pacificism do not always apply these principles in the same sphere.

GENESIS AND DYNAMICS OF WAR AND PEACE

It is impossible to consider peacemaking tendencies without taking into account the empirical evidence regarding the dynamic relation that exists between peace and war. Peace between different ethnic groups is held by many historical and ethnographic studies to be not a natural situation but a conscious creation, as Immanuel Kant proclaimed in his essay "Zum ewigen Frieden" (1795; "Toward Eternal Peace").

Causes of combat among primitives

Inherent conflict and its control. Among primitive tribes there is always present the possibility of a latent hostility, which does not always break out, however, into actual combat. This possibility is based on an ethnocentric conception of the world as comprised of simple kinship alliances, clans, tribes, and confederacies. An ingroup and an outgroup ("we" and "they"), with their corresponding inward and outward moralities, are strictly distinguished, so that friendship, solidarity, and peace exist, for the most part, only inside the ingroup, whereas xenophobia and readiness for battle prevail in attitudes directed outward. Whoever belongs to the outgroup is essentially an alien; friendly and peaceful relations with him are not excluded but have to be purposely established through certain rituals. Among these groups, military conflicts arise less from the wanton urge to attack than they do from a readiness for aggression issuing from an atmosphere of fear, tension, feelings of injury, frustration, and offense to their complex of honour and prestige. Well-defined causes and goals for war are seldom encountered—with the exception of blood revenge, with its goal of mere reprisal, which in its turn is answered by reprisal, and so on (theoretically) *ad infinitum*. An actual redemption from the feud becomes functionally possible when an authority that encroaches upon the parties obtains recognition; this authority is the root of the primitive state. Higher primitive races formulate more precise war goals: plunder, slaves, or the extension of the power of their own group by the incorporation of prisoners of war. There is a great reluctance, however, about annexing foreign areas, because the hostile foreign land is thought to be occupied by strange demons and, therefore, dangerous. But the adoption of strangers into their own group already prefigures the higher organized nations of the old Orient, the Sudan, and both Americas.

Methods of combat and peace-making

Methods of combat among primitive tribes are of two main types: first, an unprepared invasion of a hostile settlement at dawn with the aim of destroying it and, if possible, killing many inhabitants, including women and children; or, second, an open combat, announced in advance and carried out in an arena upon which both adversaries have agreed. Such combat is generally connected with certain further regulations and moderations—for instance, the prohibition of certain dangerous weapons, settlement of the combat through a match between duelists, sanctuary for cult places and channels of commerce, exemption from injury of women, children, and respected persons (older persons, priests, guests), which in practice means asylum for civil persons and a distinction between combatants and noncombatants. These regulations mark the genetic beginnings of the laws of arms with moderating and disciplining effects. It is especially the second type of combat, as practiced between traditionally hostile groups, that leads to a real and established peace. Such a peace is never amorphous but is confirmed in ceremonies and often strengthened by an exchange of women for intermarriage and the establishment of hospitable relations. In the case of primitive tribes, this created peace is perceived not as "eternal" but as having its place in a general rhythm of war and peace. It is usually ephemeral. Gauged by the ideal of positive peace, such interludes may seem trivial; but in human history they have meant much. The ethnology of war and peace suggests that relative stages of peace must be distinguished, and in reality the moderations and regulations of combat are already significant as elements of peaceful achievement.

Effect of religious sanctions

Efforts to confirm a lasting peace through religious sanctions have had little effect. Among primitive races (in parts of Polynesia, Melanesia, Africa, and North America), cultic alliances and confederations existed between different tribes whose members temporarily shunned hostilities with each other at times of a "truce of God." Such combines date far back into historical time (compare the amphictyons of the Hellenes and Israelites, associations initiated to defend a common shrine). Yet, the peace-creating effect of the common cult was never very strong. Genetically, the cult unions were the germinal forms of later supranational authoritative courts organized to try to establish a lasting peace. To sanction peace, however, involves the danger that war, if it does break out, will be marked by an additional element of quasireligious fanaticism. But if sanction systems cannot really eliminate armed conflicts, it is also true that the actual carrying out of conflicts can take place only within such a combine system, in which the adversary is no longer regarded as sovereign but as deserting or rebellious, and the struggle against him is then no longer war but federal execution, punishment, excommunication, or collective outlawing, so that the restrictions normally imposed against such an adversary by the articles of war are inoperative. This internalization is a setback in the development of martial and international law and is detrimental to the creation of a situation for a real peace. The holy wars of the Hellenic amphictyons, the religious wars of late antiquity (since Constantine the Great) between paganism and blossoming Christianity, the Crusades, and the later religious wars in West and East have always been waged with singular brutality and cruelty. Terminologically, moreover, they provide a broad opening for the perversion of language, which occupies such an important place in the semantics of "peace" rhetoric. This happens, in particular, when an element of fanaticism present in the competing religions can break in and, independent of the dogmatic content of the respective religions, announce "tolerance" and "peace."

Limiting the scope of war. There is, nevertheless, a kind of progress in the conditions for peace through the history of mankind, which is the achievement of politics with religion playing an important, though ancillary, role. Peace and security have gained ground through both the broadening of the geographical areas of peace and the expansion of the class of noncombatants.

Expansion of peace zones. By expanding smaller sovereign political units into larger ones, armed conflicts have been suspended within larger units. Naturally, this broadening has always stemmed from earlier peaceful—or, more often, military—annexations. As a consequence of such processes, a kind of imperialistic pacifism could easily be proclaimed. After many wars of conquest and annexations, for example, the emperor Aśoka of India (*c.* 265–238 BC) announced numerous edicts of tolerance, by which he made use of the Buddhistic ethics as a means of domesticating robber castes and rebellious and uncertain tribes. The same was done by Prince Shōtoku of Japan in the 6th century; the universal ethic of Buddhism aided him in shattering the traditional clan units and welding together a centralized state. The Achaemenid rulers in Iran made use of the religion of Zarathushtra for the same purpose, and Darius I wrote inscriptions comparable to the edicts of Aśoka. Such manifestations, however, are a matter of religiopolitical decrees and not of tolerance in the sense of the 18th-century Enlightenment. The regulations of the Christianized ruler Constantine the Great were also religiopolitical: a warfare of Christ and a Pax Christiana (Christian Peace) of the young church had to help confirm the Pax Romana. The fact that this process is always one of domesticating and taming of competing inner groups, parties, sects, castes, tribes, and small nations and a process of subduing or silencing the opposition within is illustrated by the epithet of the Oriental rulers, "Shepherds of the Nations"; by the comparison of pacified tribes with tamed beasts or the hope that they will be miraculously tamed in a future state (in the Old Testament); and by the Buddhist term *purisadamma* ("the human animal who can be tamed").

Imperialistic pacifism

All of these symbols suggest that a peaceful behaviour of man does not arise by itself but requires regulating and compulsory forces. This fact in no way impugns the religious sincerity of the rulers who employed such edicts and symbols, nor does it question the reality of the legal security that was guaranteed within the limits of peace of these realms. But this reality was created by war and force. The paradox of "imperial pacifism" is revealed by the insight of the philosopher Max Scheler, an early-20th-century Phenomenologist, who said during a time of declining imperialism that "the pure idea of law has no power in itself" but that right requires a power that guarantees it while it nevertheless remains right. This

Paradox of forceful pacification

principle, expressed earlier by Aristotle (and then by St. Augustine) in the formula that peace is the aim of war, is a genuine *aporia*, leading to insoluble logical difficulties; and politically seen, the principle has time and again been expressed in the paradoxical concept of pacification, which means exactly those violent actions through which an expanded area of peace shall be won and maintained. This phraseology first occurred among the Romans but was later taken over by other empires and most expressively by the Spaniards in their New World colonies, where the term *conquista* was replaced by *pacificación*. Characterizations of the imperial peace that correspond with this notion are, for instance, Pax Romana, Pax Christiana, Pax Hispanica, and Pax Britannica. In all of this terminology the heritage of the Roman conception of justice is reflected. In the case of the Romans, the terms *securitas pacis* and Pax Romana encompassed the whole area of authority of Roman laws—laws that, by their introduction among the subjugated peoples, resulted also in the introduction of the values of Roman civilization. Since Cicero and Julius Caesar, at the latest, *pacare* and *pacificatio* have been expressions denoting the military subjugation of a foreign nation and its incorporation into the jurisdiction of Rome. In this terminology, which still has aftereffects in the West, there is an inextricable interlacing of the reality of the territorially expanded peace and its security, the force and violence through which this situation was set up, and the ideological proclamation of peace and mutual tolerance as a means of domination. The opposition of those who are deprived of their political independence in the enforcement of this peace is not expressed in this terminology but is nonetheless implicitly contained in it.

Expansion of noncombatant classes. Next to the expansion of zones of peace, the recognition and expansion of the class of noncombatants during times of war is the most important step toward the goal of maintaining a relative degree of security for the life and property of the largest possible part of the population, even though it is not possible to protect the people completely from the actions of war. A privileged noncombatant class was found among primitive tribes and, indeed, existed among the Western nations as long as wars were conducted with mercenaries—*i.e.*, with minorities of combatants. It was lost in modern times by four paramount factors: (1) During the French Revolution, the principle of the equality of all citizens led to a general levy into military service and to the creation of enormous national armies. As long as the armies were small and operated with baggage-train and reinforcement columns, the largest part of the population could live in relative peace even in the midst of war. When the Napoleonic armies, however, changed over to living off the land itself, an important step was taken toward making war "total." (2) In military technology powerful mass weapons were developed that did not discriminate between combatants and noncombatants. (3) Psychological warfare was instituted to propagandize the population behind the enemy's lines and break down its resistance—the psychological equivalent of long-distance bombers. (4) And, faced with guerrilla tactics that dissolve the distinction between combatants and noncombatants (*e.g.*, by dispensing with uniforms and terrorizing populations), Western nations devised countermeasures that were similarly nondiscriminatory. Through these (and other) factors, structurally interwoven with one another, the traditional concept of a military "front" was antiquated: it became a general structural feature of modern military conflicts that the "front lines" were no longer clearly defined but were disarranged and confused. In addition, there arose a new element of national and revolutionary fanaticism, which replaced that of the earlier religious wars and was and is no less potent than the latter. The culminating point of this development is reached at the present time by the modern principles of the "revolutionary people's war."

PACIFISM IN THE HISTORY OF MAN

In ancient times. Classical antiquity knew no pacifism. The Hellenic poets complained about the evil and destruction of war; but the idea that things could be fundamentally different did not arise. The so-called pacifism of certain intellectuals of the 4th and 3rd centuries BC (such as Xenophon, a Greek essayist and soldier, and the Athenian orator Isocrates) was of an opportunistic, rather than a fundamental, nature. Protracted wars such as the Peloponnesian War, waged among the Greeks late in the 5th century BC, left an atmosphere of war weariness behind and possibly stimulated the idea that philanthrōpia ("humanity," "benevolence") and related conceptions are authentic values not only for private but for public conduct as well. Yet, the Hellenistic rulers—in particular, the Ptolemies in Egypt in their peace proclamations (*philanthrōpa*)—were the first to make amnesty a repeatedly practiced institution for rebels, emigrants, and persecuted people. The thought of a general peace presupposes a change of consciousness that would make the concept of unitary humanity possible. This consciousness had not been possible on the basis of the Hellenic polis (or city-state) but required an expansion of the horizons to the Hellenistic cosmopolis. Thus, the first timid stirrings of pacifism can be seen to emerge after the expansion of the peaceful zone in the Hellenistic states. Several intellectuals early grasped the idea of a world state without courts of justice, temples, or a monetary system. The Stoic Law of Nature showed certain very modest tendencies toward the idea of a lasting peace. The Stoic ethic, nevertheless, remained essentially a private ethic that embraced the ideals of an inner peace composed of apathy, nonaction, and strength of mind for the single individual but did not carry over such ideas to the community. The same observation can be made regarding the religiophilosophical systems of the East, where conceptions of a Law of Nature were never entirely awakened: the Taoist idea of *wu-wei* ("nonaction") and the Jainist and Buddhist demand for *ahiṃsā* ("noninjury") were demands made upon individuals and served their inner peace but remained irrelevant for political action. Gandhi was the first to interpret *ahiṃsā* positively and in the sense of a social obligation.

Greek approaches to peace

Roman juridical pax

In contrast with the Hellenistic conception of peace, which simply meant a peaceful situation, the Roman conception (pax) had from the outset a judicial character: pax is a covenant that creates a just situation, which rests upon bilateral recognition. The Roman conception of peace is therefore more political than the Greek conception, and insofar as pacifist ideas in Europe took the form of international law, they go back to the tradition of the Roman law. Additionally, near the beginning of the Christian Era there arose another complex of pacifist aspirations: "eschatological" pacifism, closely related with the conceptions of a revived "Golden Age," which was destined later to permeate all other forms of pacifism, whether humanistic, juridical, or illuminative, as a strong, emotionally resounding, messianic, and prophetically inclined undercurrent. A strong longing for peace spread abroad in the Roman Empire toward the end of the republic as a consequence of uninterrupted civil and foreign wars, a longing reflected, for example, in the myth of a divine, sunlike child who would redeem the world from the "eternal horror" (Virgil) and, as saviour and prince of peace, would lead to an era of world peace.

It was natural that in the principate of Augustus, the Pax Augusta would be interpreted with the aid of this prophecy. In Virgil, however, the renaissance of the era is portrayed as being at least as important as peace itself, which is thus not absolute because it is said that later there will be wars yet again. Clearly, these messianic-prophetic ideas were admitted into the legend of the Christian birth of the Messiah. The Pax Romana, however, was not really universal, because it was always regarded as a world peace for the civilized world, and it thus excluded barbarians; and obviously the peace would have to be defended against these barbarians. Accordingly, the peoples that were not integrated into the Roman Empire were always excluded from the peace, just as the pagans, unbelievers, and heretics were later to be excluded during the Christian Era. When, late in the 3rd century, Emperor Aurelius Probus could finally say that he no

longer needed soldiers and could demobilize his troops, this alleged pacifism was the luxurious product of a saturated imperialism. Because the threat to the frontiers of the civilized world by the barbarians and pagans never ceased, this pacifism could never become integral. When this was connected with the religious sanction, it always remained in the guise of an eschatological pacifism and cultivated the idea of a last war, which must be conducted in order to make the world ripe for an empire of eternal peace ("the war to end war"). This conception of an integral pacification through an ultimate war has also in its secularized form remained tempting through all later centuries and, in particular, in its world-revolutionary–Communistic version.

In medieval Christendom. Christianity, with its evangelical message, offered considerations in support of individual nonviolence as well as of collective peacefulness. It is questionable, however, whether the ethics of Jesus reflect the apolitical and pietistic pacifism of the Essene sect, a strict monastic brotherhood that had numerous followers in Palestine. There are words in the evangelical message that lend themselves to an interpretation of "peace" as the peace of mind of the individual ("My peace I give unto you: not as the world giveth, give I unto you"), but this unworldliness is counterbalanced by other sayings praising those who are ready for peace ("On earth peace, goodwill among men" and "Blessed are the peacemakers," in which the Greek *eirēnopoios* may be equated with the Latin *pacificus*), which can be interpreted in the sense of an integral pacifism and is, in fact, so interpreted by the many radical followers of Christ. As a rule, however, this peace was only open to minorities, or sects that practiced a rigorous ethic, while the church had to compromise with worldly necessities. Since Marcus Aurelius and Commodus, in the era of the Roman emperors, the "question of soldiers" (*i.e.*, the question whether a Christian could be or remain a soldier) had continued to perplex the faithful. The Church Father Tertullian answered the question unambiguously in the negative but testified at the same time that there were many Christians in the Roman Army and that his conception was therefore not generally shared. Though it seems that Christian soldiers who suffered a martyr's death are an exception, desertions from the army after inner or outward conflicts were, in contrast, more frequent. In consequence, to practice the Gospel strictly within the military profession was, naturally, hopeless. On the other hand, there were not only the Old Testament sources with their warlike authorities to which one could refer but also passages in the Gospels, such as those citing the centurion of Capernaum and the centurion of Caesarea, with the help of which attempts have been made to prove that the military profession could not be completely non-Christian. Subsequent to Tertullian (early 3rd century), the symbols of the military service, discipline, and combat even break into the Christian sermon, so that a Christian became understood as *miles Christi* ("soldier of Christ") or even as *miles gloriosus*, who had to fight the demons; the church became the camp of God, and the schismatics were called the gang of Korah—in reference to an ancient rebellion against Moses. Military language, in fact, became a fashion, as it did again later in the order of the Jesuits.

The "question of soldiers"

In the beginning of the 5th century, St. Augustine, bishop of Hippo, in his work *De civitate Dei* (*The City of God*), created a reconciliation between the ideas of a worldly and a supraworldly state and built some thoughts on peace into this counterpoised system, which reflects his enthusiasm about the church's victory in the world but also his impression of the threat to this world posed by the invasions of the barbarians—in particular, the invasion of the Goths (AD 410). Augustine saw the empire of a thousand years, of which the Apocalypse had spoken, as already realized in his own present epoch—but only insofar as the state is spiritually penetrated and ruled by the church. Without such influence the state is no better than a band of robbers or a city of the devil; for worldly states, taken by themselves, arose out of sin, even though this sin is called virtue. They can establish, though with the greatest of difficulty, the worldly peace, Augustine held, but not the heavenly peace, which is the greatest good and, indeed, the eternal harmony itself—a standard by which all worldly achievements are measured, including worldly peace. This worldly peace, which Augustine also called the Pax Babylonis, is in fact only a pseudo-peace, which, nonetheless, need not be despised; for, even in this disfigured form, the peace is still worth striving for, and a pious man may seek it as long as he must live in community with impious men—though only insofar as it allows him to consider the heavenly peace as well. Beyond this, in Augustine's view, the devout man does not have to take part in the worldly peace and, indeed, must even destroy it when it is contrary to God. The same applies to nations. Consequently, a war against evil men is as surely justified as the combat of the martyrs against the demons. Admittedly, such a just war is not a happy thing, and the most incomplete worldly peace remains a relative good and is, in fact, according to the interpretation of Augustine, all that can be attained for mankind on this earth. Nonetheless, this worldly state should always be interlaced with the spiritual reality of the heavenly state, by the image of which, alone, does the earthly state have a right to exist. Moreover, it is above all necessary that the worldly state shall serve the church and stand up against all peacebreakers from within and without and especially against heretics and schismatics. Thus, for Augustine, the state that is ruled by the church is a magnificent, comprehensive, and powerful sanction system.

This idea, which is that of the Catholic Church, was maintained throughout the Middle Ages. Charlemagne added the epithet *pacificus* to his emperor's title, and the German kings called themselves *Friedensfürsten* ("peace princes"). Behind these claims lay the popular millenarian expectation of an eschatological emperor, which, since the 13th century, was more closely related with the sects and orders than with the official dogma of the church. The Sibylline oracles, some of whose prophecies, though pagan, were at times taken seriously by Christian writers, reduced the doctrine of an eschatological emperor to the plane of history and saw him as a saving warrior of God who suppresses the power of the Antichrist and introduces an era of peace. Taken wholly in the sense of Augustine, however, this peace was not absolute but included the possibility and even the obligation of combatting unbelievers, heretics, and rebels; the true peace entails the just war in the name of Christ. Thus, it could happen that the reigns of several kings were praised by chroniclers as times of peace whereas in reality they were filled with war. Thus, medieval pacifism, by excluding certain categories of persons, groups, and nations, was always incomplete; and the background for this exclusion was always quite real because there were always heretics, schismatics, and, at the borders—as successors of the barbarian tribes of the Roman era—unbelievers: Arabs, Mongols, or Turks. This world view was described in terms of a dualism, or dichotomy, of good and evil spirits, Christians and non-Christians, and, as the latter were poorly known, the world view was strongly demonological. Thus, the sphere of the real as well as of the ideal and sanctioned peace was always limited, leaving outside of its province a very large part of mankind as demonic and doomed to hell. The conceptions of Dante Alighieri were also governed by this dichotomy, for he conceived of the peaceful state as one that must be outwardly defended. Moreover, the attempts to limit internal wars by a truce of God (*treuga Dei*) met with only trifling success.

The eschatological emperor

Probably the first attempt to overcome the dualistic concept was made in the late Middle Ages by Cardinal Nicholas of Cusa (Cusanus), a scholar in many fields, whose book *De pace fidei* ("On the Peace of Faith") reflects the impressions made upon him by the conquest of Constantinople by the Turks (1453). Cusanus abhorred "holy wars" and believed that the different religions should tolerate each other. He caught the vision of a kind of League of Nations, which would have to accept his philosophy of the coincidence of opposites and derive from this a peace in religious faith.

In the modern world. Since the time of the Renaissance, the possibility of a world free from wars has been considered by several intellectuals, who had very different relations, however, to the realities of politics. The humanist theologian Erasmus of Rotterdam (died in 1536), whose thought was permeated by the idea of a renewal of mankind by a deepened understanding of the Holy Scriptures, was cosmopolitan in his sentiments and an adversary of wars; he taught that the highest ideal of mankind would be peace and concord (*tranquillitas orbis Christiani*). The 17th-century French monk Émeric Crucé is important because he was the first of a series of thinkers who saw the main cause for the outbreak of wars in the strong-handed, dynastic power politics of the princes. Crucé perceived the remedy to lie in an appeal to the moderating reason of the sovereigns and in the improvement of the judicial relations between states. He thought that a transfer of the political pretensions of the rulers to the public would improve the situation. Moreover, the later legal pacifists shared, for the most part, the conviction that the public's power of reasoning was greater than that of the sovereigns. At the beginning of the 17th century, Maximilien, duc de Sully, who was a minister under King Henry IV of France, conceived a plan of European federation that was supposed to re-establish the religious unity of Christendom and to expel the unbelievers from Europe. The plan remained unrealized, notwithstanding attempts to bring about negotiations on it between the European royal courts. One hundred years later, after the Peace of Utrecht (1713), Charles, abbé de Saint-Pierre, a French social critic, again demanded the establishment of a confederation of states that would guarantee an eternal peace; but it too, was in vain. The great teachers of international law of the 17th century, the broadly learned Dutch scholar Hugo Grotius and the German publicist and jurist Samuel von Pufendorf, were not as much pacifists as they were designers of the just war, and only the presumption that the latter was considered an exception moves them toward semipacifism. The conceptions of Natural Law held by these and other thinkers reveal more and more in the following period the features of a revolutionary ideology reflecting the interests of the rising bourgeoisie as against the postfeudal or absolutistic relations of constitutional law. The abbé de Saint-Pierre had already demanded that, if necessary, the confederation to be founded had to fight those sovereigns who refused to join the new order.

Diverse peace proposals

The culminating point against the sovereigns became even more evident after the outbreak of the French Revolution. Even Kant, in his work *Vom ewigen Frieden* (1795, 1796; "Of Eternal Peace"), which contains many realistic insights, defended the claim that the monarchies tended toward wars only because the sovereigns regarded their states as their personal property and that, compared to this, a republic would be peaceful. This illusion had a long-lasting aftereffect. In the 19th century Sully's and Saint-Pierre's ideas were realized in an alliance of the great powers (1815–18), which thereupon became identified with the autocratic system of Metternich, known for its abhorrence of any essential change. This was an alliance of the sovereigns, however, not of republics. Other supranational organizations also existed after 1815; many voluntary pacifist associations were founded (New York, 1815; London, 1816; Geneva, 1830; Paris, 1841; and other cities), which organized congresses (Brussels, 1848; Paris, 1849; Frankfurt am Main, 1850; and London, 1851). The theme of pacifism thereby caught the public interest and inspired an extensive literature. General disarmament, the establishment of referees' courts to consider international conflicts, and the placing of moral stigmata on loans for purposes of war were all proposed. Some of these ideas were later realized in the Court of Arbitration in The Hague, in the League of Nations, and in the United Nations, as well as in temporary disarmament conferences. Nevertheless, these proposals remained without effective response during that time. The real maintenance of a relative peace in the 19th century was accomplished through the statesmanlike political establishment of a balance of power among the five great European states. The peace was also upheld by the grace of the Pax Britannica, which lasted as long as a British naval force existed and British finance retained its dominating position. These factors did not accomplish universal peace, but they created a certain control and localization of military conflicts. Those who fostered the system of free trade and expanding capitalism were not desirous of war, but they created the mentality of what was known as the bourgeois-liberal "Manchester Pacifism," based not on altruistic motives but on a utilitarian system of values that generally opposed an expansion in armed conflicts. The champions of this system wanted to rule but were against armed conflicts of imperial form.

Peace organizations

There is no particular correlation between Communism and pacifism. The chief theoreticians of the former, Friedrich Engels, Karl Marx, and Lenin, were influenced by the military philosophy of the Prussian strategist Gen. Carl von Clausewitz, and they nourished the idea of a world revolutionary war. For Marx, "war" and "revolution" were interchangeable terms. Lenin explicitly rejected pacifism. The idea, however, of justifiable wars against the capitalistic classes moves the ideological system of revolutionary Communism toward eschatological modes of thought, which indulge in dreams of universal peace in a classless society following the world-revolutionary victory of the international proletariat. But the accent falls less on this finality than it does on the justifiable war that will lead in this direction. Thus, the idea of national war has been accused of creating a sentiment not for universal peace but for war, pugnacity, and hate.

Marx and Communism

NONVIOLENT MOVEMENTS IN MODERN TIMES

While endeavours, valuations, ideologies, and programs are involved in pacifism, the device of nonviolence is an immediate method for the action of individuals or small but determined and convinced groups. Though chiefly a modern phenomenon, the method was understood by the Chinese Taoist wise men (such as Lao-tzu and Chuang-Tzu—in the 6th and 4th centuries BC) as negative action, nonaction, or nonresisting. In this case the method was completely apolitical and escapist. In practice, it was a course of withdrawal and elastic evasiveness. Yet, it also displayed a positive accent—"the soft water, if moving," will in the course of time "wear away the hard stone"—an accent by which this method is correlated with the virtue of patience and adapted to a sense of time that is more in accordance with Eastern religions than with Western activism. When the philosophy of nonresistance is connected with a positive goal and is practiced by an organized plurality, it becomes a nonviolent action. Nevertheless, if it preserves the terminology and ideology of negative action, as in Gandhi's philosophy of suffering, it becomes especially effective through the action that publicly interprets itself, paradoxically, as a nonaction.

Specific modern movements. In the past four centuries, nonviolent action has been proclaimed and practiced in exemplary fashion by certain Protestant sects, most convincingly by the Anabaptists, the Quakers, and the Mennonites. Theologically speaking, the purely unworldly demands and rigorous ethical principles of the Sermon on the Mount, and the religious natural law derived from them, were mobilized in radical religious groups who, esteeming themselves to be especially qualified, embraced the idea of a purely spiritual, "invisible church" and thus separated from "the world" and refused to become involved in state offices, legal courts, the oath, and military service. These refusals were partly a consequence of the ethics of the Sermon on the Mount (taken literally) and partly an offshoot from the general anti-institutionalism characteristic of this type of religious group from early Christian times. Therefore, for the English spiritualistic sects of the 16th century (for example), war was as serious a problem as the legal coercion of the government. Theologically, Calvinism, more than any other creed, found war to be a problem: if at all, only defensive war was to be allowed. Especially with the Calvinists, and above all, with the sects, a humanitarian movement grew up directed against war and in favour of its compensation by a system of covenants and arbitrations; but it was

Nonviolence as practiced by Christian religious sects

always in conflict with the belligerent and imperialistic inclinations of the majorities. Alongside the secularization and superficialization of the religious impulses, however, all of these groups and movements had to adjust to the worldly powers, which alone were able to safeguard the domination of the "right" creed from without—thus re-enacting the situation of the late Roman Empire. The Quaker community in Pennsylvania (1683 to *c.* 1770) provides a good example: it was decidedly nonviolent and pacifistic, but was entangled in the quarrels between England and France, in which it sided in warlike manner with the mother country; thus the colony lost its Quakerish stamp. While the Quakers were present in England and North America, Mennonite sectarian colonies flourished in The Netherlands, Germany, and Russia, until their recent migration to Canada. They were highly esteemed by the rulers in former times because of their industrial skill; hence their refusal of military service was tolerated. On the whole, their "pacifism," like that of many other sectarian manifestations, remained without any radiating diffusion into the broader society.

Gandhi's nonviolent activism. The most massive, comprehensive, and historically effective example of nonviolent activism is that of the movements unchained and organized by Mahatma Gandhi. Though Gandhi was not a theorist, he gave, throughout four decades, the most precise and ideal justification for nonviolence. He collected the richest experiences about practical successes and failures and designed a systematization of nonviolent campaigns, in which the methods of noncooperation and civil disobedience were conceptually and practically differentiated and graded. The religious philosophy of nonviolence has stronger Western roots than those of which Gandhi himself was conscious—in the examples, for instance, of the Puritan sects, the British spiritualists, the Quakers, the British suffragette movement, the Irish boycott practices (Sinn Féin), and in the reading of the Bible (especially the Sermon on the Mount), and such authors as Henry Thoreau, John Ruskin, and Tolstoy—although Gandhi also joined in with the methods of the Indian *dharma sitting* and the Islāmic *hijrah*. The expression *satyāgraha* (the force that is born of truth and love) was coined by him in an attempt to tie the practice of nonviolent procedures to ethical principles. While speaking as a rigorist, Gandhi declared that nonviolent actions were worthless when not related to the sentiment of love for one's enemy; but, while acting as a political practitioner, he was principally satisfied with effective action. He always attempted to control the campaigns of noncooperation and civil disobedience in their organizational phase and repeatedly interrupted them when violence occurred. Limited political goals were connected with the general goal of educating and emancipating the Indian masses for their political independence. Inner freedom and salvation were seen together with the economic deliverance from imported goods, emancipation of the untouchables, rural self-sufficiency, and political self-determination. Gandhi attempted to act in exemplary fashion as a charismatic leader—authoritative and, in practice, often authoritarian. His many visible symbolic actions, including numerous hunger strikes "unto death," were demonstrative actions employed to enforce certain political compromises. The fascinating effect that Gandhi had on his disciples and on a worldwide public of Western propagandists lay, in part, in its presentation of a doctrine that was at the same time an order for action and, in part, in its connection with mysticism and activism. This doctrine offered salvation (*mokṣa*) but simultaneously was interpreted as a revolutionary "direct action" in the sense of the French Syndicalist Georges Sorel, who envisioned an uncompromising elite who would lead the proletariat—with the only difference being that the action should be nonviolent, in contrast with the examples set in the West. Gandhi was not a pacifist in the strict sense of the word. During World War I, for example, he agreed to send Indian troops away to the European seats of war, and in World War II he agreed to set up an armed resistance in case the Japanese threatened the Indian borders.

Civil disobedience and *satyāgraha*

As far as the political effectiveness of nonviolent campaigns is concerned, it is necessary to distinguish between the limited results of single actions and the analysis of the historical results of whole chains of actions with far-reaching goals. Gandhi's single actions were often rather successful, at least as long as the British had not yet learned to adjust to these unexpected methods. The campaigns, as a whole, that were led in India from 1919–20 are to be judged with respect to the end result: the independence of India in 1947. Yet this result was not exclusively nor even predominantly produced by nonviolent actions. Since it was combined with the partition into two states, India and Pakistan, and since one of Gandhi's chief goals had been to reconcile Hindus and Muslims, this division by religions signified a serious failure. The separation of India from Great Britain was also hastened, in fact, by World War II and came at a time of general decolonization all over the world.

The pariah ethic

A more important aspect of Gandhi's policies was his identification with the role of the untouchables, such as those of the low-caste pariahs, which was part of his tendency to confound political emancipation with personal salvation and thus to make external politics an emanation from the activities of the soul. He declared that Great Britain's subjugation of India was a punishment for the treatment accorded by orthodox Hindu society, with their rigorous taboos, to the untouchables and that consequently the liberation of the untouchables (whose name he changed to Harijan, God's Children) would be a condition and a symbol for the national liberation of India. This identification illuminates the ethic of nonviolence, which is (in the words of the German sociologist Max Weber) a "pariah ethic"; for, notwithstanding Gandhi's assurances to the contrary, it is a weapon of the weak. It appeals to the situation of the underdog and the rejected while it cultivates a philosophy of suffering. It attempts to overpower the rulers by patience, love, conversion, transformation of hearts, and (speaking in modern terms) change of consciousness, especially by bewitching the adversary with a feeling of guilt or wrong. In the case of Gandhi, this was all highly sublimated, but this has not always been so. Nevertheless, these features of the pariah ethic explain, for the most part, the enormous response that the exemplary effect of nonviolent movements inspired in many countries, especially in those countries of the Third World, in which the social situation of the underdog or the "external proletariat" was similar to that of India and therefore could evoke a psychic situation similar to that of Gandhi's South Africa and India. Consequently, nonviolent direct action is a watchword for many spokesmen and demagogues of the Third World. The readiness for action heightens also the consciousness of the groups and accomplishes a turn of the pariah consciousness into the consciousness of the mission of an elite.

Gandhi's subsequent movements could not all maintain the unity of political and religious purposes. While official India built a myth around the name of Gandhi, his actual aftereffect was politically largely neutralized.

U.S. civil rights and other movements. Besides the smaller movements of Albert Luthuli, a South African tribal chief who was awarded the Nobel Peace Prize in 1960, and of Danilo Dolci in Sicily, the civil rights movement in the U.S. became the most important manifestation of nonviolent principles. Its best known leader, Martin Luther King, Jr., was strongly influenced by the doctrine and example of Gandhi. Nevertheless, the civil rights movement is grounded in a more general tradition in which the impulse of Protestant sects, especially the pacifist Quakers, and the works of the romantic nature writer Henry Thoreau and other nonconformists were just as effective as the religious ecstatic spirit of the Negro churches and sects and the pariah-elite consciousness developed among these groups, which is connected with their conviction that they are the true apostles of Christ on this earth. King was also influenced by the semipacifist and critical pacifist doctrines of A.J. Muste, of Walter Rauschenbusch, a proponent of the social gospel, of the moral philosopher Reinhold Niebuhr, and of Mordecai

Influences on Martin Luther King, Jr.

Johnson, formerly president of Howard University, who was the first one to refer him to Gandhi. The whole movement favoured, like Gandhi, symbolic actions—walks, marches, and pilgrimages; the millenaristic spirit of a departure was always nourished. Yet the element of demonstration was undoubtedly a more important factor in the public eye for awakening the democratic consciousness, even when the immediate purpose of the demonstrative actions was not always attained; the effects of the civil rights movement are, in any case, obvious. If, on the other hand, the symbolic character of actions is overly accentuated, the danger exists that the political goals will be defined at the outset by the limit of what can be attained by nonviolent methods. (This objection was posed against Gandhi by Jawaharlal Nehru, at the time of their joint nonviolent campaigns.)

In recent history there have been actions of passive resistance or obstructionism in Western countries completely without any foundation in religion, philosophy of life, or ideology. The so-called Ruhr Battle against the French occupation in Germany in 1923 was a failure. In contrast, a certain success has been attributed to the Fabian tactics of some population groups in the Western countries (The Netherlands, Denmark, Norway) that were occupied by German troops during World War II. Nonviolent resistance against the massive terror of totalitarian regimes (the German Democratic Republic, 1953; Hungary, 1956; Czechoslovakia, 1968) has been completely ineffective or impossible from the beginning. When Gandhi recommended *satyāgraha* to the Jews in Hitler's Germany in the '30s, this proposal was rejected with good reason by Martin Buber, a Jewish Existential philosopher. In cases of intelligent application, however, the nonviolent method can function if there is at least still some remnant of a joint base of reciprocally recognized game-rules—as was the case with Gandhi and the British bargainers in South Africa and India.

Tendency to turn to violence. Many movements that begin with a program of nonviolence and an irenic ideal turn to violent actions in the course of events. This fact can be established in respect to nonviolent movements in their end results as well as to the processes of single actions. Some nativistic movements, as, for example, those of certain messianic North American Indian tribes within the complex of the ghost-dance movements (which envisioned an apocalyptic cataclysm at which the dead would rise to announce the vanishing of the white man and the regeneration of pristine Indian life) started with pacifistic ideals but later became violent. The same can be observed among nativistic movements in New Guinea This was seen, too, in the case of the dualistic sects of the Bogomils and Cathari that arose in the Balkans after AD 1000, in some messianic Islāmic Mahdist movements dedicated to restoring primitive Islām, and in Gandhi's and King's campaigns.

Specific instances of the turn to violence

The turn to violence is partly caused by the reaction of the surrounding world or the adversary; but there may also be an inner trend or real dialectic in the movement of nonviolence into violence. Since many leaders of nonviolent movements are pure idealists fostering the conviction that "from good only good can arise" as "from evil only evil," they may be unable to see through a dynamic that conjures evil actions from good intentions. The idea of a "last war to end all wars"—*i.e.*, eschatological pacifism—is another great allurement; and, besides, it is difficult to keep a great multitude of people in an attitude of mere adventism when at the same time great hopes are being generated in them. It is also seldom possible to discipline the members within a nonviolent movement so that it remains throughout as mild as it was at the beginning: some leaders lose the helm while it is in their hands and cannot prevent bloody transgressions; others are manoeuvred out by a second set of leaders who take the helm and again precipitate bloody actions; and in still other cases the leader himself, at first a partisan of principles and a prophet of nonviolence, may turn into a millenarian prophet and apostle of terror. This type of leader then throws off his Gandhism and returns to the position of the authentic ideologists of violent direct action, as Georges Sorel and, presently, Frantz Fanon and the Existentialist philosopher Jean-Paul Sartre. This development is especially confirmed in the case of demagogues of the revolutionary rebellion in the Third World. There are smooth transitions from nonviolent resistance (with the use of pacifist language) to civilian defense (rural and urban guerrilla warfare) and to revolutionary people's war. This seems particularly to be a development in the poor countries and indeed a promising one—from the viewpoint of the revolutionary activists—because the combination is simple, whereas a modern military machine set against it is vulnerable. In recent times some intellectuals have also discussed the possibility of employing this new method of combat in the industrial countries. They urge that the nonviolent, political, psychological, civil, and social elements of the guerrilla war be maximized and the militaristic and territorial elements be minimized. In such a process the quantitative process of change in the guerrilla war turns at a certain point into a qualitative change, giving rise to the new quality of social defense, of which the active supporter is no longer some sort of partial organization but the total civil population.

CRITICAL EVALUATION OF THE MOVEMENTS

Ethical questions. Critics often issue warnings against simplistic views of the ethics of peace and conflict:

Relevance, scope, and misuse of ethics

1. Some critics have argued that it is difficult to correlate the practice of a certain peaceful and nonviolent behaviour with any definite, ethically positive viewpoint (as Gandhi, King, and all ethical partisans desire). As experience teaches, the results and effects of nonviolent methods are independent of their foundation in any religion or view of life; nor is there any guarantee that the representatives of peaceful ideals will personally represent these ideals. Thus pacifism and nonviolence are, from the ethical standpoint, neutral.

2. The advocates of pacifism and nonviolence, it is said, take the change from individual to collective (or social) ethics too lightly, or they do not see at all the problem of this change. According to Max Weber, this is the main problem of the ethics of the Sermon on the Mount. Thoreau acknowledged that only an individual, not a collective body, has a conscience, but he thought that this inconsistency could be resolved by merely summing up sheer individuals with a conscience. But this is a short circuit, since an organized society includes not only idealistic and ethically notable activists but also half-hearted fellow-travellers, partially resisting, or half- or totally-compelled members; in short, a sum of persons with different motives. In the actual war, guilty people and nonguilty people can be at the same front. Reinhold Niebuhr, an American moralist and theologian, has criticized as naive the democratic belief in the identity of personal and general interest.

3. Critics have charged that the methods of making a display of suffering that put an adversary in a morally bad light pose ethical problems. The formal quiet submission and suffering is supposed to evoke a readiness in the adversary to lay down his "false convictions" and be converted, an intention that is different only by degrees from the methods of brainwashing. The renunciation of physical means of combat and the moralizing of combat do not make it nonviolent but possibly only more malicious. Besides, it is difficult to draw a line between morally allowable pressure and unallowable moral extortion. The tendency to moral destruction can take the place of physical destruction. Particularly suspect is the spiritualizing of combat that turns a normal conflict, as King has said, into a combat of the powers of light against the powers of darkness instead of acknowledging the reality of the objects of group conflicts. Because of this interspersion of religious terminology, the adversary is demonized and is treated worse than he was before. Thus, a religiously founded dualism does not moderate the combat; it justifies, in principle, the total liquidation of the adversary.

4. As highly as a man may appraise the value of peace, it is wrong, according to many moralists, to make this value absolute; peace may not necessarily be a positive value in every case. As Reinhold Niebuhr showed, there

Peace as a relative value

is no peace without coercion, but there are different degrees of coercion. In nations that are pacified by too much coercion, the value of peace is not unconditionally negative, but it is disputable. There can be a "peace of the grave," which is of a dubious value. No one knows how much suffering, misery, and frustration have been accumulated among the people under despotism and dictatorships, both in the past and the present, who were condemned to remain "peacefully" still and silent. Man's world is one of conflicting values, not only between different cultures but also within the same culture. Thus, a relevant political ethic may not succumb to what Nicolai Hartmann, a dominant philosopher in Germany between the world wars, called the "tyranny of values." This ethic, therefore, may not absolutize the value of peace but has to counterbalance it with other values—in particular, with the value of personal freedom.

Pragmatic and other questions. Some critics have argued that the proponents of pacifism and nonviolence have not faced up to the realities of the human situation:

1. It has been claimed that man sometimes overestimates the goodness of human nature and overlooks the fact that an element of coercion exists in every society, that there are gradations of this coercion, and that it is therefore a question not of abolishing coercion but of favouring the gradations that are felt the least. In this view, peace is, as Niebuhr argued, always a coerced peace. Coercion, as it is, is ethically neutral and can be used either to commit or to suppress injustice. This problem cannot be precluded by moralistic claims, appeals to reason and conscience, and frontal attacks on prejudices—approaches that are useless and even dangerous because they arouse the illusion that something is being done, whereas in reality everything remains the same.

2. The demarcation of nonviolence from violence is clearly debatable. As a rule, nonviolence is understood merely as avoidance of physical attacks on an adversary. On the other hand, as Weber has charged, the theoreticians of nonviolence have overlooked the unintended indirect results of formal nonviolent actions, which can be more destructive than physical violence (among which are the destruction of life and property as a consequence of a commercial strike or boycott or the interruption of food supplies or medical care). According to this view, the supporters of nonviolent methods simply cover themselves with the language, the outer form, and visibility of nondamaging behaviour; they use a peaceful and passive terminology to achieve results that in fact are as destructive "as an earthquake" (as King himself has expressed it); and they connect a negative action that has destructive effects with a semantics of suffering.

3. It is not certain that human aggression can be abolished merely by the suppression of violent impulses, even if it is connected with a peaceful and friendly terminology. In India, the decades of Gandhi's nonviolent campaigns were followed by a wave of bloody reciprocal mass persecution of Hindus and Muslims that took the lives of hundreds of thousands. Some authors, in fact, have asserted that to intentionally dam aggression could cause even more violent outbursts of aggression in the future, a thesis that is still under discussion.

Depth of human differences

4. The spokesmen of pacifism and nonviolent movements, it is claimed, often fail to comprehend the depth of the gap that divides human groups through religion, nation, caste, and class. Consequently, they fail to grasp correctly the reality of the group conflicts, which they therefore overlook or consider more tractable than they actually are. When Nicholas of Cusa dreamed of a reconciliation of the different religions of mankind, he had no idea of the differences and the antitheses in the doctrines and in the existential sentiments of these religions. The same simplistic approach was taken by many optimistic, pacifistic, and enlightened men in later centuries. The "mankind" that was to be peacefully united was actually an unreal concept for many centuries. It was not until the 19th and 20th centuries, with their researches in comparative religions and cultural sciences, ethnology, psychology, and sociology, that it became possible to look into the depth of the differences in motivations and senses of life in various nations and societies. Conflict theories that are based on this research are still in process. Many of the attempts that have so far been made to account for group conflicts out of prejudices and to do away with these prejudices through better adjustments, edicts of toleration, and attempts to persuade may not take the reality of these conflicts seriously. Pacifistic proposals, such as the improvement of communication and information, that are suggested by many theoreticians cannot really solve such conflicts, but they can create new ones. Many forms of nonresistance and nonviolence, too, are more capable of miscarrying group conflicts than of solving them.

Besides, the attitude toward the problems of war and peace, pacifism, and violence and nonviolence might depend largely upon the personal perspective and group membership of the critic. It makes a difference whether the critic is personally involved in such problems in his own country or whether he sees them only from a distance. The activist, the simple follower, the resister, and the armchair theoretician all have different perspectives. It is often those who have not experienced dictatorship who preach nonviolent resistance against it.

Semantics of peace and war

5. The most difficult problem lies in the manipulation of the language. It has been pointed out that a falsification of peaceful language is usually established in the dynamics of war and peace itself and particularly in the course and conception of "pacification." Since peace applies to "good" and war to "evil," the political powers like to adopt an irenic and pacifistic language. They attempt to usurp the role of the attacked party in the case of conflicts by making an effort to interpret their actions as peaceful in the eyes of the world, even when they do (or have in mind) the opposite. In modern times this practice has improved through dialectic and has caused a babelized bewilderment, not only in the case of listeners and readers but also in the case of the speakers of that language themselves, who no longer know when they deceive others or when they deceive themselves. The language no longer has the function only of communication but has also that of an ideological means of combat. One and the same action can be an action of peace or an action of war, depending upon whether one or the other of the combatants is speaking. The meanings of words are no longer firmly fixed, but have become more relative in respect to the ingroup or outgroup. One can talk of peaceful coexistence when military aggression is planned and is being prepared. The nonviolent terminology is just as popular as this instrumental pacifism. A revolutionary programmatic can be draped as peace research, guerrilla war can be called civilian defense. Moreover, perfectionist pacifists are encountered who define peace so broadly that it includes social justifications that could be made to sanction armed intervention in every country. This ideological usage has evoked great uncertainty among the people.

PERSPECTIVES ON THE PROBLEM OF VIOLENCE

The technical development in the violent means of warfare had reached such a point in the decades prior to 1970 that no political goal is imaginable that would be commensurate with the potential for mutual annihilation. It does not necessarily follow, however, that a turn in human reason for the abandonment of all wars will now occur; but there exists only a certain probability that the means employed to solve international conflicts can remain below these potentials. The conflicts themselves may remain for as long as the national interests that can be conceived as realities exist; and these may prevail for as long as the idea of national sovereignty exists—an idea that, since the historical phase of decolonization, has expanded to include ever more political units.

Idealism, meliorism, and realism

Historically, there are no noticeable signs that the practice of violence has diminished among mankind. There were some short periods during Gandhi's nonviolent revolution in the '20s and the civil rights movement in the '60s when some enthusiastic people and idealists believed that at last the right method had been found to bring the settlement of group conflicts to a greater perfection. But the further development of these and other movements

did not encourage such hopes. In recent times, as violent acts within the nation's borders seem to be increasing, some observers claim to discern a certain tendency for violence to shift from foreign to domestic politics. This tendency provides no encouragement for the pacifists, however, who cannot be satisfied by the substitution of civil for foreign wars.

It may be that the perspective for pacifism would become more encouraging if it would renounce the perfectionist illusions of a positive peace and be satisfied, in a realistic way, with the intelligent planning of possibilities for a negative peace; *i.e.*, for the greatest possible reduction in the extermination of human life, the infliction of suffering, and the destruction of property. The problems that man now faces and will face in the following decades are overpopulation of the earth and pollution of the surrounding world. These problems will demand the total power of mankind to adapt plans to the remaining possibilities of the closed system that is the earth. The program calls for a hierarchy of planning, of which the reduction and moderation of armed conflicts (on the national and international planes) are an important point, but only one among others. What man can do for peace is evaluated through a consciousness of the limitations that are drawn for human endeavour. These limitations are more restrictive than most enthusiasts and idealists have assumed. The discoveries of individual and social psychology, of anthropology and sociology, are interpreted by many as demonstrating that man is not as free as the enlightened Rationalists had believed since the 18th century. The terminology of modern futurologists and decision-makers seems to suggest these limitations as well. The idea of a maximum peace that connects the abandonment of armed force with the demand in every country for the creation of institutions for social justice resuscitates, in truth, ancient millenarian hopes for an eternal peace combined with an empire of harmony and justice. In fact, this idea may not guarantee peace so much as it endangers peace to the extent that it calls for an intervention in every place where this combination does not exist as envisioned. If a maximum or integral peace cannot be attained, it is at least conceivable that, in many cases, by intelligent planning, a minimal peace can be reached, that armed conflicts can be localized and kept on a small scale, that the grossest mistreatment of civilians can be prevented, and that the belligerents can find their way back to the conclusion of peace.

BIBLIOGRAPHY. E.L. ALLEN, FRANCIS E. POLLARD, and G.A. SUTHERLAND, *The Case for Pacifism and Conscientious Objection* (1946); C.F. ANDREWS, *Mahatma Gandhi's Ideas* (1929), a classic work on Gandhi's notion of *satyāgraha;* HANNAH ARENDT, *On Violence* (1970), no pacifist book, but a significant analysis of violence; RAYMOND ARON, *Paix et guerre entre les nations*, 3rd ed. (1962; Eng. trans., *Peace and War*, 1966), a recent French contribution by one of France's most famous authors; HANS ECKEHARD BAHR (comp.), *Weltfrieden und Revolution* (1968); FRITZ BAMMEL, *Die Religionen der Welt und der Friede auf Erden* (1957), a good contribution by a Swiss Protestant theologian; C.J. CADOUX, *The Early Christian Attitude to War* (1919, reissued 1940), an important historical analysis; *Christian Pacifism Re-Examined* (1940); APRIL CARTER, DAVID HOGGETT, and ADAM ROBERTS, *Non-Violent Action, Theory and Practice: A Selected Bibliography* (1966); EMERIC CRUCE, *The New Cyneas*, ed. with an introduction and translated into English from the original French text by 1623 by T.W. BALCH (bilingual edition, 1909), an early specimen of the idea of a "league of nations"; TED DUNN (ed.), *Alternatives to War and Violence: A Search* (1963), contains Gene Sharp's "Facing Totalitarianism Without War"; G.L. DUPRAT, *Sociologie de la guerre et de la paix* (1932), considering its early date, not a bad treatise; THEODOR EBERT, *Gewaltfreier Aufstand* (1968), a study of the development of the theory of nonviolence and "democratic people's war"; FRANTZ FANON, *Les Damnés de la terre* (1961; Eng. trans., *The Damned* and *The Wretched of the Earth*, both 1963), the so-called communistic manifesto of the anti-colonialistic revolution, written by an ardent advocate of violence; G.C. FIELD, *Pacifism and Conscientious Objection* (1945), still a good study; CARL JOACHIM FRIEDRICH, *Inevitable Peace* (1948), the most important published study of Kant's celebrated 1795 tract on *Zum ewigen Frieden* (*Perpetual Peace*); ERICH FROMM, *May Man Prevail?* (1961), a socialist-pacifist treatise by a world-famed psychoanalyst; J.F.C. FULLER, *The Conduct of War, 1789–1961* (1961), an historical-critical account by a British military specialist; JOHN GALTUNG, "Violence, Peace and Peace Research," *Journal of Peace Research*, 6:169–191 (1969), the treatise of a prominent Norwegian protagonist of peace research; GANDHI, *Non-Violence in Peace and War*, 3rd ed., 2 vol. (1948–60), Gandhi's classic treatment of the subject; ROBERT GINSBERG (ed.), *The Critique of War* (1969), contemporary philosophical explorations by 18 scholars; E. GLOVER, *War, Sadism and Pacifism*, 2 vol. (1933, reprinted 1947); RICHARD GREGG, *The Power of Nonviolence*, 2nd rev. ed. (1966); ADOLF VON HARNACK, *Militia Christi: Die christliche Religion und der Soldatenstand in den ersten drei Jahrhunderten* (1905), a competent treatment by a renowned German theologian; GERALD HEARD, *The New Pacifism* (1936), far from passé, although written between the two world wars; ALDOUS HUXLEY (ed.), *An Encyclopaedia of Pacifism* (1937), the only encyclopaedia on this subject in the English language; RUFUS JONES, *The Faith and Practice of the Quakers* (1927), written by perhaps the most outstanding American Quaker of the 20th century; and (ed.), *The Church, the Gospel and War* (1948); TOYOHIKO KAGAWA, *Love, the Law of Life* (Eng. trans. 1929), a great book by the Gandhi of Japan; KARL KAISER, *Friedensforschung in der Bundesrepublik* (1970), contains a useful catalog of peace research institutions and a bibliography on science and peace; IMMANUEL KANT, *Zum ewigen Frieden* (1795; Eng. trans., *Perpetual Peace*, 1939), a universally recognized classic, though written almost two centuries ago; MARTIN LUTHER KING, *Stride Toward Freedom* (1958), by the American black martyr who won the Nobel Peace Prize; J. LEWIS, *The Case Against Pacifism* (1940); WILLIAM BROSS LLOYD, *Waging Peace, the Swiss Experience* (1958), an expedient national illustration; ALBERT LUTHULI, *Let My People Go: An Autobiography* (1962), by a South African poet who won the Nobel Prize for Literature; STAUGHTON LYND (ed.), *Nonviolence in America: A Documentary History* (1966), thus far the only existing American documentary history of nonviolence; G.H.C. MACGREGOR, *New Testament Basis of Pacifism*, rev. ed. (1960); DAVID MARTIN, *Pacifism: An Historical and Sociological Study* (1965); PETER MAYER (ed.), *The Pacifist Conscience* (1966); WILLIAM ROBERT MILLER, *Nonviolence: A Christian Interpretation* (1964); WILHELM EMIL MUHLMANN, *Mahatma Gandhi: Der Mann, sein Werk und seine Wirkung* (1950); *Gandhi and the Problem of Passive Resistance* (1968); A.J. MUSTE, *Not by Might* (1947), a powerful plea by one of America's most respected pacifists; JAYA PRAKASH NARAYAN, *Socialism, Sarvodaya and Democracy* (1964), selected writings by Gandhi's most prominent ideological and practical successor in India; REINHOLD NIEBUHR, *Moral Man and Immoral Society* (1932, reprinted 1960), a painstaking analysis by the famous American theologian; CHARLES E. RAVEN, *The Theological Basis of Christian Pacifism* (1951); P.O.P. REGAMEY, *Non-violence et conscience chrétienne* (1958; Eng. trans., *Non-Violence and the Christian Conscience*, 1966), an unusually perceptive analysis; LEYTON RICHARDS, *Christian Pacifism After Two World Wars* (1948); E.A. RYAN, "The Rejection of Military Service by the Early Christians," *Theological Studies*, 13:1–32 (1952), an excellent historical exposition; ABBE CASTEL SAINT-PIERRE, *Projet pour rendre la paix perpétuelle en Europe* (1713; Eng. trans., *A Project for Settling an Everlasting Peace in Europe*, 1714) and *Projet de traité pour rendre la paix perpétuelle entre les souverains chrétiens* (1717), two interesting treatises from the early 18th century demanding a sort of league of European nations; MAX SCHELER, *Die Idee des Friedens und der Pazifismus* (1931), a posthumously published treatise by a great German Phenomenological philosopher; GEORGES SOREL, *Réflexions sur la violence*, 3rd ed. (1912; Eng. trans., *Reflections on Violence*, 1914, reprinted 1941), a pre-World War I treatise by the great French syndicalist author; S. RUDOLF STEINMETZ, *Soziologie des Krieges* (1929), a realistic treatment by a famous Dutch sociologist—still pertinent today; WILLIAM TEMPLE, *A Conditional Justification of War* (1940); RALPH T. TEMPLIN, *Democracy and Nonviolence* (1965); NORMAN M. THOMAS, *The Conscientious Objector in America* (1923), the first major study of this subject in the United States; LEO TOLSTOY, *The Law of Love and the Law of Violence* (Eng. trans. 1948, reprinted 1970), a classic by the world-famed Russian author; ARTHUR and LILA WEINBERG (eds.), *Instead of Violence: Writings by the Great Advocates of Peace and Nonviolence Throughout History* (1963); CARL F. VON WEIZSACKER (ed.), *Kriegsfolgen und Kriegsverhütung* (1971), an analysis of the consequences as well as of the prevention of war by a contemporary German scholar; L.S. WITTNER, *Rebels Against War: the American Peace Movement, 1941–1960* (1969); QUINCY WRIGHT, *A Study of War*, 2nd ed., 2 vol. (1965), a comprehensive historical treatment of the subject.

(W.E.Mü.)

Packaging

From the beginning of commerce, packaging has been indispensable in the movement of many kinds of products. Animal skins, baskets woven from reeds, and earthenware vessels may be considered the packages of prehistoric man. The ancient world contributed glass bottles, clay amphorae, and leather bags. The cask was probably an invention of the Middle Ages. But it was not until the Industrial Revolution, which created a need for packaging great numbers of similar items for shipping, that the packaging industry became economically important.

Virtually all modern manufactured and processed goods require packaging at some stage in their production and distribution. Fresh foods need the protection and convenience that packaging gives. Specialized knowledge and skills, as well as specific machinery and facilities, are required to produce packages that meet one or more of five basic demands: protection from the environment; containment as a handleable unit; machine performance in the packaging process (such as on filling machines); communication to identify contents and to aid in marketing; and convenience to everyone concerned with the making, distribution, and use of the product; in addition, disposal of the package must be easy.

Table 1: Cost of Packages in Relation to Product Cost

product	percent of packages on the market costing		
	less than 3 percent of product cost	3–10 percent of product cost	more than 10 percent of product cost
Sugar	100	—	—
Margarine	—	100	—
Butter	38	25	37
Flour	—	50	50
Tea	33	17	50
Cocoa	—	50	50
Biscuits	—	50	50
Chocolate	33	—	67
Toilet soap	18	41	41
Detergents	4	65	31
Tobacco	17	50	33
Cement	6	79	15
Electric light bulbs	23	40	37

These basic requirements must be provided at a cost related to the selling price of the goods. Apart from certain luxury items such as cosmetics and perfumes, packaging is not an essential part of what is to be sold but only the means of conveying it. Consequently, packaging costs must be kept to the minimum necessary to do the job required. Table 1 lists approximate packaging costs in relation to typical product costs. The labour content and the machinery required are not included. For some products, *e.g.*, sugar, the packaging is at a minimum; for others (*e.g.*, butter) there is great variability according to the job the package has to do.

PACKAGING MATERIALS: THEIR CHARACTERISTICS AND MAJOR APPLICATIONS

The main packaging materials and their share of the total packaging used are shown in Table 2.

Table 2: Main Packaging Materials and Their Share of Total Packaging Used

material	percent of total value
Paper and paperboard	45
Metal	25
Plastics, moldings, and film (including cellulose)	12½
Glass	8½
Wood	3½
Textiles	2½
Miscellaneous	3

These figures relate to the United Kingdom in 1969, but similar proportions of the market will be applicable in almost all the developed countries.

The forms in which the main packaging materials are employed are shown in Table 3, which also gives some idea of the wide variety available.

Shipping containers. The principal shipping-container materials are wood, fibreboard, and metal. Glass is used for carboys of corrosive liquids such as acids, etchants, and other chemicals, but the quantities are small. Usually wood is used whenever the package is large or the product is of high density. Thus timber cases and crates are used extensively for weights above 100 kilograms (220 pounds) while below this weight fibreboard, both solid and corrugated, is the favoured material. Timber is also used for casks for wine and beer, but there is a trend toward its replacement by metal (stainless steel or aluminum) either alone or with inner liners of plastic.

Table 3: Applications of Various Packaging Materials

material	package type
Paper and board	paper wrappers, sacks, labels, etc.; fibreboard cases; boxes and folding cartons; paper bags and carrier bags
Metal	tinplate containers (cans and boxes); aluminum foil; aerosols; collapsible tubes; steel drums, boxes and crates
Glass	bottles; jars; carboys
Plastics	films; bottles, pots, jars, etc.; thermoformed trays, etc., from plain or expanded sheet; cushioning materials
Wood	boxes and crates; casks and kegs; pallets and containers
Textiles	sacks and bags; bales

There is some use of plastic for shipping containers, and these are returnable because the basic material cost is high. Plastic crates are well established in the dairy industry and are spreading to the more exacting use for bottled beers, minerals, and soft drinks. Beer crates are usually stacked higher and for longer periods than those for milk, and polypropylene is, therefore, preferred for beer, while high-density polythene has proved adequate for milk crates. Polythene is also used for carboys for many liquids, eliminating the protection required for glass. Some polythene casks (barrels) are also used.

Advantages of plastic crates

Expanded polystyrene is employed as a shipping container for tomatoes and grapes and also for both cured and fresh fish; the heat insulation properties are useful in keeping the product cold with a minimum of solid coolant. Solid and corrugated fibreboard cases are probably the most widely used, convenient, and economical shipping containers.

Solid fibreboard is made of paperboard (often waste pulp board) lined on one or both sides with kraft (strong paper, usually brown, made from sulfate pulp) or similar material of between 0.1 and 0.3 millimetres thick. The total thickness of the combined board ranges from 1 to 3 millimetres.

Corrugated fibreboard includes both double-faced board and double-wall board. Double-faced board consists of two flat sheets separated by a fluted sheet made from straw or a special hardwood pulp; different grades and types of board can be produced by varying the materials, the thickness, and the weight of the liners and the fluting medium. Double-wall board is made from two fluted sheets separated by a flat sheet and faced on both sides with a further flat sheet. Flute configurations vary in number of corrugations to the metre and hence in thickness of the combined board.

The conventional case is the one-piece, or regular slotted container, although open-tray and wrap-around styles are used extensively. The normal range of weight that corrugated and solid fibreboard cases can carry lies between five and 20 kilograms, but fibreboard cases can be made to hold loads of up to 50 kilograms (110 pounds) without any special fittings being used, and reinforced containers are capable of carrying loads of powdered or granular material of up to 500 kilograms.

In the 1960s important innovations in the movement of goods appeared, including palletization, modular packaging, and freight containers. In some instances the pallet

has superseded the wooden case or crate, and in others the shipping container has been eliminated since the goods are loaded directly into a freight container and secured. This is particularly applicable to heavy machinery.

Retail containers. Folding cartons and paperboard boxes are used extensively in the food industry (about 45 percent of all types). The second largest use is in the chemical field, particularly for pharmaceuticals, cosmetics, and detergents (about 20 percent). The cigarette and tobacco industry account for the third largest use, and the remaining use is divided among electrical equipment, clothing, toys, and other goods. In most instances where little climatic protection is needed, a carton provides the cheapest and best physical protection against crushing; it can be supplemented with a barrier against moisture by adding an inner film bag or an outer wrapper. Despite advances in printing technology for other materials, paperboard and paper remain more easily and cheaply printed to high graphic standards than most of their competitors.

Basic shapes: tube and tray

Cartons are made in a great variety of shapes and sizes. Originally cut and creased on converted printing presses, they are now produced on more efficient and sophisticated machinery. All of the hundreds of different types and styles of cartons in use are descended from two primary constructions: the tube and the tray.

A tubular carton consists of a sheet of paperboard folded over and glued at the edges to produce a rectangular tube, the ends of which can be closed or locked together. A tray-type carton consists of a sheet of board all of whose edges, usually four, have been folded up at right angles to the main sheet and secured at the corners. The tray may, in fact, be joined along one edge with another sheet of board similarly folded to form a lid. The most commonly used style of tubular carton is the glue-end carton. This is particularly suitable for automatic packaging of powdered materials, including foods and detergents. The tuck-in style of carton is often used if the package needs to be reclosed after it has been opened for the first time.

The main advantage of the tray style of carton, whether fitted with a lid or not, is the greater area initially available for filling the box; trays are preferred for applications in which the contents would be difficult to load through the narrow ends of a tube. They are used widely for biscuit assortments, cakes, and pies and are easily adapted to both semi-automatic and fully automatic packaging systems.

Although most carton styles are based on folds and creases produced from straight lines and possess four major faces, it is possible to handle curved creases and shapes based on triangles or hexagons, to improve display. Complications can arise, however, in the packing of such cartons for delivery in bulk, particularly where triangular, tetrahedral, or seven-, eight-, or ten-face constructions are produced, even if these are based on regular figures.

Tinplate containers can be divided into two classes: the cylindrical open-top variety, of which some 80 to 90 percent are in use for processed-food cans or canned beverages; and general line cans that have replaceable lids; nearly half of these are also used in food packaging, about a quarter for paints and varnishes, and most of the remainder in the tobacco, medical, and cosmetic fields.

The tin can is fabricated of steel with a thin coating of tin on each side of the sheet. For some products, it is advantageous to have a thicker coating on one or both sides, or a lacquer coat to prevent the product from attacking the metal.

Three basic styles of can are the seamless-body can, the locked-corner-body can, and the rolled-body can.

The seamless, or solid-drawn can is pressed into shape and has no joints. The body is usually finished at the open mouth by trimming, beading, or curling to receive the lid, which is frequently of a slip-on variety. The tinplate can be formed into a seamless-body can of only relatively small depth, otherwise the material will tend to pucker or wrinkle during the pressing. Limitations are imposed by the sharpness of curvature of the finished container, the thickness of the container material, and the ability of the surface finish to remain anchored to the base metal. It is possible to make round, oblong, and square shapes, but beyond about 100-millimetre (4-inch) diameter the round, seamless can tends to become uneconomic and functionally poor. The locked-corner can has, essentially, the ability to form an almost square corner, permitting efficient storage.

The rolled, built-up can must have at least two components, the side wall and the base. (It may have a third, or even a fourth component.) The side wall is often referred to as the body and the base is called the bottom, or end. A two-piece body has two joints; the side seam of the wall, and the joint between this wall and the bottom or end. Both are produced by interlocking folds. The side seam is formed by interlocking the fold on each edge and "bumping" the interlock. The bottom joint is made by a process called double seaming. The common food can is produced by securing a third component, the end, by double seaming after the can has been filled.

Aluminum is used to produce seamless body cans by impact extrusion. If the can depth required is large relative to the diameter, aluminum is the only material for producing such a container, because tinplate cannot be formed to great depths. To achieve adequate rigidity, an alloy of aluminum with about 1 percent manganese is commonly used. Aluminum does not rust and is light and pleasant to handle, but its resistance to chemical attack, particularly that of alkalies, is limited. Its application must, therefore, be carefully considered in relation to the properties of the product.

Most aerosol (pressurized) containers are based on metal cans. This package must dispense the product in the right form and quantity. Hairsprays, air fresheners, insecticides, wax polishes, deodorizers, cleaners, and coatings are among the products so packaged.

Collapsible metal, and more recently, plastic tubes are used in the cosmetics, toiletry, and pharmaceutical fields, with toothpaste accounting for the highest single use.

Aluminum is used principally as foil: in a thin form as a barrier to gases and vapours of all kinds, either alone or as a component of a laminate; or in a thicker form as a lightweight container (basins, trays, dishes) for food that must be heated before consumption. It also provides bottle caps and closures and easy-open tops for cans.

Advantages of glass

Glass is a highly inert material of great cleanliness and hence is used mainly for food and drink (about 70 percent) and for drugs, cosmetics, and toiletries. A material easily formed into almost any shape, with exceptional aesthetic potential, it is widely used to package products dispensed at intervals, over a relatively long period; *e.g.*, perfumes and toilet water. Glass containers, both bottles and jars, are mainly produced on fully automatic machines at speeds up to over a hundred a minute in a continuous process; once started, the furnace containing the molten glass remains in operation for two or three years. The furnaces operate at temperatures of up to 1,500° C (2,700° F) and hold up to 400 tons of glass. The molten glass passes from a melting chamber into the working chamber of the furnace where it cools slightly before passing into the molds of the forming machines. The molten glass is then either sucked up or, more frequently, extruded into preliminary molds where an initial or "parison" shape is formed. It then passes into the second stage where the container is blown to its final shape. After forming, the hot containers are annealed in a long tunnel to eliminate undesirable stresses.

The majority of glass containers sold today are discarded when they are empty, with the exception of milk bottles, some beer bottles, and certain carbonated-drink bottles. Successful re-use of returnable bottles depends on the control of the packer over the distribution system. Most modern developments in glass-container manufacture are related to the weight of the finished container and its strength. New production methods promise to bring considerable changes in the form of containers that will be light, strong, and shaped like an electric-light bulb, with a plastic base and a simple closure.

Of the large number of plastic materials available in the

Table 4: Comparison of Properties of Common Flexible Packaging Materials*

	durability (toughness)	convertibility	printability	heat sealability	water vapour resistance	gas transmission resistance	odour resistance	water resistance	oil or grease resistance	resistance to chemicals
Polyethylene	6	6	6	6	7	3	3	10	5	10
Polyvinyl chloride	8	6	5	4	2	5	5	9	8	6
Vinylidene chloride copolymer	9	4	5	4	10	8	8	10	8	8
Polyesters	10	4	6	2	4	6	8	10	8	10
Regenerated cellulose—grade P†	2	10	10	0	0	6	6	0	10	3
Regenerated cellulose—grade MS‡	2	10	10	4	5	6	6	2	10	2
Regenerated cellulose—grade MXXT§	2	10	10	4	6	9	9	4	10	3
Polyethylene/MS cellulose laminate‡	7	8	10	8	8	8	8	6	6	8
Cellulose acetate	4	10	8	0	0	0	2	6	6	3
Rubber hydrochloride	6	6	6	6	5	5	6	10	8	6
Kraft paper	4	10	10	0	0	0	0	4	0	0
Sulfite paper	3	10	10	0	0	0	0	4	0	0
Glassine	2	8	10	0	0	3	3	2	4	2
Wet (gloss) waxed paper	3	6	10	6	5	2	2	8	5	5
Waxed glassine	2	6	10	3	4	5	4	5	6	4
Polyethylene coated paper	6	8	8	10	7	3	3	6	5	7
Vinylidene chloride copolymer coated paper	4	8	8	8	8	8	8	6	8	7
0.009 mm aluminum foil	1	4	6	0	6	7	7	10	10	5

*Qualities graded from 0–10 where 10 is the highest grading and 0 signifies that the quality is nonexistent. †P = non-moistureproof. ‡MS = nitrocellulose heat-seal coated on both sides. §MXXT = copolymer coated on both sides.

Uses of plastic in packaging

early 1970s, only a few have made a substantial impact on packaging: polyethylene (often referred to as polythene), polystyrene, polyvinyl chloride, and polypropylene. Others, such as thermosetting resins and saran, have limited applications as closures and coatings. More than half the polythene is used in film form and much is converted into shrink film, liners, sacks, and bags. Some is used in the form of bottles, a little for larger specialized containers, and the remainder for coatings or laminates. Polystyrene is principally made into tubs for ice cream, packs for eggs, sausages, and small pots or jars for butter, jam, and cheese; or used in expanded form for packaging typewriters, record players, and other delicate machinery; or formed from expanded sheet into trays for fresh foods, cameras, or other lighter weight goods needing attractive protection against shock damage. Polyvinyl chloride is used typically for bottles for soft drinks, cooking oil, and vinegar, and as trays for chocolates. The use of polypropylene is growing rapidly, especially as a transparent overwrap, in which it has strength advantages over cellulose film.

Molded thermoplastic containers are produced by blow-molding, injection molding, or thermoforming. They generally have no special names but are given the same description as the traditional containers that they emulate. The same type of container can often be made by more than one method. Thus, bottles are produced by either blowing or injection molding, and tubs can be produced by all three processes.

One of the most important problems in the early use of plastics for the food industry was the possibility of taint, but special grades of several plastics now ensure the safe packaging of a wide range of foods.

Polystyrene is the main plastic used for injection-molded containers, but polypropylene is beginning to be employed in this field. Both are available in special grades for food. Such containers are light in weight and chemically inert, and those from polystyrene can be produced in glass-clear form if required. Polystyrene and polyvinyl chloride may also be formed by vacuum into containers for a variety of foods, including fresh fruits and vegetables. Polystyrene frequently is toughened by the addition of a small percentage of synthetic rubber, allowing it to be vacuum-formed into deep containers. Other polystyrene grades with good flow properties are used for thin-walled, injection-molded tubs for such commodities as cream or yoghurt.

The thermoforming process is suitable for making trays and fittings, as well as tubs, particularly where weight, and thus cost, must be kept to a minimum. Cellulose acetate, polystyrene, polyvinyl chloride, high density polythene, and polypropylene are all thermoformed in this way, and polystyrene can be thermoformed both from clear film and, more recently, from foamed sheet. These fittings and trays find considerable application in reducing shock and separating lighter fragile items from each other. They also provide more decorative fittings than the previously used paperboard for such items as cameras, cosmetics, and toiletries.

Plastic packaging films in very large variety have considerable uses in all fields of barrier packaging (see below). There is no single material that is equally suitable for all purposes, and the most suitable film must be chosen for the greatest advantage. Information on the properties of common flexible materials is given in Table 4.

Packaging of hazardous materials. Many chemical products have obnoxious or hazardous properties. Explosives, poisonous materials, and liquids that are corrosive or produce dermatitic or skin irritant effects, inflammable goods, materials that react either with air or water, radioactive materials, and materials that are likely to cause spontaneous combustion all require special selection of materials for packaging. In general, regulations are laid down in handbooks of road, rail, and sea transport authorities. Where export is involved, it is also necessary to comply not only with the regulations of the country of destination, but also with the regulations of countries through which the goods may have to pass.

Labels must indicate the hazardous nature of the contents.

Packaging fragile articles. The selection of materials and design of cushions and fittings for fragile products is

one of the most important problems of packaging design. Frequently a high proportion of the total packing costs must be allocated to protect against shock and vibration. Since over-packing can be expensive, it is necessary to choose the right material, use it efficiently, and provide only sufficient protection for the product. To do this requires information on the fragility of the product, the cushioning materials available, and the shocks likely to be experienced in distribution.

There are, essentially, six steps that must be taken in producing a pack for a fragile article, whether it be electronic equipment or decorative glass. First, the hazards of the transport system are assessed, and the height and the number of drops to guard against decided. Second, the fragility of the product is assessed. Third, cushioning materials are selected and their positioning calculated. Fourth, the total cost of each material is determined. Fifth, a prototype package is produced using the most economical material. And sixth, the package is tested to assess defects or deficiencies. The steps are repeated until the package is satisfactory.

Barrier packages. To maintain favourable climatic environment around goods subject to deterioration from moisture, oxygen, light, or heat, any of several barrier materials are employed, such as waxed paper, metal foil, plastic films, and flexible materials listed in Table 4. Most are good barriers to moisture and moisture vapour, some are good barriers to oxygen and other atmospheric gases. The opaque materials are barriers to light, and in some instances, for relatively short periods of time, a certain degree of heat insulation can be provided for some products.

There are four principal mechanisms by which gases and vapours penetrate flexible packaging materials: closure failure; passage through the pores and channels in the material itself, especially paper; diffusion through such material as cellulose film or polythene, and transmission of gases or vapours through pinholes in metal foils, such as aluminum. The rate at which gases or vapours are transmitted by the activated diffusion process increases as the temperature rises, but the other mechanisms are hardly affected by temperature at all.

The light transmission qualities of flexible barrier materials can be controlled by their colour, by their opacity, and by their thickness. Almost all plastic films may be treated in some way to change their light-transmission capabilities, while materials such as paper are generally opaque. Light can cause problems in the packaging of foods, particularly frozen foods, if the lighting in cool and cold display cabinets is intense. Heat insulation is possible for periods of up to 48 hours. Products sensitive to temperature need to be packed in temperature-controlled vehicles, either refrigerated or insulated.

RETAIL PACKAGE DESIGN

Four major considerations are involved in package design: the product, its market, the package production problem, and the economics of the packaging operation.

Under product considerations come such questions as susceptibility to damage or deterioration, the need for protection against temperature changes, moisture changes, oxygen, light, mold, or insects, and whether the product will interact with the potential container.

Market considerations

Market considerations include the characteristics of the customer, such as age group and income level, the unit quantities required, aspects of the retail and wholesale operations, and probable methods of handling, warehousing, and distribution. The graphics considered include brand name; whether the product is part of a range; whether there are colours that may not, or must be, used; any pertinent government regulations; and the economic effect of all these factors.

Food packages. In the food industry the processes of canning and bottling tend to be differentiated from other methods of packaging. To some extent this is a distinction between methods of food processing; *e.g.*, sterilization or pasteurization as against dehydration and quick freezing, which are much more demanding with respect to packaging.

All foods are perishable, but some ripen or mature after being packed and improve for a short period. The deterioration that inevitably occurs is hastened by unfavourable environmental conditions; much of food packaging is designed to intervene between a food and its surroundings to delay the processes of deterioration beyond the time needed for transportation, marketing, and consumption.

Factors controllable by packaging that lead to spoilage include mechanical damage, changes in moisture content, interaction with oxygen, and intrusion of foreign odours. Most spoilage involves more than one of these factors; *e.g.*, mechanical damage, in addition to physically damaging the food, may permit entry of moisture, oxygen, or foreign odours.

In addition to providing a barrier against dirt and other contamination, and protecting against physical damage, moisture, oxygen, and light, a food package is designed to function smoothly, efficiently, and economically on the production line during the packaging operation, and to supply convenience not only to the consumer but during storage and distribution. Finally, it must provide identification, information, and sales appeal.

Food packaging always involves quality control, whether through government regulation or other means. Laboratory techniques are often used, followed by field trials, in which the food is packed and stored under specified conditions for specified periods. Once the basic suitability of the package is established, some method of rapidly checking incoming packaging materials and completed packages of food is provided.

The quality of many foodstuffs (as well as other products) has reached such uniformity in standards that there is often little difference among brands. Under such circumstances, convenience of the packaging becomes paramount, at least in countries where economic competition flourishes. Provision of individual portions, such as boil-in-bag preparations, without materially affecting cost, is an important marketing factor. The flexibility of pouch packages permits remixing components that separate during storage, or, alternatively, for mixing materials packed in separate compartments. Some food products can be sold dry in a flexible package to which the requisite quantity of water, or other liquid, is added, and the mixture kneaded in the pack, eliminating the need for mixing bowls.

Consumer convenience

Convenience for the consumer, once virtually limited to easy opening and reclosing, has thus broadened in scope. Convenience for the distribution chain involves such considerations as the right size and shape for outer containers, the ease of disposing of unwanted packaging materials, and the breakdown of larger packs into simpler and more convenient units. Large savings can accrue from packaging that facilitates warehousing.

Closures. Glass containers always require a closure, and closures are produced in a great number of sizes with the precision necessary to ensure complete sealing. The majority are made of metal or plastics, although paper, rubber, glass, or combinations of these are also used. Tinplate, aluminum, and aluminum alloy are the most common metals, while Bakelite, polythene, polyvinyl chloride, and polystyrene are the most widely used plastics. Both metal and plastic closures may be rigid or semirigid, sealing by compressing an interposing wad of rubber, paper, cork, or synthetic material against the edge of the glass. Designs vary widely, including external screw caps, lug caps, internal screw stoppers, lever and snap-on caps, vacuum seals, and corks. Metal-foil caps, as used for milk and other dairy products where both the product and the pouring lip of the container are protected, and special rubber closures that can be pierced by a hypodermic syringe are also used.

Metal-can closures operate by one of five basic methods: frictional engagement, screw-thread engagement, plastic closures, vacuum seals (which rely on atmospheric pressure on the lid), and permanent mechanical interlocking of lid and body. This last is the most widely used form of tin-box closure, as used in the food can. The joint, usually called a double-seam, is ideal

in its simplicity and reliability for the exacting chemical and microbiological requirements of food canning.

Plastic containers such as bottles and jars may be closed with screw caps similar to those used on glass. Rigid and hard tubes made in polystyrene can be effectively closed with injection-molded polythene plugs, sometimes ribbed to improve the seal. Another type of closure is the push-on cover, which can be made in paperboard, plastic, or foil. Vacuum-formed plastic containers are sometimes closed by a heat-sealed flat or recessed lid. The seal is commonly made on a heated jig.

Most closures of film packages rely on the heat-seal characteristics of the film. Some films, such as regenerated cellulose, cannot be sealed with heat unless they are coated with a special heat-sealable coating, usually based on a thermoplastic lacquer. Polythene and polypropylene can be sealed by electrical-resistance heating; most of the machines currently available employ electrically heated jaws or resistance tapes pressure-closed onto the film to effect the seal. Some plastic film bags are closed by tying.

Fibreboard shipping cases may be closed successfully with cold adhesives, usually based on sodium silicate, hot-melt adhesives applied by special methods, gummed paper tape, self-adhesive plastic tapes, and stapling.

Opening devices. Despite much ingenuity used in designing opening devices, results have often been unreliable. One problem is that the easier a package is to open, the greater the chance of accidental opening. Cans, particularly for beverages, received considerable attention in the 1960s, and the easy-open beer can, not requiring an opener, became highly successful in Germany, the U.K., the U.S., France, and other Western countries. Such cans have opening tabs of aluminum, which has lower resistance to tear than tinplate. The "crown cork" for bottles also eliminates the need for a special tool and makes opening possible by tearing. In the early 1970s, a number of plastic and aluminum-foil closures of this type showed promise.

Tear-tapes for packages made of film and paper, however, have often proved unreliable, either because of poor adjustment of the machine applying the tape, or because of variations in the dimensions of the packaged product. The latter has often been the case with the film-wrapped biscuit packs. Some success has been obtained in tearing devices to convert corrugated cases into display units. Considerable further development is probable in this area.

Dispensing packages. Aerosols, or pressurized containers, are a leading example of dispensing packages. Although in theory anything that can be made into a liquid or paste form can be dispensed from an aerosol, in practice only products that derive some practical advantage from spray dispensing are so packaged.

Materials for boil-in-bag packages, besides being resistant to boiling water, must have a degree of impermeability to odours and resistance to grease. High-density polythene or polypropylene films are very suitable. Another convenience package is the bake-in tray that doubles as cooking utensil, and in one form, as serving dish. Such trays are made from foil-lined paperboard, lacquered boards coated with heat-resistant plastic, or aluminum foil.

Labelling and decoration. Plain labels, applied by adhesive spread on the machine, are used for bottles, cans, and boxes of all kinds. Pre-gummed labels are precoated on the back with a water-activated adhesive just before application. Thermoplastic backed labels have a resin adhesive that is activated by heat just before applying to the container. Pressure-sensitive or self-adhesive labels have their backs precoated with an adhesive substance that is protected by a separate backing removed just before applying the label. Pressure-sensitive labels are of two types; one remains permanently bonded, and the other may be removed without marking the surface. Thermoplastic and pressure-sensitive labels are used for plastic containers because most gummed labels will not adhere to plastic.

All printing processes are used for package decoration; the choice is determined by the package material and the type of decoration required. The principal problems are odour transfer from inks, rubbing and smearing of ink, fading by light, and colour variation.

Printing on metal

Although the most widely used form of decoration for a can is a printed paper label, both tinplate and aluminum can be printed; in fact there is little difference between printing paper or board by offset lithography and printing tinplate. Paper and board, however, are absorbent, and tinplate is not. Hence each colour printed on tinplate must be baked at a high temperature to dry it and key it to the metal. Since it is also necessary to use a base coat, a six-colour printing with a final overprint varnish requires baking eight times. Similar problems are associated with fired-on print for glass bottles.

PACKAGING MACHINERY AND TECHNIQUES

Filling and sealing bottles. Bottle-filling machines for liquids can be divided into four basic types: vacuum filling, measured dosing, gravity filling, and pressure filling.

Filling by vacuum is the cleanest and most economical way to handle many products. In spite of the care that is taken in making bottles and cleaning them, there is always a percentage with holes, chips, or cracks. Vacuum-filling machines automatically avoid such bottles. Vacuum filling is neat and efficient, and it is unnecessary to wash or wipe the bottles before the labelling. The vacuum system requires a supply tank that is below the level of the bottles to be filled; from the supply tank run pipes or lines to which are attached filling nozzles. Also connected with the supply tank is an airtight overflow receptacle. When the machine is started, vacuum is created in the overflow receptacle, and in turn in the lower end of the air line or at the suction ends of the nozzles. When the bottle is pushed into close contact with the gasket of the suction nozzle, making an airtight seal, a vacuum is created in the bottle (unless the bottle is imperfect). The vacuum draws the liquid from the supply tank, through the tube and into the bottle. When the liquid reaches the end of the overflow or suction airlines, it automatically breaks the vacuum, causing a cessation of the flow of liquid into the bottle. The filling stem is then withdrawn and the bottle passes to the closure plant.

In measured dosing, the height of fill is not constant. Each filling unit consists of a calibrated cylinder and a piston. As the piston begins its down stroke a valve opens, allowing free passage of liquid into the cylinder. At the end of the stroke the cylinder is charged with a measured quantity, and when a container is correctly positioned, a delivery valve opens and the supply valve closes. The piston discharges the liquid into the container on its return stroke, and the sequence repeats.

Gravity filling

There are two types of gravity-filling machines, one of which fills on a controlled-time cycle and the other by using a measuring chamber. In the first, the presentation of the container to the filling head opens a valve that permits the liquid to flow for a predetermined time. The valve then closes and the container is taken away. The open time is determined by the viscosity of the product and the diameter of the filling orifice; control may be mechanical, by time clock, or electronic. In the second type of gravity filler, a supply valve opens to admit liquid to a calibrated chamber. When a container is correctly positioned at the filling head the supply valve closes and a delivery valve opens, thus charging the container with the measured amount.

Pressure filling is basically similar to time-cycle gravity filling. An artificial head pressure is induced to the liquid by a pump or by air pressure within a closed tank. Gravity and pressure filling are appropriate to moderately fast filling of low viscosity liquids, such as fruit juices, but they are considerably slower than vacuum filling.

Dry goods in powdered and granular form. Granules or tablets are sometimes counted by sorting equipment with electronic counters, but more usually they are simply loaded into containers by one of two basic methods, volumetric filling, and filling after weighing. The nature of the product usually determines which method is used. Volumetric filling may be by vacuum, by the

use of an auger (screw), or the use of a flask. The vacuum method is similar to that described for liquids. In auger fillers, the quantity delivered is controlled by turning the auger a certain number of cycles. And, in flask fillers, the volume of the flask determines the quantity delivered.

Weight filling is the best way to meet the requirements of weights and measures regulations. In all the different weighing techniques (scale beams, compressed-air, or electronic measurement systems), the basic principle is to divide the supply of product into a bulk feed and a fine feed. At the beginning of the weighing cycle both the bulk and the fine feed operate to fill the weighing pan until 80 or 90 percent of the required material has been added; at this point the bulk feed stops and the fine feed continues until the exact balance is reached; as the fine feed is cut off the load is discharged, usually by tipping into the container.

Cartoning systems. A cartoning system combines a special carton with the machinery to erect it from a flat condition, fill it with a product, and close it. The machinery varies from simple hand-fed machines to automatic stations coupled with means for packing the cartons directly into cases for dispatch. Whatever the system employed, three main operations are performed.

First, forming or erecting the container. Material may be fed to the carton erection point as a continuous web, as a flat carton blank, or as a folded carton flat with a manufacturer's joint secured.

Second, loading or filling the container. In the continuous web fed and the top-loading systems the container has only one face open at the moment of filling; with an end-loading system it is possible with certain products to load the carton and close both ends afterwards.

Third, closing or sealing.

In addition to these main operations cartoning systems may be required to carry out secondary operations such as handling paper liners, embossing codes, and inserting leaflets.

All these can be performed along with the three main operations on manual, semi-automatic, or fully automatic lines.

When forming, filling, and closing must be carried out on one machine in a single operation, fully automatic machinery is usually used. The division between semi-automatic or fully automatic machines relates principally to how the filling or loading of the carton is carried out. If loading is done directly into the carton, even though an operator inserts it into the infeed conveyor by hand, the system is classed as automatic. If the rest of the operation is automatic but the load is inserted directly into the carton by hand, however, the system is described as semi-automatic. Most systems requiring higher speeds use continuous-motion machines; lower-speed systems usually use intermittent-motion machines. The latter can also be of advantage where the nature of the product demands a stationary carton at the moment of filling.

Automatic and semi-automatic loading

Film forming, filling, and closing machines. These machines use a reel of flexible material (paper, film, or foil laminate) and either form it into a tube and seal and fill it at regular intervals, or fold it lengthwise, and pinch and seal it at right angles to the fold to form a series of pockets that are filled and closed.

In high-speed operations the tolerances both in the machine and the packaging material are critical. Containers may also be produced from reels of material by thermoforming a series of trays in a web, filling them with the product, and feeding another web over the open trays to cover exactly their flanges, permitting heat-seal closures. The web of filled and closed packages is either punched out to form individual packages or slit at intervals to give a number of units joined together.

Bag manufacture, filling, and closing. A bag is a prefabricated wrapper that only needs the product and closure to complete the pack. Bags have always been useful in hand-filling operations; and recent developments in automatic bagging machines, coupled with the wider choice of materials, notably polythene, have increased their value even further.

Paper bags are available in four standard styles: flat, gusseted, self-opening satchel, and rose bottom. They are made in various types and weights of paper according to the size, weight, and nature of the contents, and other functional factors. Sulfite papers, particularly the machine-glazed variety, are used for general purposes, and kraftpapers (made from sulfate pulp) are used where strength is essential. If resistance to grease or fatty foods is necessary, bags are made in greaseproof paper, vegetable parchment, or glassine, all of which can provide some resistance to loss of flavour. Paper bags are also available in wet-strength or waxed papers; such materials give resistance to the passage of water vapour if special attention is given to the closure and seams.

Film bags offer more protection than a paper bag and also provide visibility to the product. Film-fronted or windowed bags using both film and paper are also common. Films such as polyethylene, which may be extruded in tubular form, can be produced both with or without gussets and contain no seam in the length direction of the bag. The great improvement in resistance to moisture and grease of this type of bag, and its efficient closure, have led to its wide adoption for food packaging. The principal film materials used are cellulose (in plain, moisture-proof, and heat-sealing moisture-proof grades) and low-density polyethylene, although polypropylene, polyester, vinyl- and rubber-hydrochloride, and saran films are all used.

Cotton bags also have limited uses for packaging larger quantities of some products, notably foods. They are manufactured in bleached and unbleached qualities and may be printed.

Open mesh bags are frequently used to pack products such as fresh vegetables that require complete ventilation in transport and storage. In smaller quantities, up to about 2 kilograms (5 pounds) weight, polythene film bags with perforations may be used for the same produce; but above this weight, film bags lack strength. Open mesh bags were first made in hessian (burlap), but they are now produced in yarn twisted from special paper or in mesh of tough resilient plastics.

Bag-filling and closing equipment performs four operations: feeding the flat bag to the loading point; opening the bag and keeping it open; loading the product; and making the closure.

The initial opening is usually achieved by an air-blowing device, often assisted by the incorporation of lips on the bag. Mechanical fingers are also used either alone or to assist the air-opening device. The bag may be held open and the product guided into it by a scoop or other means, or, for certain products, the scoop is inserted into the bag, and the bag pulled over the product. The principle of moving the bag rather than the product enables bagging to be performed on products with little resistance to distortion or crushing, such as sliced bread. Ties, clips, staples, tapes, and heat seals are all widely employed as closures.

Bag-in-box packages are used for protection of dried and powdered foods and also for liquid packaging such as fruit juices and even wines, in quantities up to a gallon or more. Given a wrapping material inherently water-vapour resistant and heat-sealable (whether the seal is a weld or a surface seal) it is always possible to make a bag more efficient than an overwrap for a box of a given size. Moreover, many film materials are also odour-resistant and by their immediate contact can avoid locking in potential sources of odour. Cheaper grades of boxboard may be used for cartons when an inner bag is employed.

TESTING

The three main reasons for testing packages, containers, and packaging materials are: first, to predict their performance in practice, second, to control quality, and third, to plan improvements. A single test applied repetitively, or at an increasing level of intensity, can show exactly the strengths and the weaknesses of a packaging material or container.

There are two basic ways to evaluate test performance of a new package. The first is to compare it with the

known performance of a package already in existence. The second method attempts to reproduce the actual events that are experienced by the package or the material in use, and from the results deduce what will occur in practice.

The first semiscientific test was the drum tumbler, which replaced the earlier method of rolling a package down a flight of stairs. More accurate tests are now available. The principal ones attempt to reproduce the effects of the main hazards of distribution, including impacts, vibration, and compression loads (such as are experienced in stacks) as well as climatic effects, principally the effects of moisture and temperature changes. Such tests are either put together in schedules or used alone to evaluate particular aspects of performance.

BIBLIOGRAPHY. F.A. PAINE (ed.), *Fundamentals of Packaging* (1962) and *Packaging Materials and Containers* (1967), standard British textbooks; *Modern Packaging Encyclopedia* (annual), a useful condensed summary of packaging from the marketing and retail viewpoint (used as textbook material for many U.S. colleges); A.H. WOOLLEN (ed.), *Food Industries Manual*, 20th ed. (1969), a detailed British handbook for the food industries; PRINTING, PACKAGING AND ALLIED TRADES RESEARCH ASSOCIATION (PATRA), *Evaluation of Package Performance* (1963), and G.A. GORDON, D.J. HINE, and J.H. YOUNG (eds.), *Paper and Board in Packaging* (1963), two collections of lectures by specialists experienced in particular fields of packaging; C.R. OSWIN and L.N. PRESTON, *Protective Wrappings* (1965), a useful book on the flexible packaging industry that gives detailed information on product requirements, material properties, and machinery potentials.

(F.A.P.)

Pagan

The modern village of Pagan lies on the Irrawaddy River about 90 miles (145 kilometres) southwest of Mandalay in Burma. Its population in 1970 was only about 3,000. The site of an old capital city of Burma, Pagan is unique in preserving the largest extant group of examples of a superb brick-building architecture that once stood in many parts of Southeast Asia. Its structures date from between the 9th and 13th centuries. Some, which are Buddhist shrines, have remained in use to the present day, continuously restored and redecorated; the majority, however, are shells, some well preserved but others less so. Pagan's importance lies in its heritage rather than its present. It was first built probably in AD 849 and, from the 11th century to the end of the 13th, was the capital of a region roughly the size of modern Burma. Only once, *c.* 1160–65, did it fall victim to a surprise attack. In 1287 it was overrun by the Mongols in the course of their wide-ranging conquests and never recovered its position thereafter, although a little desultory building continued on Buddhist shrines.

Topography

Old Pagan was a walled city, its western flank resting on the Irrawaddy River. It was the focus of a network of highroads by means of which its rulers could command a large region of fertile plains. They were able to dominate the other major Burmese dynastic cities, such as Pegu, Prome, Taungdwingyi, Thaton, and Tagaung, while the small Shan and Arakanese kingdoms gave allegiance. From the port of Thiripyissaya, further down the river, important overseas trade was conducted with India, Ceylon, and other regions of Southeast Asia. Studies so far undertaken suggest that the walls of the old city, within which lies a substantial area of the modern town, originally contained only royal, aristocratic, religious, and administrative buildings. The populace is thought to have lived outside in homes of light construction closely resembling those occupied by the present-day Burmese. The walled city, whose moats were fed by the Irrawaddy, was thus a sacred dynastic fortress. Although it was never as large as old Śrī Kṣetra (modern Hmawza), the circuit of its walls and river frontage is some two and a half miles, and there is evidence that perhaps as much as a third of the old city has been washed away by the river. Since building was principally in brick, decoration was carried out in carved brick, in stucco, and in terra-cotta. The earliest surviving structure is probably the 10th-century Nat Hlaung Gyaung. The shrines that stand by the Sarabha Gate in the eastern wall, although later than the wall they adjoin, are also early. These are shrines of protecting *nats*—the traditional spirit deities of the animist ethnically Burmese. Pagan lies near Mount Popa, sacred to the Burmese as the dwelling of the brother and sister spirits called the Mahagiri Nats.

Settlement

Between *c.* 500 and 950, people of the Burmese ethnic group had been infiltrating from the north into a region occupied by other peoples; these people already had been converted to Indian religion, especially the Mahāyāna Buddhism of Bihār and Bengal. Under King Anawrahta (reigned 1044–77) the ethnically Burmese finally conquered the other peoples of the region, including a people called the Mon, who were previously dominant in the south. They transported the Mon royal family and their scholars and craftsmen to Pagan, making it the capital and centre of an official, fundamentalist form of Hīnayāna (Theravāda) Buddhism adopted from Ceylon (*c.* 1056). This initiated the period of Pagan's greatness, which was sustained at first by Mon artistic traditions. The construction and existence of the enormous number of monasteries and shrines built during the next 200 years was made possible both by the enormous wealth of the royal exchequer and by the large numbers of slaves, skilled and unskilled, whose working lives were dedicated to the support of each institution. The city became one of the most important centres of Buddhist learning in the East.

Architecture

Lesser buildings are grouped around the more important pagodas and temples. Scattered around these are smaller pagodas and buildings, some of which may once have been aristocratic palaces and pavilions later adapted to monastic uses; *e.g.*, as libraries and preaching halls. All are based on Indian prototypes, modified during subsequent development by the Mon. The principal architectural themes are the Buddhist stupa, a tall bell dome, designed originally to contain near its apex the sacred relics of Buddhist saints. The other is the high, terraced plinth, which may be supplemented by stairs, gateways, extra stupas, and pinnacles and symbolizes a sacred mountain. Interior arches and vaults, both rounded and pointed, are, however, constructed by a true radiating-arch technique that was not used in India. During the course of artistic evolution the themes were frequently combined, and the combination developed into a complex rectangular hall with porticos extended from the sides, crowned by a stupa or, in some cases, by a rectangular tower of curved outline reminiscent of the contemporary Indian Hindu shrine tower. A vista across the site of Pagan shows a series of variations and combinations of the themes. Many buildings, especially those no longer in use and hence unrestored, bear substantial remains of external, decorative stucco and terra-cotta, adding flamboyance to the finely proportioned rectilinear structures,

Van Bucher—Photo Researchers

The Ananda temple, Pagan.

and internal paintings and terra-cottas recording Buddhist legend and history.

Anawrahta constructed the Shwezigon pagoda. Nearby he built a *nat* shrine with images. The Shwezigon is a huge, terraced pyramid, square below, circular above, crowned by a bell-shaped stupa of traditional Mon shape and adorned with stairways, gates, and decorative spires. It is still much revered and famous for its huge golden umbrella finial encrusted with jewels. Also revered are the late-12th-century pyramidal Mahābodhi, built as a copy of the temple at the site of the Buddha's enlightenment at Bodh Gayā, in India, and the Ananda temple mountain just beyond the east gate, founded in 1091 under King Kyanzittha. By the time the Thatpyinnyu temple was built (1144), Mon influence was waning, and a Burmese architecture had evolved. Its four stories, resembling a two-staged pyramid, and its orientation are new. Its interior rooms are spacious halls, rather than sparsely lit openings within a mountain mass, as in the earlier style. This building combined the functions of stupa, temple, and monastery. The Burmese style was further developed in the great Sulamani temple and culminated in the Gawdawpalin, dedicated to the ancestral spirits of the dynasty (late 12th century), whose exterior is decorated with miniature pagodas, the interior with extremely lavish, coloured surface ornament.

BIBLIOGRAPHY. The physical and geographical characteristics, with plans of the city, are discussed by DAW THIN KYI, "The Old City of Pagan," in *Essays Offered to G.H. Luce . . .*, vol. 2, pp. 179–188 (1966). WIM SWAAN, "Pagan," in *The Lost Cities of Asia*, pp. 90–120 (1966), gives an account of the architectural and artistic history. These are set into broader historical perspective and related to the arts of the rest of Southeast Asia by A.B. GRISWOLD in the "Burma" section of *Art of Burma, Korea, Tibet* (1964).

(P.S.R.)

Pageant and Parade

Pageants and parades are related phenomena, often one and the same thing, although a pageant may include events other than processions or omit them altogether. Basically both are entertainments, often in the open air, involving spectacle that is sometimes extravagantly devised; they are usually used as a means of expressing national, communal, or other kinds of group purpose or identity. The word pageant is derived from the Latin *pagina*, "page," and the form is, especially in its premodern manifestations, essentially a kind of illustration.

Three early usages of the word in English are recorded to denote (1) the individual episodes of a mystery or miracle play cycle, (2) the wagons upon which these plays were performed in the streets of certain English towns, and (3) any piece of stage scenery or machinery used to produce spectacular effects in the indoor court masques. Later, the word came to have broader implications, and parades or processions, even those solemn in nature, partake of the essence of pageantry.

ORIGIN AND DEVELOPMENT

Communal celebrations. In a primitive society a procession is one of the most basic demonstrations of communal unity. The occasions for such demonstrations vary greatly—fertility rites, the casting out of evil spirits, or a display of military strength are but a few—as do the formations the paraders may adopt. The simplest of these is for a whole tribe or community to form the procession. A more sophisticated variant is to include only selected individuals such as priests, witch doctors, warriors, or rulers who in their persons represent vicariously the interests of the community at large or their own particular status and function. It is characteristic of pageantry, as it is of heraldry, to represent in symbolic form the various classes and castes of society. Thus the common people will develop forms of pageantry that proclaim their own interests, whereas the religious, merchant, and ruling classes will produce other, but equally characteristic, forms. Occasions of national rejoicing, such as religious feasts or patriotic celebrations, tend to bring these various manifestations together into a more or less prolonged period of rejoicing.

Processions

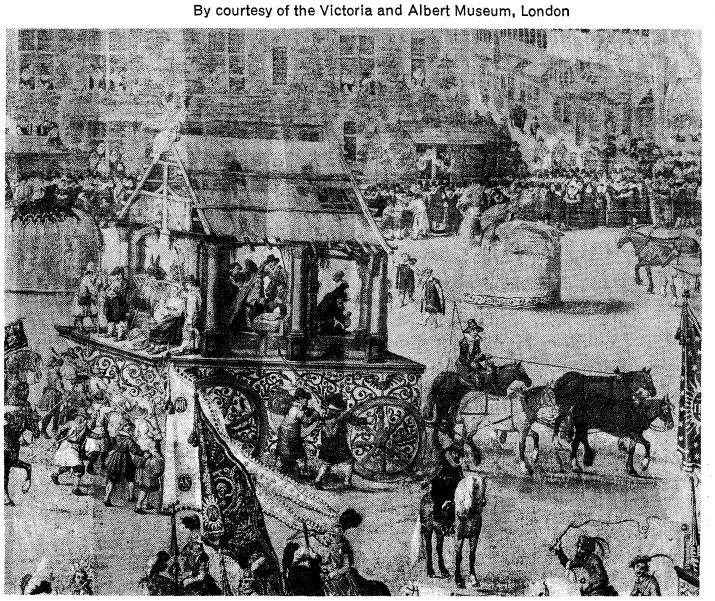
Once such periods of festivity have been established, they and the customs and processions connected with them tend to pass easily from one culture to another with little regard to the purposes for which they were instituted. Thus, for example, the carnival processions that precede Lent in many Roman Catholic countries are probably derived from the heathen feasts of Saturnalia, Lupercalia, and Bacchanalia. In the same way, Communist countries adopted May 1 as the occasion for great military parades, though May Day had been the occasion for fertility rites and processions since the dawn of history. The history of pageants and parades, therefore, is essentially one of eclecticism and adaptation. Any large modern parade, indeed, comprises items and floats that appear at first sight to have no necessary connection with one another: the origin and significance of some will lie in primitive history, that of others in more recent tradition, and that of still others in the demands of commercial advertising. Giants, dragons, Arabs, Indians, professional soldiers, drum majorettes, and wagons advertising milk and beer may follow one another in mystifying assortment. Moreover, the features of one procession may be taken over by another and their original significance either reinterpreted or lost altogether.

Pageant drama. The eclecticism of pageantry in general becomes even more pronounced in pageant drama. Music of one kind or another was an easily arranged adjunct to a simple procession. When it did not serve to express such spontaneous communal emotion as grief or bellicosity, it was probably intended to frighten off evil spirits. Once the theme of the procession came to be illustrated with spoken words or simple dramatic action, however, the parade had, of necessity, to be halted. "Stations" would be adopted and ceremonies or short scenes enacted at given points along the route. But such scenes and ceremonies would not necessarily be any more self-explanatory than the individual emblems and items that made up the parade. They would tend to remain enacted illustrations of a wider theme or story, the significance of which was usually apparent only to those in the group or community, who would already be fully conversant with the matter in hand.

Herein lies the difficult distinction to be made between drama proper and what in Western civilization has come to be the subdivision known as pageant drama. All forms of drama, of course, may be regarded as originating in religious rites—more particularly, and usually, in the desire to illustrate a particular theme or point as a kind of gloss to the rite itself. Thus, in the Christian Church, there arose the so-called tropes or additions to the mass; some of these were originally simple dialogues, based on the Gospel text, that were sung antiphonally

Relation to other drama

By courtesy of the Victoria and Albert Museum, London



Pageant car of the Nativity in the Brussels "Triumph" for the Archduchess Isabella, 1615. Detail from a painted panel by Denis van Alsloot in the Victoria and Albert Museum, London.

by the choir. Illustrative properties, such as the wooden ass, or *Palmesel,* used on Palm Sunday, were introduced, and a number of pageant-like ceremonies were evolved that eventually grew into complete mystery and miracle cycles. As such, they developed such fully rounded characters as Noah's wife in some English plays, a fact indicative of their tendency toward a more self-contained form of drama than that of the pageant episode. Certainly Western drama, once it had passed beyond the influence of the church and, more especially, as it was influenced by later Neoclassical scholarship in the 16th and 17th centuries, rapidly adopted the Aristotelian characteristic of having "a beginning, a middle, and an end." Strict pageant drama, on the other hand, differs from other dramatic forms in lacking this structure, since it is made up of loosely connected episodes.

Analogies in the East

This distinction between pageant drama and other forms is a peculiarly Western one. In Japanese Nō theatre, for example, and in its descendant jōruri and kabuki forms, it is customary for a performance to comprise individual scenes from several plays rather than a single, unified play. Although each incident, like the episode of a sophisticated Western pageant drama, may be more or less complete in itself, the full significance of the action is clear only to those well versed in Eastern myth or legend. The jōruri theatre, performances of which are still occasionally given, takes the form of the singing of dramatic ballads, the action of which is illustrated by puppets whose manipulators are always visible to the audience. Indeed, the frankly stylized nature of the puppet stage, with its conventionalized movements and often fragmentary narrative, communicated itself everywhere to other forms of Eastern theatre, not only in Japan but also in China and India. Thus the classical dramas of these countries have much in common with what Western theatregoers would class as pageants.

The pageant dramas of the West have tended to be largely open-air performances given in front of mass audiences, and hence they have stressed spectacle and broad effects at the expense of subtlety of characterization. The idea of a pageant as a collection of spectacular historical episodes illustrating a given theme or tradition is a modern one; it probably dates from the work of Louis Napoleon Parker in England, who produced a number of such entertainments in the first two decades of the 20th century.

TYPES AND OCCASIONS

Fertility rites. Fertility rites and military demonstrations must be counted among the earliest occasions for the holding of a parade. An instance of fertility procession has been recorded in the practice of a Snake tribe in the Punjab: an effigy of the snake god made of dough was carried from dwelling to dwelling and finally laid in the earth and worshipped. Similar processions have been recorded in all parts of the world, though sometimes a live animal was substituted for the effigy. Even as late as the 19th century, travelers to eastern Siberia witnessed such a procession, in this case with a bear who was subsequently sacrificed, among the Gilyaks, a Tunguzian people. In Paris a fat ox was paraded through the streets on Mardi Gras (Shrove Tuesday) in imitation of a Roman sacrificial procession. The ancient Chinese made wood and paper effigies of their dragon gods and carried them through the streets to procure rain, and until modern times the peasants of Sicily carried the effigy of St. Francis of Paolo through the market gardens with a similar intention.

Contests. A more sophisticated development of processions designed to influence the will of heaven can be found in a ceremonial race or contest in which the outcome is uncertain, presumably awaiting the decision of a supernatural power. It has even been suggested that the game of English football, or soccer, is derived from such a contest: after the sacrifice of the divine animal, teams or tribes contended with each other to obtain the sacred head, which would then be buried in their own territory to ensure its fertility for another year. If this is so, the processions of victorious football teams, returning to their home grounds with their trophy high over the heads of their exultant supporters, have an ancient history behind them. Certainly many modern forms of pageantry have the form of a contest, notable among them being the dragon-boat races of Hong Kong and the *palio* of Siena, Italy. The former takes place on the fifth day of the fifth month, when boats 80 to 100 feet in length, decorated to resemble dragons and crewed by up to 50 highly trained oarsmen, compete in an atmosphere of general festivity. The origin of the practice is said to be an attempt to stimulate the dragon gods to fight in heaven and hence produce rain.

Sporting events

The *palio* is an annual horse race held in the centre of Siena on July 2, the Feast of the Madonna of the Province, and on August 16, the day following the Feast of Our Lady's Assumption. It is remarkable for its colourful medieval costumes. Each *contrada,* or ward, of the city enters a horse, and the prize given to the winner is the *palio,* or banner, that invariably depicts a religious subject. The victors carry their prize home and keep it safely until the next contest. It is then taken in solemn procession to church on the day preceding the race and placed near the high altar during mass as a sign of its sacred significance.

Certain aspects of pageantry are also to be found in the *tauromaquia,* or bullfight, which has become a national sport of Spain and other Spanish-speaking countries. Both the parade of the brilliantly clad toreadors, matadors, and civic officials that precedes the fight itself, and the ceremonies that lead up to the releasing of the bull into the ring, are occasions of great splendour. Unlike the dragon-boat races and the *palio,* however, the bullfight is not chiefly popular today because it re-enacts a traditional ceremony. Its pageantry is essentially subservient to its importance as a contest, and although any popular sporting event may attract to itself elements of pageantry, as in the parades preceding football bowl games in the U.S., it is important to distinguish between contests that take place for their own sake and those that are staged for a more specifically mimetic or illustrative purpose.

A clear and extreme example of the latter type of contest, of a contest turned into a pageant, is the game of dice, played by Richard II of England when a procession of 130 citizens of London greeted him at Kennington at Christmas 1377. The procession itself was a mumming or disguising, words that in 14th-century vocabulary did not denote anything specifically dramatic but merely implied the wearing of colourful costumes for the Christmas revelry. When the mummers—dressed as knights and squires, "24 arrayed like cardinals," one dressed like an emperor, another like a pope—reached the King, they invited him to play at dice with them, a common feature of a mumming procession since those so disguised were popularly held to bring good luck with them. On this occasion, however, to ensure this good fortune the dice had been loaded so that Richard and other members of his party might be sure of winning the rich gifts the mummers had brought with them. The line dividing contest from pageant cannot always be so sharply drawn and depends ultimately on whether the contest or its attendant decoration is the main attraction.

Mumming

Particularly illustrative of the connection between contests and pageants were the jousts and tournaments of medieval Europe. Originally the tournament was nothing more than a means of training warriors, but as it evolved into a bloody and brutal exercise it encouraged the maintenance of private armies by the nobility, which was likely to lead to sedition. The consequent disapproval of both the church and heads of state led gradually to the adoption of literary and decorative elements, thus prolonging the existence of tournaments under the guise of civilized and harmless ornament. No longer might one end in the death of a contestant: artificial barriers, elaborately decorated, were set up in the streets, and the presence of the ladies as spectators led to the development of allegory of the kind encouraged by the cults of courtly love and chivalric honour. Knights would arrive in pageant cars decked out to pay fanciful compliments to the ladies of their choice, and such emphasis came to be put upon sheer equestrian skill that by the end of the 16th

Tournaments

century jousts and tournaments had disappeared altogether, their place being taken by the carrousel or horse ballet. This entertainment has at least three descendant forms in the modern world: the horse ballet of the circus; the carrousel or merry-go-round of the fair ground, whose wooden horses represent the live originals; and the displays of equestrian art given in the Spanish Riding School of Vienna.

Gymkhanas and tattoos. Equestrian skill is also much in evidence in gymkhanas and tattoos. The former originated under the British rule in India, combining displays of horsemanship with general athletics. The latter are a form of military display highly developed in Great Britain, at Aldershot and as part of the Edinburgh Festival, during the 20th century. The word tattoo, in this sense, is said to be derived from the Dutch phrase *doe den tap toe* ("turn the tap off"), which was used by the innkeepers of Flanders when the British were fighting there in the 17th century: as a sign of the closing of the taverns, it sent processions of soldiery through the streets on the way back to quarters.

State occasions. As a modern phenomenon, however, the spectacular tattoo must be considered in conjunction with other manifestations of pageantry on occasions of state celebration, which have always generated parades and lavish display. One of the earliest examples of this was the Roman triumph, with its procession of prisoners and victorious soldiers. Modern examples include the victory parades held in the Allied capitals after World Wars I and II and the many parades to celebrate national independence, such as those held in the United States on July 4, since 1950 in India on January 26, and in other nations. Such parades, like those of the various "old comrades" associations throughout the world (*e.g.*, the American Legion), serve to commemorate occasions of general national importance. Others, however, may celebrate only the interests of one section or ethnic group within the larger community—*e.g.*, the St. Patrick's Day parades or the ceremonial "piping in" of the haggis on "Burns's night," a tradition maintained by Scottish communities throughout the world.

Victory parades

Coronations, royal weddings, and state visits have been the occasion for pageantry since the dawn of history and were enhanced in medieval Europe by the development of spectacular pageant machines and devices. In 1481, for example, Isabella of Spain was greeted at Barcelona by a pageant wagon of St. Eulalia, remarkable for its device of three revolving spheres. In 1514, when Louis XII of France was married to Mary Tudor, sister of Henry VIII of England, a pageant was erected outside the Church of the Holy Innocents in Paris to represent the occasion in allegorical form. The pageant had two levels: on the lower stage a rose representing England, on the upper a lily for France. The rose had a bud that grew until it reached the second level, where it opened to reveal a girl representing Mary herself. At the same time the lily opened to show a youth who symbolized the French king.

By courtesy of the trustees of the British Museum

Double pageant stage with operating medieval stage machinery. The rosebud grew, opened its petals and ascended to the height of the upper stage. Illumination from a medieval manuscript in the British Museum (Cot. Vesp. B. II).

Such pageantry was frequent in the major cities of Europe during the 16th century, but thereafter it declined in frequency. The machinery involved could stand little comparison with the effects that were obtained in the court masques or even, perhaps, the public theatres, and hence static pageantry in the streets tended to give way to moving processions. Samuel Pepys, for example, found the pageants offered for the Lord Mayor's procession of 1660 "good for such kind of things, but in themselves but poor and absurd." The Lord Mayor of London still has his "show" at the commencement of his year of office, but the use of mechanical devices at given stations has vanished. A procession connected with merchant guilds, that of Sächselauten (literally, "ringing of the bell at six o'clock"), persists in Zürich, Switzerland, but does not reach a station until its end in the heart of the city, where the figure of winter is ceremonially burned.

Court masques. Court masques, an example of pageantry confined to one section of the populace, became increasingly elaborate during the 17th century. Essentially they were entertainments that led up to a dance or masked ball: the masquers would conclude their display by inviting the spectators to dance with them, and it was therefore essential that they themselves be of high birth. The names of Ben Jonson and Inigo Jones are closely connected with this form of display in England, which illustrated a simple theme with lavish embellishments of music, poetry, costumes, and ever-changing scenery. In Italy, where stage machinery was first devised in the West, the masque reached even more spectacular heights. At Parma, on the occasion of the marriage of Ranuccio II to Margherita of Tuscany in 1628, an entertainment was given in the ducal theatre that combined opera, tournament, ballet, and regatta. Its climax was the flooding of the floor of the Teatro Farnese and the entry of several pageant cars representing sea monsters and islands, each propelled by men who guided them from beneath. After the display the waters were drained away as miraculously as they had appeared.

In Austria, too, such pageants were popular: when the emperor Leopold I welcomed the Spanish infanta Margaretha Elisabeth as his bride in 1666, the "Contest between Air and Water" was staged at the Hofburg. This display, which included a carrousel that had taken months to rehearse, reached its climax when the Temple of Eternity descended from the skies to reveal within itself a cupbearer, 8 live guardsmen, 16 grooms, 12 trumpeters, and the monarch himself, dressed like a Roman emperor and on horseback. The entire spectacle comprised numerous pageant cars, much music, and over 1,300 performers.

"Contest between Air and Water"

Carnival. The use of such lavish pageantry by a single privileged class of society has necessarily disappeared in the modern world. Where much money is devoted to display, all sections of the community demand equal rights to enjoyment of the splendour. Nowhere can the democratic tendencies of modern pageantry be seen more clearly than in carnival processions. In the West, carnival originated in Italy, sharing characteristics with ancient Roman feasts. The Roman Saturnalia permitted slaves to treat their masters as equals, a license that recurred in both the Feast of Fools and carnival itself. Each occasion has its unique procession. On the Feast of Fools, observed by the subdeacons in great cathedrals during January, it was the custom in 13th-century France for a girl carrying a baby to ride into church on an ass, in imitation of the flight into Egypt. In the ensuing mass,

celebrant and choir mimicked the braying of an ass, and the proceedings were sometimes concluded by the throwing of buckets of water. Carnival processions at first exhibited equal disorder but soon acquired elaborate artistic display. In modern Italy the processions at Rome and Venice are especially famous, the Venetian ones inevitably being held on the canals. An equally renowned carnival is held at Nice, whose culture is predominantly Italian. One of Nice's main attractions is the Battle of Flowers at carnival and other occasions. Floats are lavishly decorated, and spectators and paraders pelt one another with blossoms. The origin of the custom may lie in the ancient habit of exchanging gifts of foliage at fertility festivals, but the throwing of flowers, bonbons, confetti, rice, and ticker tape is common to many parades, and too much stress must not be laid on its having primitive origins. Like the ubiquitous balloon—in origin, possibly the bladder taken from a sacrificial animal—such features endure by the simple pleasure they afford, a reminder of the ease with which features of one kind of parade are incorporated into another without preexistent rites or traditions to support their new presence. The Battle of Flowers, for example, has been copied in the Tournament of Roses parade at Pasadena, California, since 1890, though it has acquired a more distinctly American flavour by association, since 1916, with the Rose Bowl football game. Students returning from Paris in 1827 introduced a Mardi Gras parade to New Orleans; decorated floats were added in 1837, and annual parades there date from 1857. In Rio de Janeiro, a similar pre-Lent carnival has long been customary. In Germany, too, the carnival, known as Fasching, with its attendant processions, was re-established in the 19th century in both the Rhineland and Bavaria, especially in Munich and Cologne.

Eastern carnivals

Carnival parades also occur in the East. Hindus observe the feast of Holī in March, originally a fertility celebration, when revellers set out in groups to squirt coloured water at everyone they meet or to smear their faces with coloured powders, customs recalling the throwing of water at the Feast of Fools and the blackening of the face in European parades. In China and in Chinese communities such as that in San Francisco, the New Year is observed with parades distinguished by elaborate dragon masks. In Hong Kong the festival of Tin Hau, Queen of Heaven, sees a procession of "lion dancers" accompanied by hooters and drums. Such dances originated in religious ritual and elsewhere have developed into dramatic form, as in Japanese Nō theatre. But Japan is also rich in pageantry of a less specifically dramatic kind. At the Gion festival in July enormous floats and magnificent costumes are exhibited, and at Miyazu Toro Nagahi in August thousands of lanterns are set floating on the sea in honour of the dead. The anthropologist Sir James Frazer suggested that carnival giants like those once popular at Douai and Dunkirk, France, and Antwerp, Belgium, originated in the sacrifice to the nature god of human beings, carried along in huge osier frames decorated with greenery, much in the fashion of the basket-clad figure of Jack-in-the-Green in English May Day parades. But not all parades originated in fertility rites. Some arose from practical necessity, like the British "trooping of the colour," which came from the need to familiarize mercenaries with the flag they were to follow. And the trooping ceremony, held in London on the sovereign's official birthday, uses mounted soldiers to give the tall effects essential to any parade.

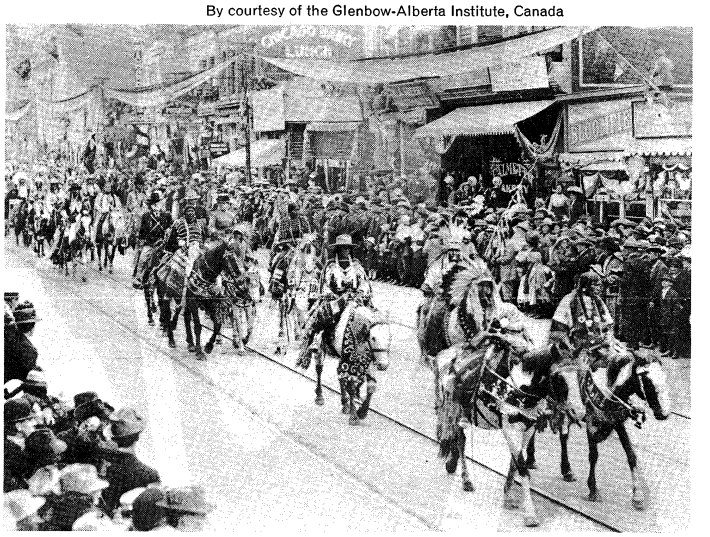
By courtesy of the Japan National Tourist Organization

Decorated shrine car in the Gion Matsuri festival in Kyoto, Japan, 1968. The festival originated in the 9th century as a religious procession.

Such height is interestingly met at the annual Bun festival in Hong Kong by bandaging children to supporting frames and carrying them above the procession. Most contemporary giants are, however, very tall and often mechanized figures, long serpents and dragons, perhaps belching smoke, or are enormous balloons in the form of a traditional or comic-strip character.

Historical re-enactments. Many modern forms of pageantry unite parades with some form of drama. The Calgary Stampede in Alberta includes not only a procession, competitions, and a carnival atmosphere but also re-enactments of battle between cowboys and Indians. A similar historical re-enactment was presented for Queen Elizabeth II of England at Waitiang, New Zealand, in 1963, when she was greeted by a "Maori challenge." In India the parade at the autumn festival of Dussera culminates in plays based on incidents in the life of Rāma, the prince of Hindu mythology. In Delhi the performances are especially spectacular, and one incident of the prince's life is dramatized each day. As a culmination of these celebrations in northern India, the effigy of the demon-king Rāvaṇa is packed with fireworks and burned.

By courtesy of the Glenbow-Alberta Institute, Canada



The Indian section of the first Calgary Stampede, Alberta, 1912.

Christian themes

In the Christian religion simple plays and tableaux, such as the procession of the Magi at Limoges in the 11th century and the introduction of the Christmas crib three centuries earlier, were incorporated into church services. Like the Hindu Rāma plays and the Islāmic plays of Ḥusayn (a grandson of Muḥammad) still performed in Shīʿah towns, such Christian manifestations show the essential pageant characteristic of being fully understood only in relation to a wider context that is not itself dramatized. Even when Christian drama left the church proper and became, in such English cities as York and Coventry, part of a miracle cycle ranging from the Creation to the Day of Judgment, this characteristic remained: although one episode might have dramatic connection with the next, the different episodes were performed in different parts of the town.

Pageant in modern drama. In recent times distinguished dramatists have contributed to pageant drama, notably T.S. Eliot, who wrote the choruses for *The Rock*, a religious pageant staged in England in 1934. Other dramatists have written dialogue and commentary for *son et lumière*, the most modern development of pageant form. In the U.S., outdoor dramas on historical themes are staged annually. Notable among these are Paul Green's *The Lost Colony* on Roanoke Island; *The Common Glory* at Williamsburg, Virginia; and *Unto These Hills*, dealing with the Cherokee Indian tribe, at Cherokee, North Carolina. Such entertainments are sometimes termed epic dramas, but the designation is misleading, implying as it does a connection with the work and theories of the German playwright Bertolt Brecht. Although Brecht's own plays were deliberately episodic and eschewed Aristotelian precepts of unity, their basic intention was opposite to that of most pageantry. Brecht used a fragmentary method to provoke his audience into critical thought, whereas the usual aim of a pageant is to awaken nostalgic and uncritical identification with past glories.

Reinhardt's contributions

Not Brecht but Max Reinhardt must be seen as the essential figure behind modern pageant drama. From the beginning of the present century Reinhardt was advocating *Sprengung des Bühnenrahmens* ("exploding the confines of the stage") and experimenting with performances on such unusual stages as the circus ring and the exhibition hall. These experiments led others to realize that drama could indeed be staged, as it had been before the advent of the proscenium arch in the 17th century, in many different locales. Reinhardt's production of *Jedermann*, the adaptation by Hugo von Hofmannsthal of the medieval morality play *Everyman*, has been performed in front of Salzburg Cathedral annually since 1920; it is distinguished by elaborate crowd movement and audacious dramatic effects. His staging of *The Miracle* in London, in 1911, with music by Engelbert Humperdinck, was one of the most elaborate and successful examples of modern pageantry.

Restaged classics. Not all modern pageant drama, of course, is especially written or composed for the occasion. Works that are so written tend by their very nature to be of only local or temporal interest, although Mozart's music for student parades in Salzburg and Handel's *Water Music* and *Music for the Royal Fireworks* show that individual contributions to pageantry can enter the repertory of the classics.

An interesting modern development of pageantry is the restaging of classical works in open-air theatres of countries closely connected with the genre or theme presented. The operas of Verdi and other composers are given annually during the summer in the Arena at Verona and in the other Roman monuments throughout Italy. In Greece, classical plays are staged in the ancient theatres with a maximum of spectacular effect. At Bregenz in Austria, operettas are performed on a stage that stands in the waters of Lake Constance, transforming them, as it were, into water pageants. In Switzerland annual open-air performances are given of Schiller's *Wilhelm Tell*. Although all these works lend themselves to such spectacular production, it is significant that none was originally conceived as a pageant in the strict sense of the word. The nature of the presentation alone makes them so, coupled with the fact that they seem uniquely representative of the national spirit of the country concerned.

Oberammergau

A similar manifestation of local feeling is represented by the famous passion play (*Passionsspiel*) that is presented every ten years at Oberammergau in Bavaria. This custom dates from 1634, originating in a pledge made by the villagers in gratitude for the ending of the plague, in 1633. The actors are all local amateurs, about 500 in number; the action of the play lasts more than seven hours; and the entire proceeding is informed by a strong sense of religious devotion. In the intervening years other plays are presented, usually on Old Testament subjects, to give the actors training and experience. The *Passionsspiel* is given in a modern open-air theatre that is reserved for it alone. Since 1680 it has been presented in decimal years.

It will be seen, therefore, that pageantry is by no means a dead art nor a mere relic of things past. The instincts that created it in the past—religious belief, superstition and fear, militarism, great national occasions, the desire of a particular community to demonstrate its traditions and ethos—are still central forces in human existence and will doubtless continue to develop and adapt pageantry. Although many of the forms they take may be ephemeral, pageants and parades must be regarded as an abiding part of man's artistic endeavour.

BIBLIOGRAPHY. Major works on theatre history that discuss pageants and parades include E.K. CHAMBERS, *The Medieval Stage*, 2 vol. (1903), *The Elizabethan Stage*, 4 vol. (1923); G.W.G. WICKHAM, *Early English Stages, 1300 to 1660*, 2 vol. (1959–63); and A. NICOLL, *World Drama* (1949). E.J. STERN, *My Life, My Stage*, trans. from German by E. FITZGERALD (1951), covers the work of Max Reinhardt, theatrical producer; J.W. GASSNER, "Outdoor Pageant-Drama: Symphony of Sight and Sound," *Theatre Arts*, 38:80–83 (1954), considers modern U.S. pageant drama.

(Jo.P.)

Pain, Theories of

Pain is the experience one associates with injury or some type of damage to the body: blows, bangs, cuts, and so forth. It is described in this way because the experience of pain is psychological and subjective. A good deal is known about the anatomical and physiological basis for pain, however. A minimal effective condition for the experience of pain is any stimulus that can cause tissue damage, slight and easily reversible, perhaps, but nevertheless damage. There is evidence that persons learn to discriminate the experience of pain as a result of events in infancy. Withdrawal from damaging stimuli occurs readily and early both in man and other animals. There are indications, however, that undue protection from the experience of damage reduces the ability to feel pain or to discriminate the stimuli that give rise to it. Not only does the unburned child not fear the fire, he may less readily have pain or know what it is when he is ultimately burned.

There are upward of 65 reports of persons who were born without the ability to feel pain at all, or else very little. Although they may survive without disability, they are at greater risk of suffering, or neglecting, burns, fractures, or infections. A similar effect has been produced in one recorded case by accidental damage to part of the midbrain. The life expectancy in these cases is shorter than average.

Even in people without this disability severe injury quite often does not cause pain; on the other hand, pain in adults often occurs for psychological reasons, in circumstances such that a physical basis for this bodily experience is, apparently, either absent or secondary. Thus an important proportion of persons with pain have their pain not because of any physical disease but because of emotional disorders. Nevertheless, pain is primarily associated with injury. To allow for this, for the learning of pain, for its usual subjective patterns, and for the frequent lack of evidence of a physical basis, pain has been defined operationally as an unpleasant experience, primarily associated with injury to tissue, and described in terms of such damage.

Pain without physical disease

Pain helps to preserve the organism by securing withdrawal from harmful agents. Alternatively, in sick people, pain may secure the rest that is needed for healing. Nature often is overabundant, however, and pain at times appears unnecessarily severe. This happens especially in some cases of damage to the nervous system.

MANIFESTATIONS OF PAIN

Pain is presumed to occur when a person takes avoiding action after stimulation from a potentially damaging source or when he gives signs of distress suggesting the effects of pain. The external signs of pain thus include avoidance behaviour, verbal complaints, cries, yelps, and so on. Another reaction to damaging stimuli is not avoidance but attack on neighbouring organisms; this may happen among rats that are subjected to electric shock. None of these signs is exclusive to pain. For many pur-

The private nature of pain

poses pharmacologists and other experimenters can work with laboratory animals, testing analgesics (pain-relieving drugs) on the assumption that their cries represent painful experience. But there can never be any certainty of this since inevitably the supposed experiences of others are private; only their unreliable external activities are public. One is only able to feel directly his own headache. Thus in human research, pain can only be inferred or assessed from what the sufferer does or says about it. After allowing for some discrepancies, it can be shown that people may grade their pains approximately as mild, moderate, or severe and that these descriptions tend broadly to conform to the degree of disorder, physical or psychological, from which it may be ascertained that they are suffering. Further, the measures required to relieve pain (*e.g.*, by analgesics) tend to conform to what would be predicted from the descriptions obtained.

The introspective report of pain is most consistent with regard to severity. The quality of pain is extremely hard to describe. Usually, its site can be denoted, then its intensity (*e.g.*, mild or very bad), and, with somewhat more difficulty, whether it is dull, throbbing, piercing, stabbing, and so forth. When the sufferer offers many descriptions of his unpleasant subjective state, it seems to be an indication of the severity that he attributes to the pain.

An indirect or inferential measurement of someone else's pain may be made by means of so-called pain thresholds. All techniques available for measuring these are open to criticism. They require the individual to undergo a safe stimulus that is varied in intensity until it just begins to feel painful; critics observe this to be a very different matter from spontaneous pain (as in cancer) in which the potential harm may be unclear and the feared risk very alarming. A popular technique for measuring pain thresholds employs a device (dolorimeter) that consists of a lamp as a source of heat, focussed onto the skin, and operated by a timing mechanism. An even safer device is a pressure algometer, a rod with a flat end loaded against a calibrated spring. Pressure is applied with this rod against a flat bony surface (*e.g.*, the shin) and increased steadily until pain is reported. Thresholds simply specify the directly observable degree of stimulation (intensity of heat or pressure) reported as pain. Under controlled conditions it has been shown that women are more sensitive (have lower thresholds) than men, office workers more than manual workers, and anxious people more than calm people. Thresholds tend to increase (sensitivity diminishes) somewhat in old age, although spontaneous pain at that time of life is relatively frequent.

NEUROPHYSIOLOGICAL BASES

Before the 19th century, theorists, physicians, and others tended to think of pain as an emotion (*q.v.*) rather than as a sensory experience. Anatomical and neurophysiological discoveries overwhelmingly resulted in its later classification as a sensory activity. More recent work, however, does not sustain such a limited sensory classification, although the neurophysiological basis remains extremely important. Pain is generally reported after injurious stimuli (or others, such as electric shock, which are not always damaging) are applied to normal skin or other tissues that contain microscopic structures called free nerve endings (see MECHANORECEPTION; NERVES AND NERVOUS SYSTEMS). Free nerve endings seem to subserve other experiences besides pain and are widely distributed to internal and external body surfaces; from them nerve fibres conduct impulses to the spinal cord. A stimulated point on the skin that is supplied by free nerve endings can, under varying circumstances, give rise to reports of other experiences (*e.g.*, heat, cold) as well as pain. Thus it can be argued that pain can best be related to variable anatomic and physiological *patterns* rather than to specific, unchanging neural pathways, each resembling the tract of an isolated single railway line. Nevertheless, there is also evidence that some nerve fibres in the vicinity of their receptors (*i.e.*, near such peripheral parts of the body as the skin) do respond in a fairly specific way to noxious stimulation. Accordingly, even at the most peripheral level it seems that there are some fixed elements that contribute to the physiological basis for the experience of pain. Their effects, however, seem to be modified by other concomitant patterns of activity deeper within the nervous system. In general, the nerve fibres that respond in a fixed way to noxious stimulation belong to the smaller myelinated and to the unmyelinated groups of fibres. (Myelinated fibres are those covered with a fatty sheath that seems to have an insulating function.) These conduct impulses more slowly than the larger fibres, which have been related to such sensory experiences as touch and pressure. Some of the so-called pain fibres are distributed to what is named (from its appearance) the posterior horn of gray matter in the spinal cord. Here they appear to be in close relationship to the substantia gelatinosa, an area lying at the tip of the posterior horn. Additional fibres cross from the two posterior horns to the opposite side of the spinal cord where they run in an area of white matter known as the anterolateral column. Several tracts run in this column, including a spino-thalamic tract that goes to the thalamus, which serves as a kind of relay centre in the brain. There is a wealth of evidence that interruption of this tract either by injury or disease (sometimes very highly localized damage) results in loss of the ability to feel pain from stimuli applied to the opposite side of the body in parts below the level of interruption.

The role of free nerve endings

Such loss is found to be an invariable consequence of full interruption of the pathways mentioned; more spatially restricted loss occurs after injuries to individual nerves. In a small proportion of cases, damage to the pathways may result in the spontaneous occurrence of pain, referred to the area supplied by the pathway, as well as a loss of sensory input. When this pain follows from damage to peripheral nerves, it is usually reported as burning in character and is called causalgia. After a limb is lost, the amputee normally still experiences the impression that it remains; this effect of a phantom limb is also sometimes accompanied by phantom pain (often very unpleasant). Spontaneous discomfort similar to causalgia can occur when there is damage to more centrally located structures (such as the spinal cord or the thalamus) and is known as central pain. It is convenient to call central pain, phantom-limb pain, and causalgia, jointly, neurogenic pain.

The sensation of pain in a missing leg

Neurogenic pains usually occur together with some general loss of sensory activity in the affected part of the body, thresholds for touch being raised particularly, a circumstance known as anaesthesia dolorosa. Study of such cases has led to the development of a theory that pain results from an imbalance of input between large, rapidly conducting myelinated ("fast") fibres and smaller, slower conducting myelinated and unmyelinated ("slow") fibres. If the slow fibres are heavily stimulated or the fast fibres are not stimulated enough, pain is believed to develop. Physiological evidence suggests that the balance is controlled by certain cells in the spinal cord that operate a "gating" mechanism that modulates, or regulates, the fast and slow input. These cells in turn appear susceptible to the influence of higher level structures (*i.e.*, cells farther up in the nervous system) that may increase or reduce their activity.

A coherent theory is thus available to account for experienced variations in pain by invoking the influence both of local peripheral events and of higher levels in the nervous system. Stated in neurophysiological language, this theory accounts for many of the observed phenomena. While it allows for fixed peripheral input, it also places particular weight on the pattern of input and on the possibility of psychological influence associated with relevant brain processes.

PSYCHOLOGICAL THEORIES

Some theories of pain are given primarily in psychological terms without specifying possible neural bases for them. As noted above, pain frequently is attributed to psychological (*e.g.*, emotional) factors. Secondly, an emotional state (*e.g.*, anxiety) often can enhance or abate pain that arises from observable physical sources. In consequence it is often convenient to talk of pain in

introspective terms whenever its possible neurophysiological basis remains incompletely explained.

Limited evidence indicates that persistent muscle tension will give rise to pain, as in holding an unfamiliar posture on a long journey. Similarly, there may be contraction of muscles as a result of anxiety, particularly in the neck and scalp. Tension in these muscles produces pain, notably headache, which generates further anxiety and alarm; this makes the pain even worse, in what seems to be a vicious circle. Relaxation and the easing of anxiety frequently bring relief. Such pain may be called psychosomatic; *i.e.*, it seems to be a bodily effect stemming from a psychological disturbance.

Sometimes pain seems to arise from an idea. The psychoanalyst Sigmund Freud reported the case of a man who attended an operation in which his brother's hip joint was to be straightened under an anesthetic. The joint was frozen in place by some abnormal adhesions that had to be broken; when a loud crack was heard, the man said he felt a pain in his own hip at the same moment, in what might be interpreted as sympathetic identification. Another physician reported a patient who imagined a nail or hammer at the back of his head and who experienced pain whenever he thought of the nail. Pain can arise as a hallucination as well; this may happen among people who are severely depressed and occasionally among those who suffer from schizophrenia or epilepsy.

All of these examples represent psychological routes for the production of pain. A psychological theory of pain that allows both for the physical and psychological description of events has been offered by the Hungarian-U.S. psychoanalyst Thomas Stephen Szasz. According to Szasz, pain functions as a danger signal and arises when a threat appears to the integrity of the body. Pain is held to arise as a consequence of the individual's awareness of a specific threat of loss in terms of bodily function. While it is wholly psychological, this theory seems to account for the apparent discrepancies between the intensity of pain and the extent of bodily lesions observed in people who report pain. Pain is regarded as "organic" or as "neurotic" according to the observer's opinion of the reality of the threat to the body. The actual presence or absence of structural disturbance is considered not to be relevant to the psychological mechanism but only to the individual's assessment of the reality of the threat.

MODIFICATION OF PAIN

It has been indicated that, in addition to physical lesions in the body generally and in the nervous system especially, the factors related to pain include age, sex, occupation, social circumstances, learning effects, and anxiety. There is much evidence that pain can be alleviated by drugs and psychiatric techniques designed to reduce anxiety. It is probable that analgesics such as acetylsalicylic acid (aspirin) have a special pharmacological effect in the relief of pain. But any drug or other measure that increases confidence and abates fear also will ease pain.

Effects of trance states, hypnosis, and drugs

The abolition of pain by trance states, including hypnosis, has long been accepted. Without taking all claims at their face value (which is probably inflated), it can be accepted that the modification of attitudes does at times abate pain. While this holds for moderate and slowly rising intensities of pain, it seems much less effective for sudden or severe pain. Nevertheless, under special circumstances (as in battle), severe wounds may not be noticed until later. As the French writer Michel de Montaigne put it, one feels a single cut from the surgeon's scalpel more than ten strokes of the sword in the heat of battle.

When drugs work by altering mood, their effect is partly attributable to the effects of suggestion and the expectation that relief will occur. This commonplace placebo effect (*e.g.*, relief by swallowing sugar pills) seems actually to result in less pain to the extent that the sufferer believes he will improve, becomes more confident, has more hope, and less anxiety. With such drugs as aspirin, part of the benefit is attributable to direct pharmacological effects on anxiety and hence on pain.

In one study, wounded soldiers were compared with civilians who had surgical wounds of comparable size. To the soldiers, their incapacitating wounds apparently meant an honourable release from danger. At any rate, their demands for analgesics were much more modest than those of the civilians for whom surgery had brought no comparable advantages, only serving to interrupt their lives. This shows how circumstances can alter pain. Pain seems alleviated to a very similar extent by hypnosis, placebos, suggestion, psychotherapy, and other means that relieve anxiety, including sedatives and the milder analgesics. For effective relief, more severe pain with a mainly organic basis tends to require potent drugs, local anesthesia, or neurosurgery. Treatments for depression also relieve chronic pain in many cases. Among current treatments available, some surgical operations on selected portions of the brain may be effective in pain relief; their application remains a matter for very careful consideration, however, and is certainly not routine.

MASOCHISM

Feelings of pain and pleasure are not simple opposites; but a surprising mixture of the two has often been suggested as occurring in masochism. By one definition, masochists are people who find sexual gratification in being humiliated or physically injured. Among many who seek injury or humiliation, however, the goal certainly does not seem to be immediately sexual. This is not to say that physical trauma in sexual activity is uncommon; indeed, among some animals (*e.g.*, the mink) it seems essential for normal arousal and orgasm. Among humans, it is most abnormal to require such injury for full sexual arousal, although acceptance of mild pain in sexual encounters may be normal. In addition, the term masochism may be extended to self-damaging behaviour, as among people who pursue martyrdom. Some psychiatric patients appear to experience pain as a result of such broadly defined masochistic motives. Whatever such behaviour represents, to some students of behaviour it seems a contradiction in terms; *i.e.*, the enjoyment of the unenjoyable.

Most often it is the humiliation or subservient sexual role that is sought by the extreme masochist, rather than any actual physical damage. It is thus possible to consider many manifestations of masochism as relatively normal (or at least harmless) in relation to the biological economy. Insofar as females do experience pain in such activities as defloration and childbirth, it would seem advantageous to the survival of species that they should have evolved to some experience of pain as rewarding.

BIBLIOGRAPHY. H. MERSKEY and F.G. SPEAR, *Pain: Psychological and Psychiatric Aspects* (1967), a survey of the history, medical characteristics, and clinical and experimental aspects of pain; W. NOORDENBOS, *Pain: Problems Pertaining to the Transmission of Nerve Impulses Which Give Rise to Pain* (1959), a penetrating and original consideration of pain mechanisms, based upon clinical data of neurological investigations; R. MELZACK and P.D. WALL, "Pain Mechanisms: A New Theory," *Science*, 150:971–979 (1965), a paper presenting a new theory, partly derived from that of Noordenbos but taking into account additional physiological observations and the relationship of pain to psychological factors; J.D. HARDY, H.G. WOLFF, and H. GOODELL, *Pain Sensations and Reactions* (1952), the classic report of investigations conducted during the period 1940–50; K.D. KEELE, *Anatomies of Pain* (1957), a scholarly and valuable source book dealing with the history of pain and its mechanisms; W.H. SWEET, "Pain," in J. FIELD (ed.), *Handbook of Physiology*, vol. 1 (1959), a standard account of the physiology of the subject; R.S. KNIGHTON and P.R. DUMKE (eds.), *Pain* (1966), a symposium —good in parts; T.S. SZASZ, *Pain and Pleasure: A Study of Bodily Feelings* (1957), an original, thoughtful book by a psychoanalyst, very clear on the conceptual and semantic problems of pain.

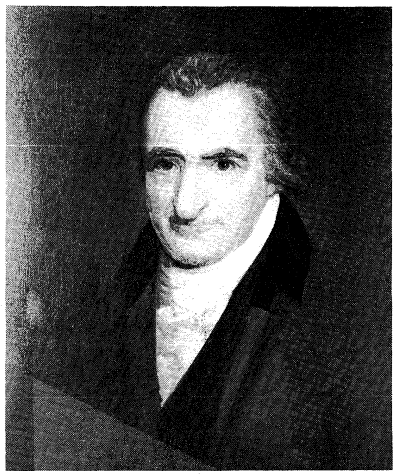
(H.Me.)

Paine, Thomas

Thomas Paine, political journalist, patriot, and champion of the rights of the common man, first achieved fame with the pamphlet "Common Sense," a stirring plea for American independence. Other works that gained him his

reputation as one of the greatest political pamphleteers in history were "The American Crises" papers, which rejuvenated the dispirited Continental Army; *Rights of Man*, a defense of the French Revolution and of republican principles; and *The Age of Reason*, an exposition of the place of religion in society.

By courtesy of the Thomas Paine National Historical Association



Paine, portrait by John Wesley Jarvis. In the Thomas Paine Memorial House, New Rochelle, New York.

Life in England and America. Paine was born at Thetford, Norfolk, on January 29, 1737, of a Quaker father and an Anglican mother. His formal education was meagre, just enough to enable him to master reading, writing, and arithmetic. At 13 he began work with his father as a corset maker and then tried various other occupations unsuccessfully, finally becoming an officer of the excise. His duties were to hunt for smugglers and collect the excise taxes on liquor and tobacco. The pay was insufficient to cover living costs, but he used part of his earnings to purchase books and scientific apparatus.

Paine's life in England was marked by repeated failures. He had two brief marriages. He was unsuccessful or unhappy in every job he tried. He was dismissed from the excise office after he published a strong argument in 1772 for a raise in pay as the only way to end corruption in the service. Just when his situation appeared hopeless, he met Benjamin Franklin in London, who advised him to seek his fortune in America and gave him letters of introduction.

Arrival in Philadelphia

Paine arrived in Philadelphia on November 30, 1774. His first regular employment was helping to edit the *Pennsylvania Magazine*. In addition Paine published numerous articles and some poetry, anonymously or under pseudonyms. One such article was "African Slavery in America," a scathing denunciation of the African slave trade, which he signed "Justice and Humanity."

Paine had arrived in America when the conflict between the colonists and England was reaching its height. After blood was spilled at the Battle of Lexington and Concord, April 19, 1775, Paine argued that the cause of America should not be just a revolt against taxation but a demand for independence. He put this idea into "Common Sense," which came off the press on January 10, 1776. The 50-page pamphlet sold more than 500,000 copies within a few months. More than any other single publication, "Common Sense" paved the way for the Declaration of Independence, unanimously ratified July 4, 1776.

During the war that followed, Paine served as volunteer aide-de-camp to General Nathanael Greene. His great contribution to the patriot cause was the 16 "Crisis" papers issued between 1776 and 1783, each one signed "Common Sense." "The American Crisis. Number I," published on December 19, 1776, when George Washington's army was on the verge of disintegration, opened with the flaming words: "These are the times that try men's souls." Washington ordered the pamphlet read to all the troops at Valley Forge.

In 1777 Congress appointed Paine secretary to the Committee for Foreign Affairs. He held the post until early in 1779, when he became involved in a controversy with Silas Deane, a member of the Continental Congress, whom he accused of seeking to profit personally from French aid to the United States. But in revealing Deane's machinations, Paine was forced to quote from secret documents to which he had access as secretary of the Committee for Foreign Affairs. As a result, despite the truth of his accusations, he was forced to resign his post.

Paine's desperate need of employment was relieved when he was appointed clerk of the General Assembly of Pennsylvania on November 2, 1779. In this capacity he had frequent opportunity to observe that American troops were at the end of their patience because of lack of pay and scarcity of supplies. Paine took $500 from his salary and started a subscription for the relief of the soldiers. In 1781, pursuing the same goal, he accompanied John Laurens to France. The money, clothing, and ammunition they brought back with them were important to the final success of the Revolution. Paine also appealed to the separate states to cooperate for the well-being of the entire nation. In "Public Good" (1780) he included a call for a national convention to remedy the ineffectual Articles of Confederation and establish a strong central government under "a continental constitution."

Efforts to aid the American troops

At the end of the American Revolution, Paine again found himself poverty stricken. His patriotic writings had sold by the hundreds of thousands, but he had refused to accept any profits in order that cheap editions might be widely circulated. In a petition to Congress endorsed by Washington, he pleaded for financial assistance. It was buried by Paine's opponents in Congress, but Pennsylvania gave him £500 and New York a farm in New Rochelle. Here Paine devoted his time to inventions, concentrating on an iron bridge without piers and a smokeless candle.

In Europe: "Rights of Man." In April 1787 Paine left for Europe to promote his plan to build a single-arch bridge across the wide Schuylkill River near Philadelphia. But in England he was soon diverted from his engineering project. In December 1789 he published anonymously a warning against the attempt of Prime Minister William Pitt to involve England in a war with France over Holland, reminding the British people that war had "but one thing certain and that is increase of taxes." But it was the French Revolution that now filled Paine's thoughts. He was enraged by Edmund Burke's attack on the uprising of the French people in his *Reflections on the Revolution in France*; and, though Paine admired Burke's stand in favour of the American Revolution, he rushed into print with his celebrated answer, *Rights of Man* (March 13, 1791). The book immediately created a sensation. At least eight editions were published in 1791, and the work was quickly reprinted in America, where it was widely distributed by the Jeffersonian societies. When Burke replied, Paine came back with *Rights of Man, Part II*, published on February 17, 1792.

What began as a defense of the French Revolution evolved into an analysis of the basic reasons for discontent in European society and a remedy for the evils of arbitrary government, poverty, illiteracy, unemployment, and war. Paine spoke out effectively in favour of republicanism as against monarchy, and went on to outline a plan for popular education, relief of the poor, pensions for aged people, and public works for the unemployed, all to be financed by the levying of a progressive income tax. To the ruling class Paine's proposals spelled "bloody revolution," and the government ordered the book banned and the publisher jailed. Paine himself was indicted for treason, and an order went out for his arrest. But he was en route to France, having been elected to a seat in the National Convention, before the order for his arrest could be delivered. Paine was tried in absence, found guilty of seditious libel, and declared an outlaw; and *Rights of Man* was ordered permanently suppressed.

Paine's support of Republicanism

In France Paine hailed the abolition of the monarchy but deplored the terror against the royalists, and fought unsuccessfully to save the life of King Louis XVI, favour-

ing banishment rather than execution. He was to pay for his efforts to save the King's life when the radicals under Robespierre took power. Paine was imprisoned from December 28, 1793, to November 4, 1794, when, with the fall of Robespierre, he was released and, though seriously ill, readmitted to the National Convention.

Imprisonment in France

While in prison, the first part of Paine's *Age of Reason* was published (1794), and it was followed by Part II after his release (1796). Although Paine made it clear that he believed in a Supreme Being and as a deist opposed only organized religion, the work won him a reputation as an atheist among the orthodox. The publication of his last great pamphlet, "Agrarian Justice" (1797), with its attack on inequalities in property ownership, added to his many enemies in establishment circles.

Paine remained in France until September 1, 1802, when he sailed for the United States. He quickly discovered that his services to the country had been all but forgotten and that he was widely regarded only as the world's greatest infidel. Despite his poverty and his physical condition, worsened by occasional drunkenness, Paine continued his attacks on privilege and religious superstitions. He died in New York City on June 8, 1809, and was buried in New Rochelle on the farm given to him by the state of New York as a reward for his Revolutionary writings. Ten years later, William Cobbett, the political journalist, exhumed the bones and took them to England, where he hoped to give Paine a funeral worthy of his great contributions to humanity. But the plan misfired, and the bones were lost, never to be recovered.

Assessment

At Paine's death most American newspapers reprinted the obituary notice from the *New York Citizen*, which read in part: "He had lived long, did some good and much harm." This remained the verdict of history for over a century following his death, but in recent years the tide has turned: on January 30, 1937, *The Times* of London referred to him as "the English Voltaire," and on May 18, 1952, Paine's bust was placed in the New York University Hall of Fame.

MAJOR WORKS

POLITICAL AND RELIGIOUS: "Common Sense" (1776); "The American Crisis," 15 pt. (1776–83); "Prospects on the Rubicon" (1787); *Rights of Man*, 2 pt. (1791–92); "Letter Addressed to the Addressers on the Late Proclamation" (1792); *The Age of Reason*, 2 pt. (1794–95); "Dissertation on the First Principles of Government" (1795); "Letter to George Washington, President of the United States of America, on Affairs Public and Private" (1796); "Atheism Refuted" (1798); "Letters to the Citizens of the United States," 4 letters (dated 1802).

ECONOMIC AND FINANCIAL: "Dissertation on the Affairs of the Bank" (1786); *The Decline and Fall of the English System of Finance* (1796); "Agrarian Justice" (1797).

BIBLIOGRAPHY. The first comprehensive edition of Paine's works is that of MONCURE D. CONWAY, *The Writings of Thomas Paine*, 4 vol. (1894–96). This has been replaced by PHILIP S. FONER, *The Complete Writings of Thomas Paine*, 2 vol. (1945). Several later discoveries of Paine's writings are in A.O. ALDRIDGE, "Some Writings of Thomas Paine in Pennsylvania Newspapers," *American Historical Review*, 56:832–838 (1951). RICHARD GIMBEL, *Thomas Paine Fights for Freedom in Three Worlds* (1961), offers the best annotated bibliography of Paine's works. The first worthwhile biography, though entirely uncritical, was MONCURE D. CONWAY, *The Life of Thomas Paine*, 2 vol. (1892). There is still no full-length study of Paine that replaces Conway's biography, but a number of brief studies of the man are more valuable in terms of an overall evaluation of his contributions. Worth consulting are VERNON L. PARRINGTON, "Tom Paine: Republican Pamphleteer," in *Main Currents in American Thought*, pp. 327–341 (1927); HARRY H. CLARK'S Introduction to *Thomas Paine: Representative Selections*, rev. ed. (1961); and PHILIP S. FONER'S Introduction to *The Complete Writings*. A recent book that helps break down the tradition of the religious infidel and political demagogue is ROBIN MCKOWN, *Thomas Paine* (1962).

(P.S.F.)

Painting, Art of

The art of painting is the expression of ideas and emotions, with the creation of certain aesthetic qualities, in a two-dimensional visual language. The elements of this language—its shapes, lines, colours, tones, and textures—are used in various ways to produce sensations of volume, space, movement, and light on a flat surface. These elements are combined into expressive patterns in order to represent real or supernatural phenomena, to interpret a narrative theme, or to create wholly abstract visual relationships. The artist communicates his visual message in terms of the sensuous qualities and expressive possibilities and limitations of a particular medium, technique, and form.

Earlier cultural traditions—of tribes, religions, guilds, royal courts, and states—largely controlled the craft, form, imagery, and subject matter of painting and determined its function, whether ritualistic, devotional, decorative, entertaining, or educational. Painters were employed more as skilled artisans than as creative artists. Later, the Far East and Renaissance Europe saw the emergence of the fine artist, with the social status of scholar and courtier, who signed his work, who decided its design and often its subject and imagery, and who established a more personal, if not always amicable, relationship with his patron.

The painter as skilled artisan and fine artist

During the 19th century the painter in Western societies began to lose his social position and secure patronage. Generally, he can now reach an audience only through commercial galleries and public museums, although his work may be occasionally reproduced in art periodicals. He may be also assisted by financial awards or commissions from industry and the state. He has, however, gained the freedom to invent his own visual language and to experiment with new forms and unconventional materials and techniques. The restless endeavour to extend the boundaries of expression in Western art produces continuous international stylistic changes. The often bewildering succession of new movements in painting is further stimulated by the swift interchange of ideas by means of international art journals, travelling exhibitions, and art centres.

This article is concerned with the elements and principles of design in painting and with the various mediums, forms, imagery, subject matter, and symbolism employed or adopted or created by the painter.

This article is divided into the following sections:

ELEMENTS AND PRINCIPLES OF DESIGN

The design of a painting is its visual format: the arrangement of its lines, shapes, colours, tones, and textures into an expressive pattern. It is the sense of inevitability in this formal organization that gives a great painting its self-sufficiency and presence.

The colours and placing of the principal images in a

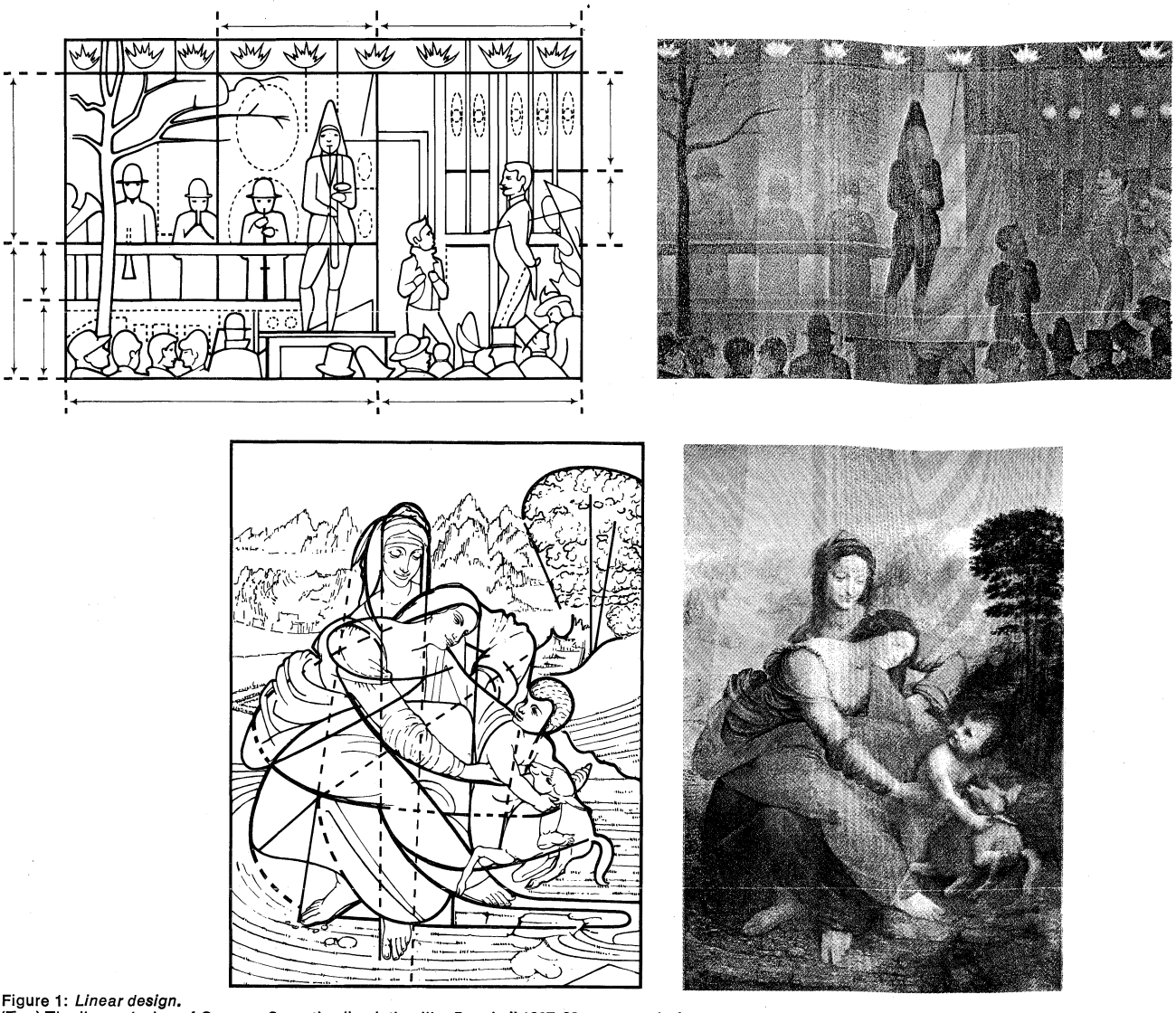


Figure 1: *Linear design.*
(Top) The linear design of Georges Seurat's oil painting "La Parade," 1887–88, composed of ovals and rectangles arranged in an overall grid pattern, based on the Golden Mean system of proportions. In the Metropolitan Museum of Art, New York. 1 m × 1.5 m. (Bottom) The interwoven, linear pattern of Leonardo da Vinci's panel painting "Virgin and Child with St. Anne," 1500. In the Louvre, Paris. 1.68 m × 1.3 m.
By courtesy of (top right) the Metropolitan Museum of Art, New York, bequest of Stephen C. Clark, 1961; photograph, (bottom right) Giraudon

Formal interplay of colours and shapes

design may be sometimes largely decided by representational and symbolic considerations. Yet, it is the formal interplay of colours and shapes that alone is capable of communicating a particular mood, producing optical sensations of space, volume, movement, and light and creating forces of both harmony and tension, even when a painting's narrative symbolism is obscure.

Elements of design. *Line.* Each of the design elements has special expressive qualities. Line, for example, is an intuitive, primeval convention for representing things; the simple linear imagery of young children's drawings and prehistoric rock paintings is universally understood. The formal relationships of thick with thin lines, of broken with continuous, and of sinuous with jagged are forces of contrast and repetition in the design of many paintings in all periods of history. Variations in the painted contours of images also provide a direct method of describing the volume, weight, spatial position, and textural characteristics of things. The finest examples of this pictorial shorthand are found in Japanese ink painting, where an expressive economy and vitality of line is closely linked to a traditional mastery of calligraphy.

In addition to painted contours, a linear design is composed of all the edges of tone and colour masses, the axial directions of images, and the lines implied by alignments of shapes across the picture. The manner in which these various kinds of line are echoed and repeated animates the design (Figure 1). The artist, consciously or intuitively, also places them in relationship to one another across the picture, so that they weave a unifying rhythmic network throughout the painting (Figure 1).

Apart from the obvious associations of some linear patterns with particular actions—undulating lines suggesting buoyant movement, for instance—emotive sensations are produced by certain linear relationships. Thus, lines moving upward express feelings of joy and aspiration, those directing the eye downward evoke moods of sadness or defeat, while lines at angles opening to the right of a design are more agreeable and welcoming than those spreading outward to the left.

Shape and mass. Shape and mass, as elements of design, include all areas of different colour, tone, and texture, as well as individual and grouped images.

Children instinctively represent the things they see by geometrical symbols. Not only have sophisticated modern artists, such as Paul Klee and Jean Dubuffet, borrowed this primitive imagery, but the more arresting and expressive shapes and masses in most styles of painting and those to which most people intuitively respond will generally be found to have been clearly based on such archetypal forms. A square or a circle will tend to dominate a design and will therefore often be found at its focal

Response to archetypal forms

centre—the square window framing Christ in Leonardo da Vinci's "Last Supper," for example, the hovering "sun" in an Adolph Gottlieb abstract, or the halo encircling a Christian or Buddhist deity. A firmly based triangular image or group of shapes seems reassuring, even uplifting, while the precarious balance implied by an inverted triangular shape or mass produces feelings of tension. Oval, lozenge, and rectangular forms suggest stability and protection and often surround vulnerable figures in narrative paintings.

There is generally a cellular unity, or "family likeness," between the shapes and masses in a design similar to the visual harmony of all units to the whole observed in natural forms—the gills, fins, and scales in character with the overall shape of a fish, for example.

The negative spaces between shapes and masses are also carefully considered by the artist, since they can be so adjusted as to enhance the action and character of the positive images. They can be as important to the design as time intervals in music or the voids of an architectural facade.

Colour. In many styles and periods of painting, the functions of colour are primarily decorative and descriptive, often serving merely to reinforce the expression of an idea or subject communicated essentially in terms of line and tone. In much of 20th-century painting, however, colour has gained in importance and is the primary expressive element.

The principal dimensions of colour in painting are the variables or attributes of hue, tone, and intensity. Red, yellow, and blue are the basic hues from which all others on the chromatic scale can be made by mixtures. These three opaque hues are the subtractive pigment primaries and should not be confused with the behaviour of the additive triads and mixtures of transparent, coloured light. Mixtures of primary pairs produce the secondary hues of orange, violet, and green. By increasing the amount of one primary in each of these mixtures, the tertiary colours of yellow-orange, orange-red, red-violet, violet-blue, blue-green, and green-yellow, respectively, are made. The primary colours, with their basic secondary and tertiary mixtures, can be usefully notated as the 12 segments of a circle (Figure 2). The secondary and tertiary colour segments between a pair of parent primaries can then be seen to share a harmonious family relationship with one another—the yellow-orange, orange, and orange-red hues that lie between yellow and red, for example.

Local hues are the inherent and associative colours of things. In everyday life, familiar things are described by particular colours, and these often are identified by reference to familiar things; the green of grass and the grass green of paint, for instance. Although, as the Impressionists demonstrated, the inherent colours of forms in the real world are usually changed by effects of light and atmosphere, many of the great primitive and classical styles of representational painting are expressed in terms of local hues.

Tone is a colour's relative degree, or value, of lightness or darkness. The tonal pattern of a painting is shown in a monochrome reproduction. A painting dominated by dark colours, such as a Rembrandt, is in a low tonal key, while one painted in the pale range of a late Claude Monet is said to be high keyed. The tonal range of pigments is too narrow for the painter to be able to match the brightest lights and deepest darks of nature. Therefore, in order to express effects of illumination and dense shadow, he must lower the overall tonal key of his design, thus intensifying the brightness value of his lightest pigment colours.

Tonal pattern of a painting

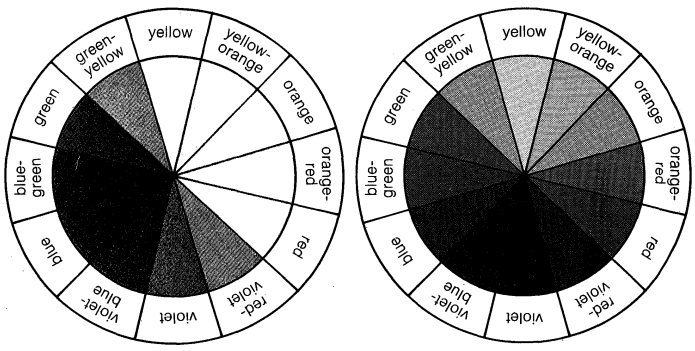


Figure 2: *Colour.*
(Left) Colour wheel made up of the primary colours and their basic secondary and tertiary mixtures. (Right) Colour wheel with approximate, inherent tonal values.

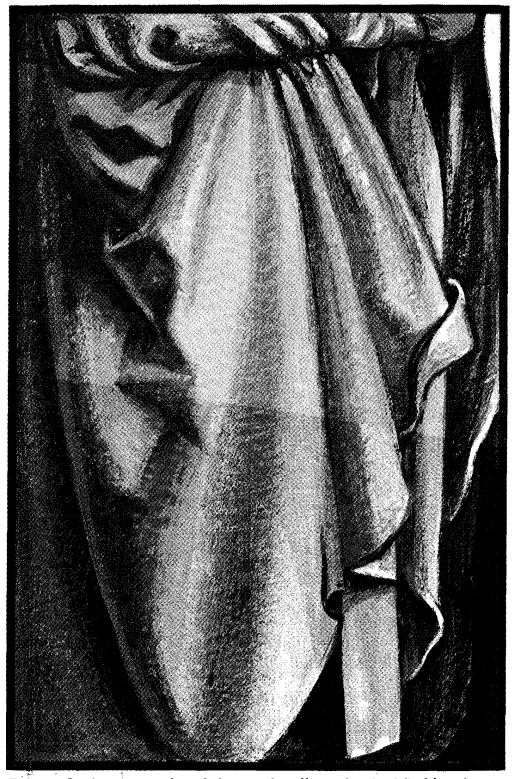


Figure 3: An example of the early oil method of (left) colour glazing a (right) monochrome painting.

The Greco-Roman, Renaissance, and Neoclassical method of representing volume and space in painting was by a system of notated tonal values, the direction of each plane in the design being indicated by a particular degree of lightness or darkness. Each tonal value was determined by the angle at which a plane was meant to appear to turn away from an imaginary source of light. The tonal modelling, or shading, of forms was often first completed in a monochrome underpainting. This was then coloured with transparent washes of local hues, a technique similar to that of colour tinting a black-and-white photograph (Figure 3).

Each hue has an intrinsic tonal value in relation to others on the chromatic scale; orange is inherently lighter than red, for instance, and violet is darker than green (Figure 2). Any reversal of this natural tonal order creates a colour discord. An optical shock is therefore produced when orange is juxtaposed with pink (a lighter tone of red) or pale violet is placed against dark green. Such contrasts as these are deliberately created in paintings for the purpose of achieving these dramatic and disturbing effects.

The intensity of a colour is its degree of purity or hue saturation. The colour of a geranium, therefore, is said to be more intense, more highly saturated with pure orange-red than is mahogany. The pigment vermilion is orange-red at maximum intensity; the brown earth pigment burnt sienna is grayer and has a lower degree of orange-red saturation.

Intense hues are termed chromatic colours. The achromatic range is made up of hues reduced in intensity by the addition of white, making the tints, or pastel colours, such as cream and pink; or of black, producing the shades, or earth colours, such as mustard and moss green; or of both

white and black, creating the neutralized hues, or colour-tinged grays, such as oatmeal and charcoal.

An achromatic colour will seem more intense if it is surrounded by neutralized hues or juxtaposed with its complementary colour. Complementaries are colour opposites. The complementary colour to one of the primary hues is the mixture of the other two; the complementary to red pigment, for example, is green—that is, blue mixed with yellow. The colour wheel (Figure 2) shows that the tertiaries also have their colour opposites, the complementary to orange-red, for instance, being blue-green. Under clear light the complementary to any chroma, shade, or tint can be seen if one "fixates," or stares at one colour intently for a few seconds then looks at a neutral, preferably white, surface. The colour afterimage will appear to glow on the neutral surface (Figure 4). Mutual

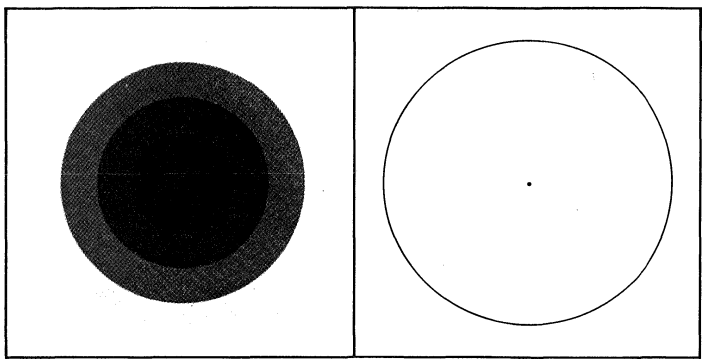


Figure 4: Coloured afterimages: if a person stares for about 30 seconds at the coloured disk under a clear light and then fixes upon the empty space of the adjacent circle, coloured afterimages will appear.

Enhancement of colour intensity

enhancement of colour intensity results from juxtaposing a complementary pair, red becoming more intensely red, for instance, and green more fiercely green when these are contiguous than either would appear if surrounded by harmonious hues. The 19th-century physicist Michel-Eugène Chevreul referred to this mutual exaltation of opposites as the law of simultaneous contrast. Chevreul's second law, of successive contrast, referred to the optical sensation that a complementary colour halo appears gradually to surround an intense hue. This complementary glow is superimposed on surrounding weaker colours, a gray becoming greenish when juxtaposed with red, reddish in close relationship with green, yellowish against violet, and so on (Figure 5).

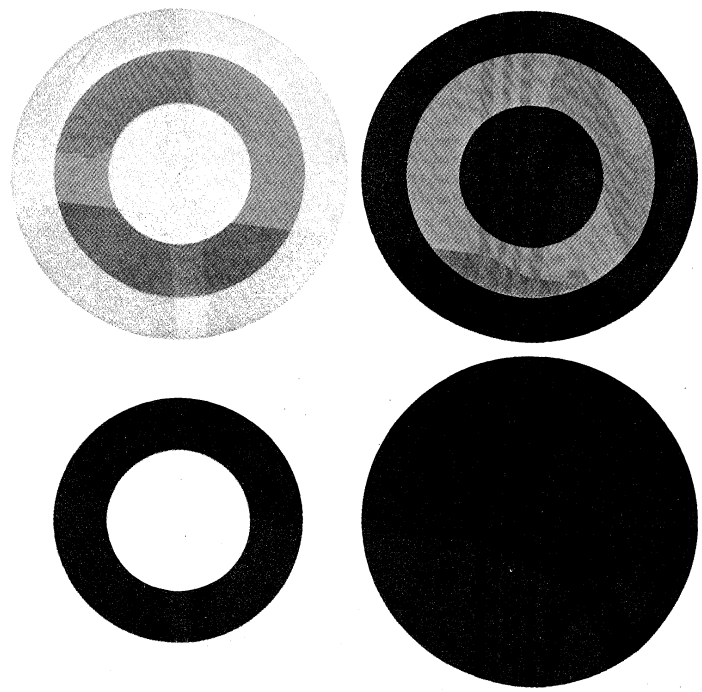


Figure 5: *Optical colour change.*
(Top) By complementary action, the same gray pigment will appear greenish when adjacent to red but reddish if adjacent to green. (Bottom) A green hue will seem cool if surrounded by yellow, but warm when surrounded by blue-green.

Hues containing a high proportion of blue (the violet to green range) appear cooler than those with a high content of yellow or red (the green-yellow to red-violet range). This difference in the temperature of hues in a particular painting is, of course, relative to the range and juxtaposition of colours in the design. A green will appear cool if surrounded by intense yellow, while it will seem warm against blue-green (Figure 5). The optical tendency for warm colours to advance before cold had been long exploited by European and Oriental painters as a method of suggesting spatial depth. Changes in temperature and intensity can be observed in the atmospheric effects of nature, where the colours of distant forms become cooler, grayer, and bluish, while foreground planes and features appear more intense and usually warmer in colour.

The apparent changes in a hue as it passes through zones of different colour has enabled painters in many periods to create the illusion of having employed a wide range of pigment hues with, in fact, the use of very few. And, although painters had applied many of the optical principles of colour behaviour intuitively in the past, the publication of research findings by Chevreul and others stimulated the Neo-Impressionists and Post-Impressionists and the later Orphist and Op Art painters to extend systematically the expressive possibilities of these principles in order to create illusions of volume and space and vibrating sensations of light and movement. Paul Cézanne, for example, demonstrated that subtle changes in the surface of a form and in its spatial relationship to others could be expressed primarily in facets of colour, modulated by varying degrees of tone, intensity, and temperature and by the introduction of complementary colour accents.

While the often-complex religious and cultural colour symbologies may be understood by very few, the emotional response to certain colour combinations appears to be almost universal. Optical harmonies and discords seem to affect everyone in the same way, if in varying degrees. Thus, an image repeated in different schemes of colour will express a different mood in each change (Figure 6).

Texture. Pointillism was a term given to the Neo-Impressionist system of representing the shimmer of atmospheric light with spots of coloured pigment. This technique produced an overall granular texture. As an element of design, texture includes all areas of a painting enriched or animated by vibrating patterns of lines, shapes, tones, and colours, in addition to the tactile textures created by the plastic qualities of certain mediums. Decorative textures may be of geometrical repeat patterns, as in much of Indian, Islāmic, and medieval European painting and other art, or of representations of patterns in nature, such as scattered leaves, falling snow, and flights of birds.

Perceptual and conceptual levels of understanding spatial relationships

Volume and space. The perceptual and conceptual methods of representing volume and space on the flat surface of a painting are related to the two levels of understanding spatial relationships in everyday life.

Perceptual space is the view of things at a particular time and from a fixed position. This is the stationary window view recorded by the camera and represented in the later periods of ancient Greek and Roman paintings and in most Western schools of painting since the Renaissance. Illusions of perceptual space are generally created by use of the linear perspectival system, based on the observations that objects appear to the eye to shrink and parallel lines and planes to converge as they approach the horizon, or viewer's eye level (Figure 7).

Young children and primitive artists, however, do not understand space in this way and represent it conceptually. Their paintings, therefore, show objects and surroundings independently of one another and from the views that best present their most characteristic features. The notion of scale in their pictures is also subjective, the

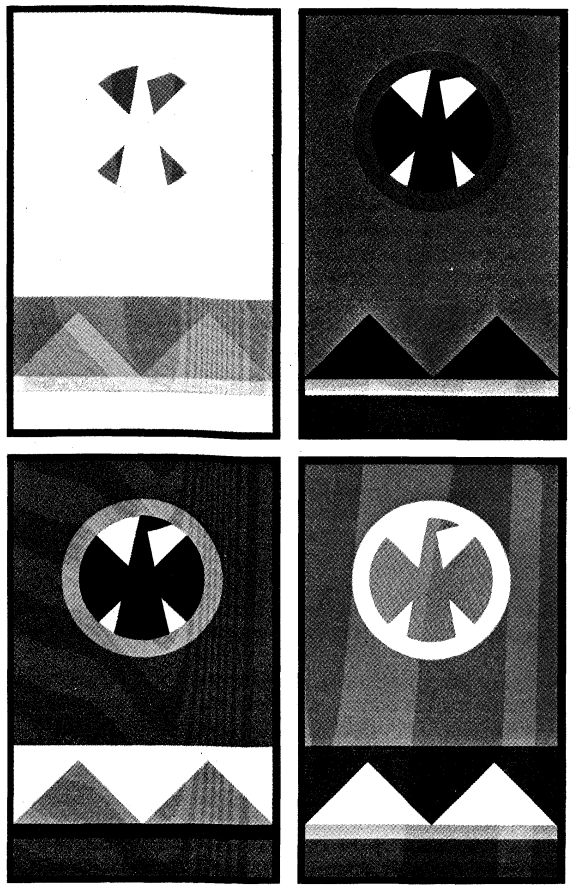


Figure 6: Emotive colour relationships: an identical pattern of shapes may express a different emotional mood through each colour variation.

relative size of things being decided by the artist either by their degree of emotional significance for him or by their narrative importance in the picture (interest perspective).

The conceptual, polydimensional representation of space has been used at some period in most cultures. In much of ancient Egyptian and Cretan painting, for example, the head and legs of a figure were shown in profile, but the eye and torso were drawn frontally. And in Indian, Islāmic, and pre-Renaissance European painting, vertical forms and surfaces were represented by their most informative elevation view (as if seen from ground level), while the horizontal planes on which they stood were shown in isometric plan (as if viewed from above) (Figure 7). This system produces the overall effect that objects and their surroundings have been compressed within a shallow space behind the picture plane.

By the end of the 19th century Cézanne had flattened the conventional Renaissance picture space (Figure 7), tilting horizontal planes so that they appeared to push vertical forms and surfaces forward from the picture plane and toward the spectator (Figure 8). This illusion of the picture surface as an integrated structure in projecting low relief was developed further in the early 20th century by the Cubists. The conceptual, rotary perspective of a Cubist painting shows not only the components of things from different viewpoints but presents every plane of an object and its immediate surroundings simultaneously. This gives the composite impression of things in space that is gained by having examined their surfaces and construction from every angle.

In Pop Art painting, both conceptual and perceptual methods of representing space are often combined. And, where the orbital movement of forms in European design since the Renaissance was intended to hold the spectator's attention within the frame (Figure 7), the expanding picture space in recent mural-size abstract paintings directs his eye outward to the surrounding wall, and their shapes and colours seem about to invade his own territory (Figure 9).

Time and movement. Time and movement in painting are not restricted to representations of physical energy but are elements of all design. Part of the viewer's full experience of a great painting is to allow the arrangement of lines, shapes, and accents of tone or colour to guide him across the picture surface at controlled tempos and rhythmic directions that contribute to the expression of a particular mood, vision, and idea.

Kinetic sensations on a flat surface

Centuries before cinematography, painters attempted to produce kinetic sensations on a flat surface. A mural of 2000 BC in an Egyptian tomb at Beni Hasan, for instance, is designed as a continuous strip sequence of wrestling holds and throws, so accurately articulated and notated that it might be photographed as an animated film cartoon. The gradual unrolling of a 12th-century Japanese hand scroll produces the visual sensation of a helicopter flight along a river valley, while the experience of walking to the end of a long, processional Renaissance mural by Andrea Mantegna or Benozzo Gozzoli is similar to that of having witnessed a passing pageant as a standing spectator.

In the Eastern and Western narrative convention of continuous representation, various incidents in a story were depicted together within one design, the chief characters

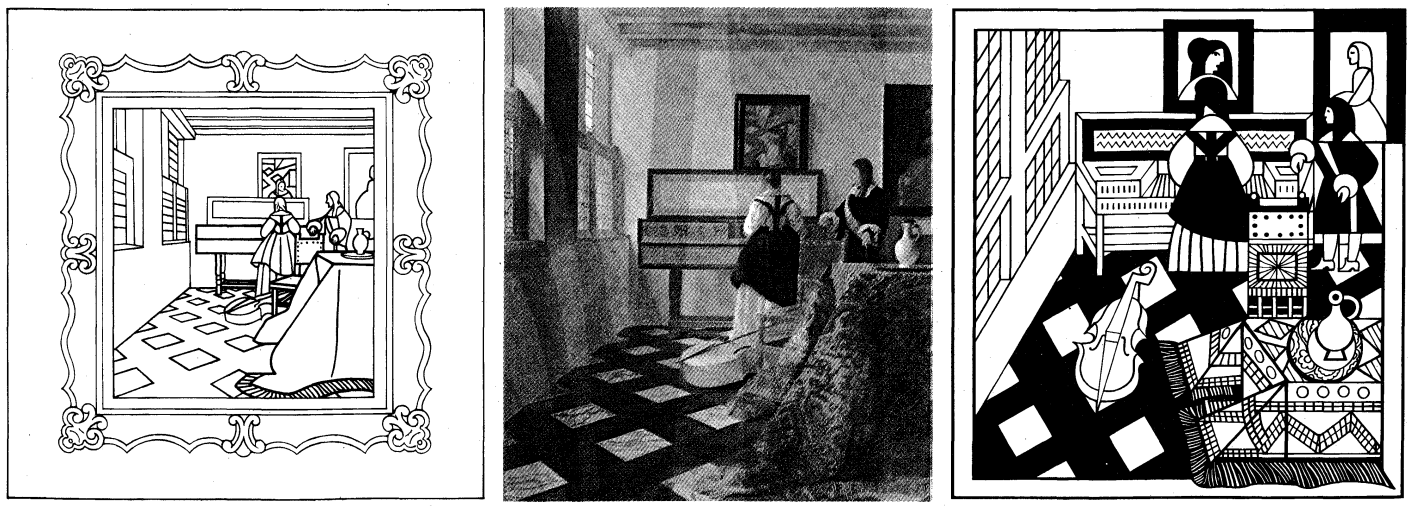


Figure 7: *Perceptual and conceptual space.*
(Left) Perceptual space: the illusion of an interior space as it might be recorded by a camera and as it is represented by the optical laws of Renaissance perspective. (Centre) Example based on Jan Vermeer's oil painting "The Music Lesson," *c.* 1660. In Buckingham Palace, London. 74 cm × 64 cm. (Right) Conceptual space: how the same subject might be represented by polydimensional and "interest" perspectival systems.

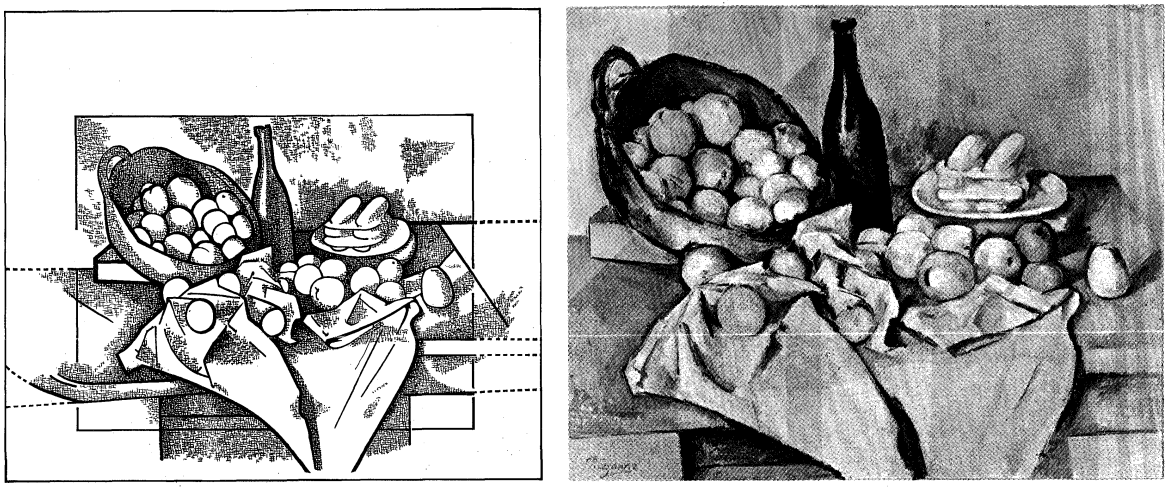


Figure 8: *Simultaneous viewpoints.*
(Left) Paul Cézanne's method of representing things in a painting as if seen from different directions and at varying eyelevels, over a period of time. In contrast to the receding planes of Vermeer's design (Figure 7), the illusion is created, by tone and colour contrasts, that planes and forms are advancing toward the spectator. (Right) Example based on Cézanne's oil painting "The Basket of Apples," 1890–94. In the Art Institute of Chicago. 65 cm × 81 cm.
(Right) By courtesy of the Art Institute of Chicago, Helen Birch Bartlett Memorial Collection

in the drama easily identified as they reappeared in different situations and settings throughout the painting. In Byzantine murals and in Indian and medieval manuscript paintings, narrative sequences were depicted in grid patterns, each "compartment" of the design representing a visual chapter in a religious story or a mythological or historical epic.

The Cubists aimed to give the viewer the time experience of moving around static forms in order to examine their volume and structure and their relationships to the space surrounding them. In paintings such as "Nude Descending a Staircase," "Girl Running on a Balcony," and "Dog on Leash," Marcel Duchamp and Giacomo Balla combined the Cubist technique of projected, interlocking planes with the superimposed time–motion sequences of cinematography. This technique enabled the artists to analyze the structural mechanics of forms, which are represented as moving in space past the viewer.

Principles of design. Because painting is a two-dimensional art, the flat pattern of lines and shapes is an important aspect of design, even for those painters concerned with creating illusions of great depth. And, since any mark made on the painting surface can be perceived as a spatial statement—for it rests upon it—there are also qualities of three-dimensional design in paintings composed primarily of flat shapes. Shapes in a painting, therefore, may be balanced with one another as units of a flat pattern and considered at the same time as components in a spatial design, balanced one behind another. A symmetrical balance of tone and colour masses of equal weight creates a serene and sometimes monumental design, while a more dynamic effect is created by an asymmetrical balance.

Geometrical shapes and masses are often the basic units in the design of both "flat patterns," such as Byzantine and Islāmic paintings, and "sculptural compositions," such as Baroque and Neoclassical figure tableaux. The flat, overlapping squares, circles, and triangles that create the pattern of a Romanesque mural, for example, become the interlocking cubic, spherical, and pyramidal components that enclose the grouped figures and surrounding features in a Renaissance or a Neoclassical composition (Figure 10).

An emphasis upon the proportion of the parts to the whole is a characteristic of Classical styles of painting. The Golden Mean, or Section, has been used as an ideal proportion on which to base the framework of lines and shapes in the design of a painting. The Renaissance mathematician Lucas Pacioli defined this aesthetically satisfying ratio as the division of a line so that the shorter part is to the longer as the longer is to the whole (approximately 8 to 13). His treatise (*Divina proportione*) influenced Leonardo da Vinci and Albrecht Dürer. The Neo-Impressionists Georges Seurat (Figure 1) and Paul Signac based the linear pattern of many of their compositions upon the principle of this "divine proportion." Golden Mean proportions can be discovered in the design of many other styles of painting, although often they may have been created more by intuitive judgment than by calculated measurement.

Tension in paintings

Tension is created in paintings, as it is experienced in everyday life, by the anticipation of an event or by an unexpected change in the order of things. Optical and psychological tensions occur in passages of a design, therefore, when lines or shapes almost touch or seem about to collide, when a harmonious colour progression is interrupted by a sudden discord, or when an asymmetrical balance of lines, shapes, tones, or colours is barely held.

Contrasts in line, shape, tone, and colour create vitality; rectilinear shapes played against curvilinear, for instance, or warm colours against cool. Or a painting may be composed in contrasted overall patterns, superimposed in counterpoint to one another—a colour scheme laid across contrasting patterns of lines and tones, for example (Figure 11).

Design relationships between painting and other visual arts. The philosophy and spirit of a particular period in painting usually have been reflected in many of its other visual arts. The ideas and aspirations of the ancient cul-

By courtesy of the trustees of the Tate Gallery, London; photograph, A.C. Cooper Ltd.

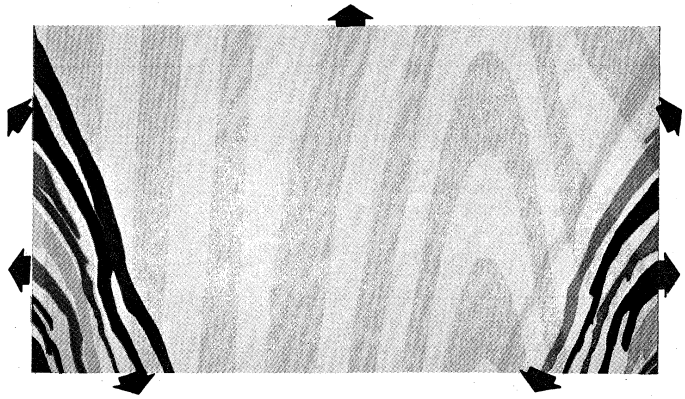


Figure 9: Opening out the picture space: the movement of shapes in Morris Louis' acrylic painting "Alpha-Phi," 1961, directs the spectator's eye outside the picture surface. In the Tate Gallery, London. 2.59 m × 4.58 m. In contrast, Vermeer's design (Figure 7) is contrived to hold the spectator's attention within the frame.

tures, of the Renaissance, Baroque, Rococo, and Neoclassical periods of Western art and, more recently, of the 19th-century Art Nouveau and Secessionist movements were expressed in much of the architecture, interior design, furniture, textiles, ceramics, dress design, and handicrafts, as well as in the fine arts, of their times. Following the Industrial Revolution, with the redundancy of handcraftmanship and the loss of direct communication between the fine artist and society, idealistic efforts to unite the arts and crafts in service to the community were made by William Morris in Victorian England and by the Bauhaus in 20th-century Germany. Although their aims were not fully realized, their influences, like those of the short-lived de Stijl and Constructivist movements, have been far-reaching, particularly in architectural, furniture, and typographic design.

Michelangelo and Leonardo da Vinci were painters, sculptors, and architects. Although no artists since have excelled in so wide a range of creative design, leading 20th-century painters have expressed their ideas in many other mediums. In graphic design, for example, Pierre Bonnard, Henri Matisse, and Raoul Dufy produced posters and illustrated books; André Derain, Fernand Léger, Marc Chagall, Mikhail Larionov, and Kazimir Malevich designed for the theatre; Joan Miró, Georges Braque, and Chagall worked in ceramics; Braque and Salvador Dalí designed jewelry; and Dalí, Hans Richter, and Andy Warhol have made films. Many of these, with other modern painters, have also been sculptors and printmakers and have designed for textiles, tapestries, mosaics, and stained glass, while there are few mediums of the visual arts that Pablo Picasso has not worked in and revitalized.

In turn, painters have been stimulated by the imagery, techniques, and design of other visual arts. One of the earliest of these influences was possibly from the theatre, where the ancient Greeks are thought to have been the first to employ the illusions of optical perspective. The discovery or reappraisal of design techniques and imagery in the art forms and processes of other cultures has been an important stimulus to the development of more recent styles of Western painting, whether or not their traditional significance have been fully understood. The influence of Japanese woodcut prints on Synthetism and the Nabis, for example, and of African sculpture on Cubism and the German Expressionists helped to create visual vocabularies and syntax with which to express new visions and ideas. The invention of photography introduced painters to new aspects of nature, while eventually prompting others to abandon representational painting altogether. Painters of everyday life, such as Edgar Degas, Henri de Toulouse-Lautrec, Edouard Vuillard, and Bonnard, exploited the design innovations of camera cutoffs, close-ups, and unconventional viewpoints in order to give the spectator the sensation of sharing an intimate picture space with the figures and objects in the painting.

Influence of photography

TECHNIQUES AND METHODS

Whether a painting reached completion by careful stages or was executed directly by a hit or miss alla prima method (in which pigments are laid on in a single application) was once largely determined by the ideals and established techniques of its cultural tradition. For example, the medieval European illuminator's painstaking procedure, by which a complex linear pattern was gradually enriched with gold leaf and precious pigments, was contemporary with the Sung Chinese Zen practice of immediate, calligraphic brush painting, following a contemplative period of spiritual self-preparation. More recently, the artist has himself decided the technique and working method best suited to his aims and temperament. In France in the 1880s, for instance, Seurat might be working in his studio on innumerable drawings, tone studies, and colour schemes in preparation for a large composition at the same time that, outdoors, Monet would be endeavouring to capture directly the transitory effects of afternoon light and atmosphere, while Cézanne analyzed the structure of the Montagne (mountain) Sainte-Victoire with deliberated brush strokes, laid as irrevocably as mosaic tesserae (small pieces, such as marble or tile).

The kind of relationship established between artist and patron, the site and subject matter of a painting commission, and the physical properties of the medium employed may also dictate working procedure. Peter Paul Rubens, for example, followed the business-like 17th-century custom of submitting a small oil sketch, or *modella*, for his client's approval before carrying out a large-scale com-

(Left) Archivo Mas, Barcelona, (right) SCALA, New York

Figure 10: *Principles of design.*
(Left) The flat pattern design of the anonymous Spanish panel painting "Virgin and Child," 12th century. In the Museo Arqueológico Artístico Episcopal, Vich, Spain. (Right) The three-dimensional design of interlocking pyramids of Andrea del Sarto's panel painting "Madonna of the Harpies," 1517. In the Uffizi, Florence. 2.07 m × 1.78 m.

Figure 11: *Visual counterpoint in design.*
(Left) Linear arabesques drawn across (centre) patterns of tone and colour. (Right) Example based on Henri Toulouse-Lautrec's oil paint on cardboard painting "At the Nouveau Cirque: Five Stuffed Shirts," 1891. In the Philadelphia Museum of Art. 119 cm × 87 cm.
(Right) By courtesy of the Philadelphia Museum of Art, John D. McIlhenny Fund

mission. Siting problems peculiar to mural painting, such as spectator eye level and the scale, style, and function of a building interior, had first to be solved in preparatory drawings and sometimes with the use of wax figurines or scale models of the interior. Scale working drawings are essential to the speed and precision of execution demanded by quick-drying mediums, such as buon' fresco (see below *Mediums*: *Fresco*) on wet plaster and acrylic resin on canvas. The drawings traditionally are covered with a network of squares, or "squared-up," for enlarging on the surface of the support. Some modern painters prefer to outline the enlargement of a sketch projected directly onto the support by epidiascope (a projector for images of both opaque and transparent objects).

In Renaissance painters' workshops, pupil assistants not only ground and mixed the pigments and prepared the supports and painting surfaces but often laid in the outlines and broad masses of the painting from the master's design and studies.

Preservation of a painting

The inherent properties of its medium or the atmospheric conditions of its site may themselves preserve a painting. The wax solvent binder of encaustic paintings (in which after application, the paint is fixed by heat [see below *Mediums*], for example) both retains the intensity and tonality of the original colours and protects the surface from damp. And, while prehistoric rock paintings and buon' frescoes are preserved by natural chemical action, the tempera pigments thought to be bound only with water on many ancient Egyptian murals are protected by the dry atmosphere and unvarying temperature of the tombs. It has, however, been customary to varnish oil paintings, both to protect the surface against damage by dirt and handling and to restore the tonality lost when some darker pigments dry out into a higher key. Unfortunately, varnish tends to darken and yellow with time into the sometimes disastrously imitated, "Old Masters' mellow patina." Once cherished, this amber-gravy film is now generally removed to reveal the colours in their original intensity. Glass began to replace varnish toward the end of the 19th century, when painters wished to retain the fresh, luminous finish of pigments applied directly to a pure white ground. The air-conditioning and temperature-control systems of modern museums make both varnishing and glazing unnecessary, except for older and more fragile exhibits.

The frames surrounding early altarpieces, icons, and cassone panels (painted panels on the chest used for a bride's household linen) were often structural parts of the support. With the introduction of portable easel pictures, heavy frames not only provided some protection against theft and damage but were considered an aesthetic enhancement to a painting, and frame making became a specialized craft. Gilded gesso moldings (consisting of plaster of paris and sizing that forms the surface for low relief) in extravagant swags of fruit and flowers certainly seem almost an extension of the restless, exuberant design of a Baroque or Rococo painting. A substantial frame also provided a proscenium (in a theatre, the area between the orchestra and the curtain) in which the picture was isolated from its immediate surroundings, thus adding to the window view illusion intended by the artist. Deep, ornate frames are unsuitable for many modern paintings, where the artist's intention is for his forms to appear to advance toward the spectator rather than be viewed by him as if through a wall aperture. In contemporary Minimal paintings, no effects of spatial illusionism are intended; and, in order to emphasize the physical shape of the support itself and to stress its flatness, these abstract, geometrical designs are displayed without frames or are merely edged with thin protective strips of wood or metal.

MEDIUMS

By technical definition, mediums are the liquids added to paints to bind them and make them workable. They are discussed here, however, in the wider meaning of all the various paints, tools, supports, surfaces, and techniques employed by painters. The basis of all paints is variously coloured pigment, ground to a fine powder. The different expressive capacities and characteristic final surface texture of each medium are determined by the vehicle with which it is bound and thinned, the nature and surface preparation of the support, and the tools and technique with which it is handled.

Sources of pigments

Pigments are derived from various natural and artificial sources. The oldest and most permanent pigments are the blacks, prepared from bone and charcoal, and the clay earths, such as raw umber and raw sienna, which can be changed by heating into darker, warmer browns. In early periods of painting, readily available pigments were few. Certain intense hues were obtainable only from the rarer minerals, such as cinnabar (orange-red vermilion), lapis lazuli (violet-blue ultramarine), and malachite (green). These were expensive and therefore reserved for focal accents and important symbolic features in the design. The opening of trade routes and the manufacture of synthetic substitutes gradually extended the range of colours available to painters.

Tempera. A tempera medium is dry pigment tempered with an emulsion and thinned with water. It is a very ancient medium, having been in constant use in most world cultures, until in Europe it was gradually superseded, during the Renaissance, by oil paints. Tempera was the original mural medium in the ancient dynasties of

Egypt, Babylonia, Mycenean Greece, and China and was used to decorate the early Christian catacombs. It was employed on a variety of supports, from the stone stelae (or commemorative pillars), mummy cases, and papyrus rolls of ancient Egypt to the wood panels of Byzantine icons and altarpieces and the vellum leaves of medieval illuminated manuscripts.

True tempera is made by mixture with the yolk of fresh eggs, although manuscript illuminators often used egg white and some easel painters added the whole egg. Other emulsions have been used, such as casein glue with linseed oil, egg yolk with gum and linseed oil, and egg white with linseed or poppy oil. Individual painters have experimented with other recipes, but few of these have proved successful; all but William Blake's later tempera paintings on copper sheets, for instance, have darkened and decayed, and it is thought that he mixed his pigment with carpenter's glue.

Distemper is a crude form of tempera made by mixing dry pigment into a paste with water, which is thinned with heated glue in working or by adding pigment to whiting, a mixture of fine-ground chalk and size. It is used for stage scenery and full-size preparatory cartoons for murals and tapestries. When dry, its colours have the pale, mat, powdery quality of pastels, with a similar tendency to smudge. Indeed, damaged cartoons have been retouched with pastel chalks.

Egg tempera is the most durable form of the medium, being generally unaffected by humidity and temperature. It dries quickly to form a tough film that acts as a protective skin to the support. In handling, in its diversity of transparent and opaque effects, and in the satin sheen of its finish, it resembles the modern acrylic resin emulsion paints.

Traditional process of tempera painting

Traditional tempera painting is a lengthy process. Its supports are smooth surfaces, such as planed wood, fine set plaster, stone, paper, vellum, canvas, and modern composition boards of compressed wood or paper. Linen is generally glued to the surface of panel supports, additional strips masking the seams between braced wood planks. Gesso, a mixture of plaster of paris (or gypsum) with size, is the traditional ground. The first layer is of gesso grosso, a mixture of coarse, unslaked plaster and size. This provides a rough, absorbent surface for ten or more thin coats of gesso sotile, a smooth mixture of size and fine plaster previously slaked in water to retard drying. This laborious preparation results, however, in an opaque, brilliant white, light-reflecting surface, similar in texture to hard, flat icing sugar.

The design for a large tempera painting traditionally was executed in distemper on a thick paper cartoon. The outlines were pricked with a perforating wheel so that when the cartoon was laid on the surface of the support, the linear pattern was transferred by dabbing, or "pouncing," the perforations with a muslin bag of powdered charcoal. The dotted contours traced through were then fixed in paint. Medieval tempera painters of panels and manuscripts made lavish use of gold leaf on backgrounds and for symbolic features, such as haloes and beams of heavenly light. Areas of the pounced design intended for gilding were first built up into low relief with gesso duro, the harder, less absorbent gesso compound used also for elaborate frame moldings. Background fields were often textured by impressing the gesso duro, before it set, with small, carved, intaglio wood blocks, to create raised, pimpled, and quilted repeat patterns that glittered when gilded. Leaves of finely beaten gold were pressed onto a tacky mordant (adhesive compound) or over wet bole (reddish-brown earth pigment) that gave greater warmth and depth when the gilded areas were burnished.

Colours were applied with sable brushes in successive broad sweeps or washes of semitransparent tempera. These dried quickly, preventing the subtle tonal gradations possible with watercolour washes or oil paint; effects of shaded modelling had therefore to be obtained by a crosshatching technique of fine brush strokes. According to the Italian painter Cennino Cennini, the early Renaissance tempera painters laid the colour washes across a fully modelled monochrome underpainting in terre vert (olive-green pigment), a method developed later into the mixed mediums technique of tempera underpainting followed by transparent oil glazes.

The luminous gesso base of a tempera painting, combined with the accumulative effect of overlaid colour washes, produces a unique depth and intensity of colour. Tempera paints dry lighter in value, but their original tonality can be restored by subsequent waxing or varnishing. Other characteristic qualities of a tempera painting, resulting from its fast drying property and disciplined technique, are its steely lines and crisp edges, its meticulous detail and rich linear textures, and its overall emphasis upon a decorative flat pattern of bold colour masses.

The great Byzantine tradition of tempera painting was developed in Italy in the 13th and 14th centuries by Duccio di Buoninsegna and Giotto. Their flattened picture space, generously enriched by fields and textures of gold leaf, was extended by the Renaissance depth perspectives in the paintings of Giovanni Bellini, Piero della Francesca, Carlo Crivelli, Sandro Botticelli, and Vittore Carpaccio. By that time, oil painting was already challenging the primacy of tempera, Botticelli and some of his contemporaries apparently adding oil to the tempera emulsion or overglazing it in oil colour.

20th-century revival of tempera techniques

Following the supremacy of the oil medium during succeeding periods of Western painting, the 20th century saw a revival of tempera techniques by such U.S. artists as Ben Shahn, Andrew Wyeth, and George McNeil and by the British painter Edward Wadsworth. It would probably have been the medium also of the later hard-edge abstract painters had the new acrylic resin paints not proved more easily and quickly handled.

Fresco. Fresco (Italian: "fresh") is the traditional medium for painting directly onto a wall or ceiling. It is the oldest known painting medium, surviving in the prehistoric mural decorations and perfected in 16th-century Italy in the buon' fresco method.

The cave paintings are thought to date from 13,000–20,000 BC. Their pigments probably have been preserved by a natural sinter process of rainwater seeping through the limestone rocks to produce saturated bicarbonate. The colours were rubbed across rock walls and ceilings with sharpened solid lumps of the natural earths, (yellow, red, and brown ochre). Outlines were drawn with black sticks of wood charcoal. The discovery of mixing dishes suggests that liquid pigment mixed with fat was also used and smeared with the hand. The subtle tonal gradations of colour on animals painted in the Altamira and Lascaux caves appear to have been dabbed in two stages with fur pads, natural variations on the rock surface being exploited to assist in creating effects of volume. Feathers and frayed twigs may have been used in painting manes and tails.

These were not composite designs but separate scenes and individual studies that, like graffiti drawings, were added at different times, often one above another, by various artists. Paintings from the Magdalenian period (*c.* 10,000 BC) exhibit astonishing powers of accurate observation and ability to represent movement. Women, warriors, horses, bison, bulls, boars, and ibex are depicted in scenes of ritual ceremony, battle, and hunting. Among the earliest images are imprinted and stencilled hands. Vigorous meanders, or "macaroni" linear designs, were traced with fingers dipped in liquid pigment.

Fresco secco. In the fresco secco, or lime-painting, method, the plastered surface of a wall is soaked with slaked lime. Lime-resistant pigments are applied swiftly before the plaster sets. Secco colours dry lighter than their tone at the time of application, producing the pale, mat, chalky quality of a distempered wall. Although the pigments are fused with the surface, they are not completely absorbed and may flake in time, as in sections of Giotto's 14th-century S. Francesco murals at Assisi. Secco painting was the prevailing medieval and early Renaissance medium and was revived in 18th-century Europe by artists such as Giovanni Battista Tiepolo, François Boucher, and Jean-Honoré Fragonard.

Buon' fresco. Buon', or "true," fresco is the most durable method of painting murals, since the pigments are

completely fused with a damp plaster ground to become an integral part of the wall surface. The stone or brick wall is first prepared with a brown trullisatio scratch coat, or rough-cast plaster layer. This is then covered by the arricciato coat, on which the linear design of the preparatory cartoon is pounced (see above *Tempera*) or engraved by impressing the outlines into the moist, soft plaster with a bone or metal stylus. These lines were usually overworked in reddish sinopia pigment. A thin layer of fine plaster is then evenly spread, allowing the linear design to show through. Before this final intonaco ground sets, pigments thinned with water or slaked lime are applied rapidly with calf-hair and hog-bristle brushes; depth of colour is achieved by a succession of quick-drying glazes. Being prepared with slaked lime, the plaster becomes saturated with an aqueous solution of hydrate of lime, which takes up carbonic acid from the air as it soaks into the paint. Carbonate of lime is produced and acts as a permanent pigment binder. Pigment particles crystallize in the plaster, fusing it with the surface to produce the characteristic lustre of buon' fresco colours. When dry, these are mat and lighter in tone. Colours are restricted to the range of lime-resistant pigments.

Arricciato coat

The intonaco coat is laid only across an area sufficient for painting before the plaster sets. The joins between each successive "day piece" are sometimes visible. Alterations must be made by immediate washing or scraping; minor retouching to set plaster is possible with casein or egg tempera, but major corrections necessitate breaking away the intonaco and replastering. The swift execution demanded stimulates bold designs in broad masses of colour with a calligraphic vitality of flecked, spotted, and scribbled brush marks.

No ancient Greek buon' frescoes now exist, but forms of the technique survive in the Pompeian villas of the 2nd century AD, in Chinese tombs at Liao-yang, Manchuria, and in the 6th-century Indian caves at Ajantā. Among the finest buon' fresco murals are those by Michelangelo in the Sistine Chapel and by Raphael in the Stanze of the Vatican. Other notable examples from the Italian Renaissance can be seen in Florence: painted by Andrea Orcagna in the Museo dell'Opera di Sta. Croce, by Gozzoli in the chapel of the Palazzo Medici-Riccardi, and by Domenico Ghirlandajo in the church of Sta. Maria Novella. Buon' fresco painting is unsuited to the damp, cold climate of northern countries, and there is now some concern for the preservation of frescoes in the sulfurous atmosphere of even many southern cities. In recent times, buon' fresco has been successfully revived by the Mexican mural painters Diego Rivera, José Orozco, and Rufino Tamayo.

Sgraffito. Sgraffito (Italian *graffiare*, "to scratch") is a form of fresco painting for exterior walls. A rough plaster undercoat is followed by thin plaster layers, each stained with a different lime-fast colour. These coats are covered by a fine-grain mortar finishing surface, about two inches thick (five centimetres). The plaster is then engraved with knives and gouges at different levels to reveal the various coloured layers beneath. The sintered-lime process binds the colours. The surface of modern sgraffito frescoes is often enriched with textures made by impressing nails and machine parts, combined with mosaics of stone, glass, plastic, and metal tesserae.

Sgraffito has been a traditional folk art in Europe since the Middle Ages and was practiced as a fine art in 13th-century Germany. It has been recently revived in northern Europe.

Oil. Oil paints are made by mixing dry pigment powder with refined linseed oil to a paste, which is then milled in order to disperse the pigment particles throughout the oil vehicle. According to the 1st-century Roman scholar Pliny the Elder, whose writings the Flemish painters Hubert and Jan van Eyck are thought to have studied, the Romans used oil colours for shield painting. The earliest use of oil as a fine-art medium is generally attributed to 15th-century European painters, such as Giovanni Bellini and the van Eycks, who glazed oil colour over a tempera underpainting. It is also thought probable, however, that medieval manuscript illuminators had been using oil glazes in order to achieve greater depth of colour and more subtle tonal transitions than their tempera medium allowed.

Earliest use as a fine-art medium

Oils have been used on linen, burlap, cotton, wood, hide, rock, stone, concrete, paper, cardboard, aluminum, copper, plywood, and processed boards, such as masonite, pressed wood, and hardboard. The surface of rigid panels is traditionally prepared with gesso and that of canvas with one or more coats of white acrylic resin emulsion or with a coat of animal glue followed by thin layers of white-lead oil primer. Oil paints can be applied undiluted to these prepared surfaces or can be used thinned with pure gum turpentine or its substitute, white mineral spirit. The colours are slow drying; the safest dryer to speed the process is cobalt siccative.

An oil glaze is a transparent wash of pigment, traditionally thinned with stand oil (a concentrate of linseed oil). Glazes can be used to create deep, glowing shadows and to bring contrasted colours into closer harmony beneath a unifying tinted film. Scumbling is the technique of scrubbing an undiluted, opaque, and generally pale pigment across others for special textural effects or to raise the key of a dark-coloured area.

Hog-bristle brushes are used for much of the painting, with pointed, red sable-hair brushes generally preferred for outlines and fine details. Oils, however, are the most plastic and responsive of all painting mediums and can be handled with all manner of tools. The later works of Titian and Rembrandt, for example, appear to have been executed with thumbs, fingers, rags, spatulas, and brush handles. With these and other unconventional tools and techniques, oil painters create pigment textures ranging from delicate tonal modulations to unvarying, mechanical finishes and from clotted, impasto ridges of paint to barely perceptible stains.

The tempera-underpainting–oil-glaze technique was practiced into the 17th century. Artists such as Titian, El Greco, Rubens, and Diego Velázquez, however, used oil pigments alone and, employing a method similar to pastel painting, applied them directly to the brownish ground with which they had tinted the white priming. Contours and shadows were stained in streaks and washes of diluted paint, while lighter areas were created with dry, opaque scumbles, the tinted ground meanwhile providing the halftones and often remaining untouched for passages of local or reflected colour in the completed picture. This use of oil paint was particularly suited to expressing atmospheric effects and to creating chiaroscuro, or light and dark, patterns. It also encouraged a bravura handling of paint, where stabs, flourishes, lifts, and pressures of the brush economically described the most subtle changes of form, texture, and colour according to the influence exerted by the tinted ground through the varying thicknesses of overlaid pigment. This method was still practiced by the 19th-century painters such as John Constable, J.M.W. Turner, Eugène Delacroix, and Honoré Daumier. The Impressionists, however, found the luminosity of a brilliant white ground essential to the alla prima technique with which they represented the colour intensities and shifting lights of their plein air (open air) subjects. Most oil paintings since then have been executed on white surfaces.

Deterioration of Leonardo's "Last Supper"

The rapid deterioration of Leonardo's 15th-century "Last Supper," which was painted in oils on plaster, probably deterred artists from using the medium directly on a wall surface. And the likelihood of eventual warping also prohibited using the large number of braced wood panels required to make an alternative support for an extensive mural painting in oils. Canvas, however, can be woven to any length, and, since an oil-painted surface is elastic, mural paintings could be executed in the studio and rolled and restretched on a wooden framework at the site or marouflaged (fastened with an adhesive) directly onto the wall surface. In addition to the immense studio canvases painted for particular sites by artists such as Jacopo Tintoretto, Paolo Veronese, Delacroix, Pierre-Cécile Puvis de Chavannes, and Monet, the use of canvas has made it possible for mural-size, modern oil paintings to be transported for exhibition to all parts of the world.

Tempera

"Wind from the Sea," tempera painting by Andrew Wyeth, 1947. In a private collection. 47 × 70 cm.

"Christ Discovered in the Temple," panel painting by Simone Martini, 1342. In the Walker Art Gallery, Liverpool. 49.5 × 35 cm.

"King Khusrau Seated Upon his Throne," Persian miniature from the *Khamsa of Nizami*, 1524–25. In the Metropolitan Museum of Art, New York City. 32 × 23 cm.

A leaf from the *Calendar of a Book of Hours (April)*, manuscript illumination by Simon Bennick, early 16th century. In the Victoria and Albert Museum, London. 14 × 9.6 cm.

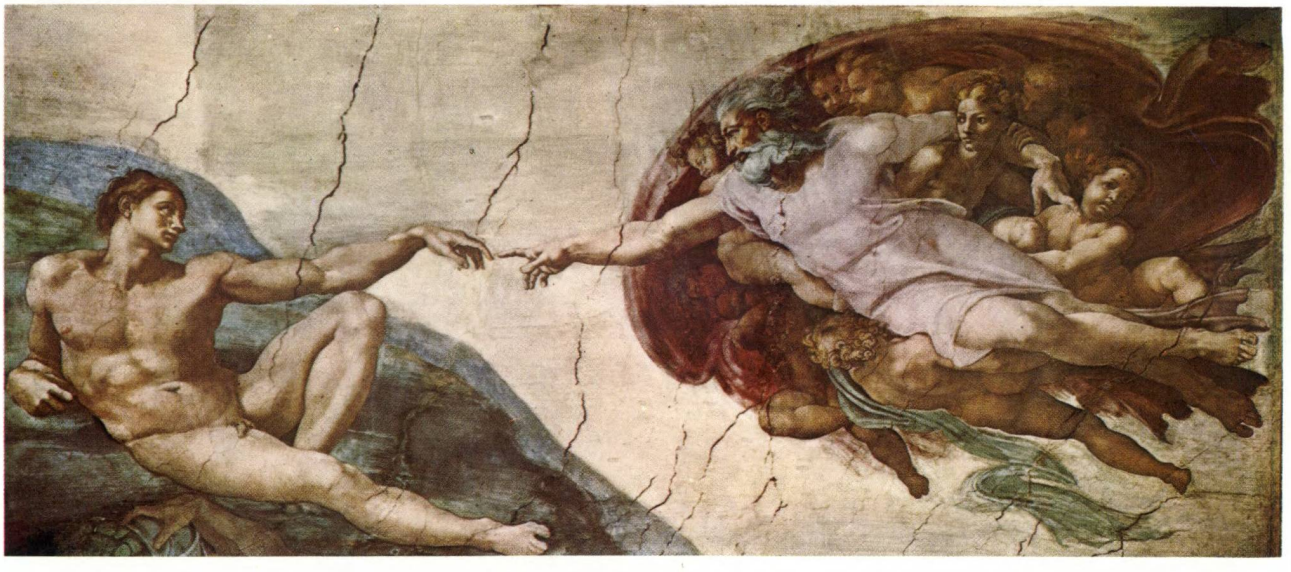
Plate 1: By courtesy of (top right) the Walker Art Gallery, Liverpool, (bottom left) the Metropolitan Museum of Art, New York, gift of Alexander Smith Cochran, 1913, (bottom right) the Victoria and Albert Museum, London; photograph, (bottom right) John Webb

Fresco of a court scene from the *Mahājanaka Jataka*, Cave I, Ajantā, India, 600–700.

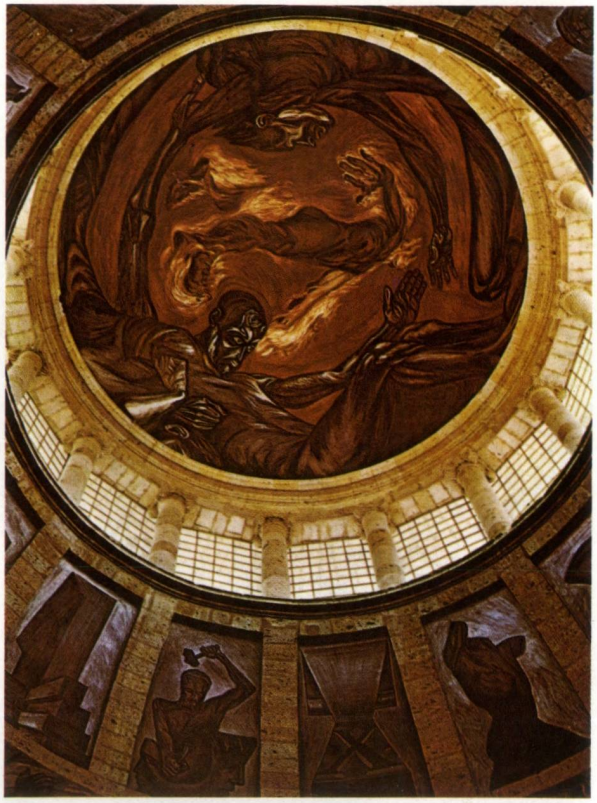
Frescoes by Giovanni Battista Tiepolo, 1750–52, decorating the Kaisersaal Residenz, Würzburg, West Germany, designed by Balthasar Neumann, 1719–44.

"Fowling in the Marshes," detail of a wall painting from the tomb of Amenemhab, Thebes, Egypt, *c.* 1450 BC, XVIII dynasty. In the British Museum.

"The Creation of Adam," detail of the ceiling fresco in the Sistine Chapel, Vatican, by Michelangelo, 1508–12.



Wall painting



Frescoes by José Clemente Orozco, 1938–39, decorating the ceiling and drum of the dome of the Hospicio Cabañas, Guadalajara, Mexico.

Fresco cycle depicting scenes from the life of the Virgin and the life of Christ by Giotto in the Arena Chapel, Padua, Italy, *c.* 1305–09.

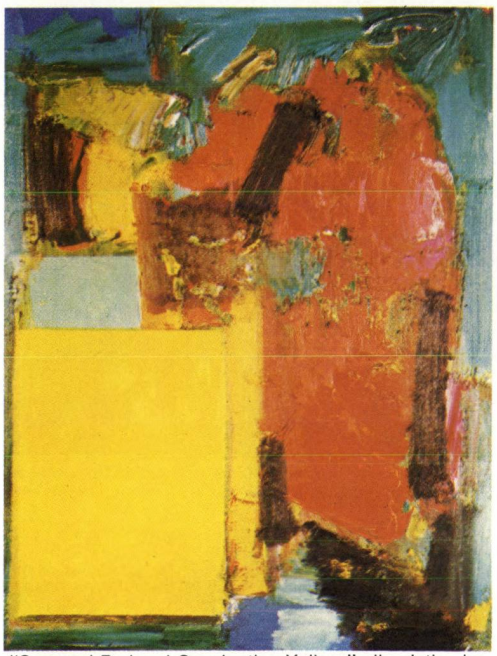
Folk art fresco on the Adam and Eve house, Ardez, Switzerland, 1647.

Plate 2: By courtesy of (centre left) the trustees of the British Museum; photographs, (top left) V. Panjabi—Shostal, (top right) A.F. Kersting, (bottom) SCALA, N.Y.

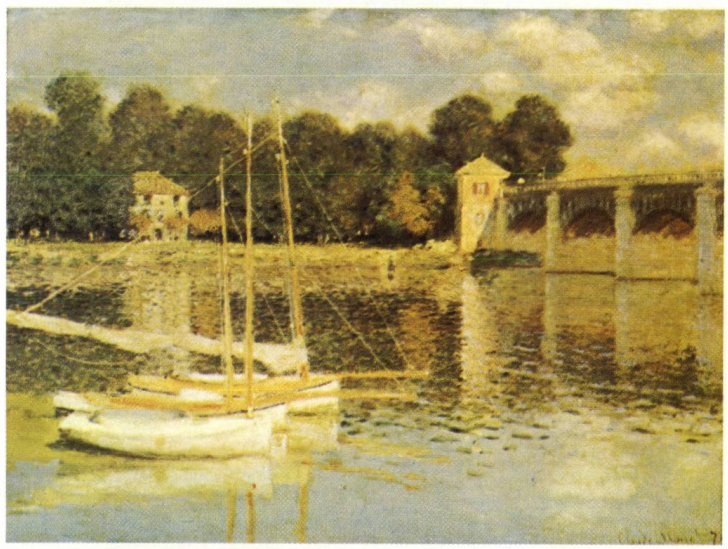
Plate 3: By courtesy of (top right) Fondo Editorial de la Plastica Mexicana; photographs, (centre left) SCALA, N.Y. (bottom right) Toni Schneiders

Oil and encaustic

"Salisbury Cathedral," oil painting by John Constable, *c.* 1825. In the National Gallery, London. 53 × 77 cm.



"Smaragd Red and Germinating Yellow," oil painting by Hans Hofmann, 1959. In the Cleveland Museum of Art. 1.4 × 1 m.



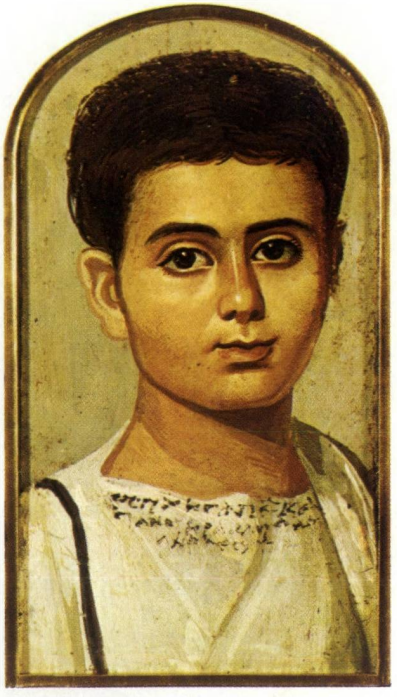
"The Bridge at Argenteuil," oil painting by Claude Monet, 1874. In the Louvre, Paris. 60 × 80 cm.

"Venus and the Lute Player," oil painting by Titian, *c.* 1560. In the Metropolitan Museum of Art, New York City. 1.7 × 2.1 m.

Plate 4: By courtesy of (top left) the Contemporary Collection of the Cleveland Museum of Art, (top right) the trustees of the National Gallery, London, (bottom) the Metropolitan Museum of Art, New York, Munsey Fund, 1936; photograph, (centre right) Cliche Musees Nationaux

Plate 5: By courtesy of (top left) the Rijksmuseum, Amsterdam, (top right) the Metropolitan Museum of Art, New York, gift of Edward S. Harkness, 1917–18, (bottom left) the National Gallery of Art, Washington, D.C., Andrew W. Mellon Collection, 1937; photograph, (bottom right) Joseph Martin—SCALA, N.Y.

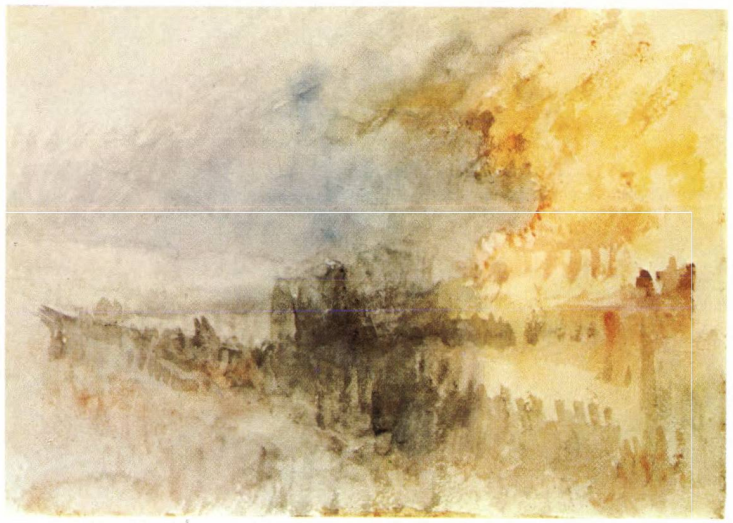
"The Bridal Couple," oil painting by Rembrandt, *c.* 1665. In the Rijksmuseum, Amsterdam. 1.2 × 1.7 m.



Portrait of a boy, encaustic painting from al-Fayyūm, Egypt, probably 2nd century. In the Metropolitan Museum of Art, New York City. 33 × 18.5 cm.

"The Annunciation," oil painting by Jan van Eyck, *c.* 1434. In the National Gallery of Art, Washington, D.C. 93 × 36.5 cm.

"Mystic Marriage of St. Catherine," oil painting by Peter Paul Rubens, 1627–28. In St. Augustine Church, Antwerp. 5.7 × 4 m.



"The Burning of the Houses of Parliament," watercolour painting by J.M.W. Turner, 1834. In the British Museum. 23.4 × 32.4 cm.

"A Young Man Among Roses," watercolour miniature by Nicholas Hilliard, *c.* 1588. In the Victoria and Albert Museum, London. 13.4 × 7.4 cm.

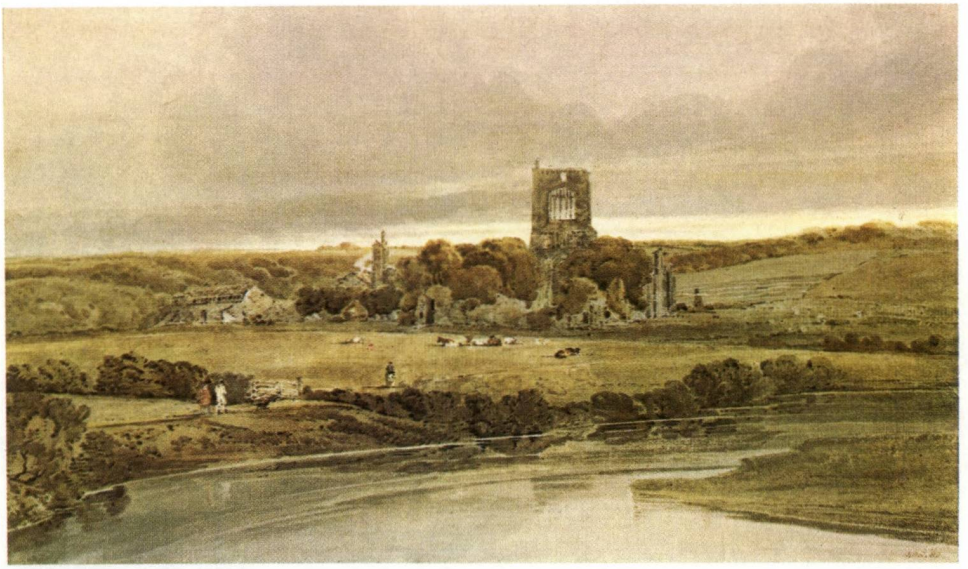
Ink, gouache, and watercolour

"The Annunciation," anonymous Austrian or Swabian woodcut, hand-coloured with watercolour, 1450–70. In the National Gallery of Art, Washington, D.C. 27.3 × 19.2 cm.

"The Monumental Turf," watercolour painting by Albrecht Dürer, 1503. In the Albertina, Vienna. 41 × 31.5 cm.

Plate 6: By courtesy of (top left) the Victoria and Albert Museum, London, (bottom left) the National Gallery of Art, Washington, D.C., Rosenwald Collection, (bottom right) the Albertina, Vienna; photograph, (top right) Hans Hinz, Basel

Plate 7: By courtesy of (top, centre right) the Victoria and A'bert Museum, London, (bottom) the Museum of Fine Arts, Boston, Ross Collection; photographs, (top, centre right) John Webb, (centre left) Holle Bildarchiv, Baden-Baden

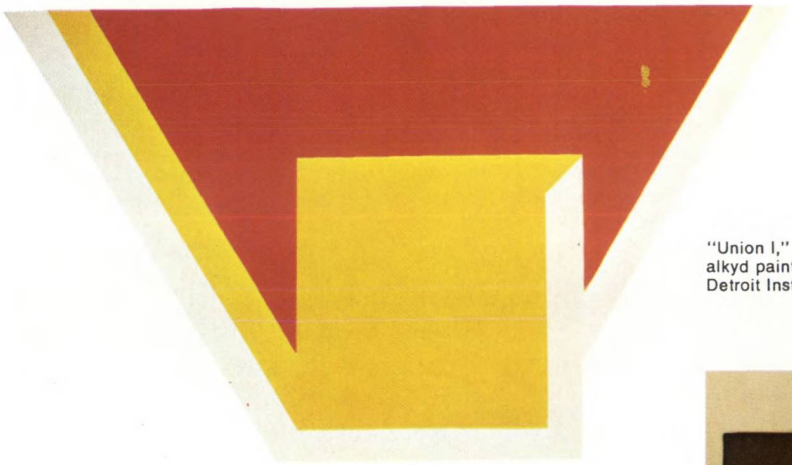


"Kirkstall Abbey, Yorkshire—Evening," watercolour painting by Thomas Girtin, 1801. In the Victoria and Albert Museum, London. 30.4 × 51 cm.

"Landscape," detail of a Japanese handscroll, ink on paper by Sesshū (1420–1506). In the National Museum, Tokyo. Dimensions of entire scroll 147.3 × 35.6 cm.

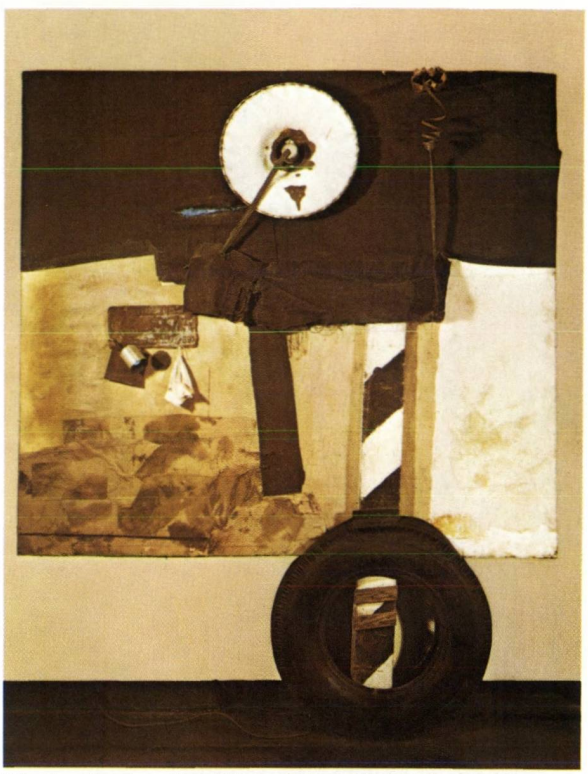
French fan, gouache on paper, sticks of carved mother-of-pearl inlaid with gilt and set with pastes, *c.* 1750. In the Victoria and Albert Museum, London.

"The Return of Wen-chi to Ying-chuan," scene from the Chinese handscroll "The Life of Lady Ts'ai-chi," anonymous, ink and colour on silk, 12th century, Sung dynasty. In the Museum of Fine Arts, Boston. Detail of the scroll 25 × 55.8 cm.

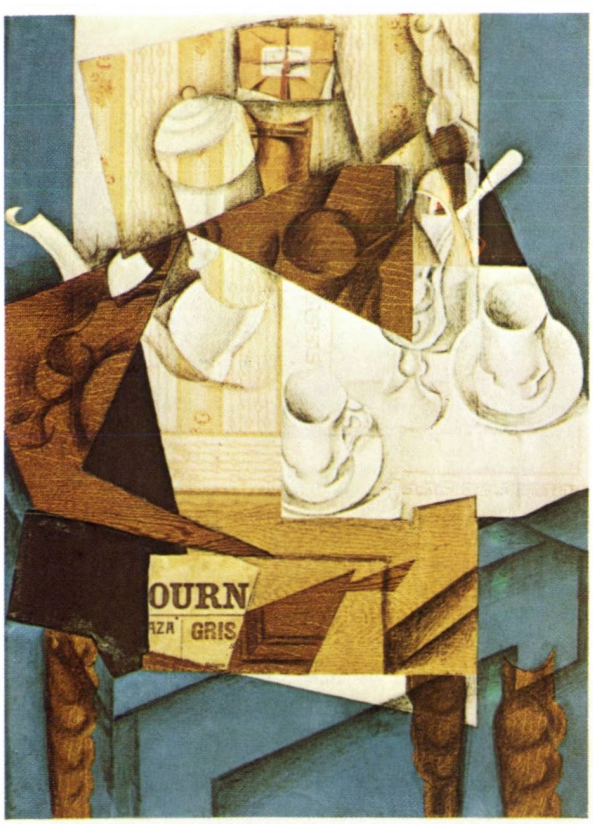


"Union I," epoxy paint and fluorescent alkyd painting by Frank Stella, 1966. In the Detroit Institute of Arts. 2.6 × 4.5 m.

Synthetic and mixed media



"First Landing Jump," combine painting by Robert Rauschenberg, 1961. In the Museum of Modern Art, New York City. 2.3 × 1.8 m.



"Breakfast," pasted paper, crayon, and oil paint collage by Juan Gris, 1914. In the Museum of Modern Art, New York City. 81 × 60 cm.

"Oh, Jeff . . . I Love You, Too . . . But . . .," acrylic painting by Roy Lichtenstein, 1964. In the Harry N. Abrams Family Collection, New York City. 1.2 × 1.2 m.

Plate 8: By courtesy of (top left) the Detroit Institute of Arts, gift of Friends of Modern Art, (top right) the Leo Castelli Gallery, New York, (bottom left) the Museum of Modern Art, New York, acquired through the Lillie P. Bliss Bequest, (bottom right) the Harry N. Abrams Family Collection, New York City

The tractable nature of the oil medium has sometimes encouraged slipshod craftsmanship. Working over partly dry pigment or priming may produce a wrinkled surface. The excessive use of oil as a vehicle causes colours to yellow and darken, while cracking, blooming, powdering, and flaking can result from poor priming, overthinning with turpentine, or the use of varnish dryers and other spirits. Colour changes may also occur through the use of chemically incompatible pigment mixtures or from the fading of fugitive synthetic hues, such as the crimson lakes used by Pierre-Auguste Renoir.

Watercolour. Watercolours are pigments ground with gum arabic and gall and thinned with water in use. Sable and squirrel ("camel") hair brushes are used on white or tinted paper.

Three hundred years before the late 18th-century English watercolorists, Albrecht Dürer had anticipated their technique of transparent colour washes in a remarkable series of plant studies and panoramic landscapes. Until the emergence of the English school, however, watercolour became a medium merely for colour tinting outlined drawings or, combined with opaque body colour to produce effects similar to gouache (see below) or tempera, was used in preparatory studies for oil paintings.

The chief exponents of the English method were Thomas Girtin, John Sell Cotman, John Robert Cozens, Richard Parkes Bonington, David Cox, and Constable. Their contemporary J.M.W. Turner, however, true to his unorthodox genius, added white to his watercolour and used rags, sponges, and knives to obtain unique effects of light and texture. Victorian watercolorists, such as Birket Foster, used a laborious method of colour washing a monochrome underpainting, similar in principle to the tempera–oil technique. Following the direct, vigorous watercolours of the French Impressionists and Postimpressionists, however, the medium has been established in Europe and America as an expressive picture medium in its own right. Notable 20th-century watercolorists have been Wassily Kandinsky, Klee, Dufy, and Georges Rouault; the U.S. artists Thomas Eakins, Maurice Prendergast, Charles Burchfield, John Marin, Lyonel Feininger, and Jim Dine; and the English painters John and Paul Nash, Eric Ravilious, Edward Bawden, Edward Burra, and Patrick Procktor.

The "pure" watercolour technique

In the "pure" watercolour technique, often referred to as the English method, no white or other opaque pigment is applied, colour intensity and tonal depth being built up by successive, transparent washes on damp paper. Patches of white paper are left unpainted to represent white objects and to create effects of reflected light. These flecks of bare paper produce the sparkle characteristic of pure watercolour. Tonal gradations and soft, atmospheric qualities are rendered by staining the paper when it is very wet with varying proportions of pigment. Sharp accents, lines, and coarse textures are introduced when the paper has dried. The paper should be of the type sold as "handmade from rags"; this is generally thick and grained. Cockling is avoided when the surface dries out if the dampened paper has been first stretched across a special frame or held in position during painting by an edging of adhesive tape.

Ink. Ink is the traditional painting medium of China and Japan, where it has been used with long-haired brushes of wolf, goat, or badger on silk or absorbent paper. Oriental black ink is a gum-bound carbon stick that is ground on rough stone and mixed with varying amounts of water to create a wide range of modulated tones or applied almost dry, with lightly brushed strokes, to produce coarser textures. The calligraphic brush technique is expressive of Zen Buddhist and Confucian philosophies, brushstroke formulas for the spiritual interpretation of nature in painting dictating the use of the lifted brush tip for the "bone," or "lean," structure of things and the spreading belly of the hairs for their "flesh," or "fat," volumes. The Far Eastern artist poises the brush vertically above the paper and controls its rhythmic movements from the shoulder. Distant forms represented in landscapes painted on silk were sometimes brushed on from the reverse side in order to create a mysterious illusion of depth.

In the Western world, ink has been used rather more for preparatory studies and topographical and literary illustrations than as a medium for easel paintings. Western artists have generally combined ink washes with contours and textures in quill or steel pen. Among the finest of these are by Rembrandt, Nicolas Poussin, Tiepolo, Samuel Palmer, Constable, and Edouard Manet. Claude Lorrain, Turner, and Daumier and, in the 20th century, Braque, Picasso, Reginald Marsh, Henri Michaux, and John Piper are some of those who have exploited its unique qualities.

Gouache. Gouache is opaque watercolour, known also as poster paint and designer's colour. It is thinned with water for applying, with sable- and hog-hair brushes, to white or tinted paper and card and, occasionally, to silk. Honey or starch is sometimes added to retard its quick-drying property. Liquid glue is preferred as a thinner by painters wishing to retain the tonality of colours (which otherwise dry slightly lighter in key) and to prevent thick paint from flaking. Gouache paints have the advantages that they dry out almost immediately to a mat finish and, if required, without visible brushmarks. These qualities, with the capacities to be washed thinly or applied in thick impasto and a wide colour range that now includes fluorescent and metallic pigments, make the medium particularly suited to preparatory studies for oil and acrylic paintings. It is the medium that produces the suede finish and crisp lines characteristic of many Indian and Islāmic miniatures, and it had been used in Western screen and fan decoration and by modern artists such as Rouault, Klee, Dubuffet, and Morris Graves.

Medium particularly suited to preparatory studies

Encaustic. Encaustic painting (from the Greek: "burnt in") was the ancient method, recorded by Pliny, of fixing pigments with heated wax. It was probably first practiced in Egypt about 3000 BC and is thought to have reached its peak in Classical Greece, although no examples from that period survive. Pigments, mixed with melted beeswax, were brushed onto stone or plaster, smoothed with a metal spatula, and then blended and driven into the wall with a heated iron. The surface was later polished with a cloth. Leonardo and others attempted unsuccessfully to revive the technique. North American Indians used an encaustic method, whereby pigments mixed with hot animal fat were pressed into a design engraved on smoothed buffalo hide.

A simplified encaustic technique uses a spatula to apply wax mixed with solvent and pigment to wood or canvas, producing a ridged, impasto surface. This is an ancient and most durable medium, Coptic mummy portraits from the 1st and 2nd centuries AD retaining the softly blended, translucent colouring typical of waxwork effigies. In the 19th century, Vincent van Gogh also used this method to give body to his oil pigment; the Neo-Impressionist artist Louis Hayet applied encaustic to paper, and it was used by the U.S. Pop Art painter Jasper Johns for his "Flag" canvases. Coloured wax crayons, also, have been used by modern painters such as Picasso, Klee, Arshile Gorky, and David Hockney (see above *Mediums: Fresco*).

Casein. Casein, or "cheese painting," is a medium in which pigments are tempered with the gluey curd of cheese or milk precipitate. For handling, an emulsion of casein and lime is thinned with water. The active element of casein contains nitrogen, which forms a soluble caseate of calcium in the presence of lime. It is applied in thin washes to rigid surfaces, such as cardboard, wood, and plastered walls.

Casein colours dry quickly, although lighter in tone than when first applied. Since they have more body than egg-tempera paints, they can be applied with bristle brushes to create impasto textures not unlike those of oils. Casein paints were used in ancient Rome. They are now available ready-made in tubes and have been used by such modern artists as Robert Motherwell and Claes Oldenburg.

Casein is also an ingredient of some charcoal and pastel fixatives and was a traditional primer for walls and panels.

Synthetic mediums. Synthetic mediums, developed by industrial research, range from the Liquitex fabric dyes

used on canvas by the U.S. abstract painter Larry Poons to the house enamel paints employed at times by Picasso and Jackson Pollock.

Acrylic resin emulsion

The most popular medium and the first to challenge the supremacy of oils is acrylic resin emulsion, since this plastic paint combines most of the expressive capabilities of oils with the quick-drying properties of tempera and gouache. It is made by mixing pigments with a synthetic resin and thinning with water. It can be applied to any sufficiently toothed surface with brush, roller, spray gun, spatula, sponge, or rag. Acrylic paints dry quickly, without brush marks, to form a mat, waterproof film that is also elastic, durable, and easily cleaned. They show little colour change in drying, nor do they darken in time. While they lack the surface textural richness of oil or encaustic, they can be built up with a spatula into opaque impastos or thinned immediately into transparent colour glazes. Polyvinyl acetate (PVA) is applied for priming, although it is claimed that acrylic paints can be safely applied directly onto unprepared raw canvas or cotton. The wide range of intense hues is extended by fluorescent and metallic pigments. Polymer paints are particularly suitable for the precise, immaculate finish demanded by Op Art and Minimal painters, such as Bridget Riley, Morris Louis, Kenneth Noland, and Frank Stella.

Other mediums. *French pastels.* French pastels, with the sharpened lumps of pigment used by Ice Age artists, are the purest and most direct painting materials. Pastel pigments are mixed only with sufficient gum to bind them for drying into stick molds. Generally, they are used on raw strawboard or on coarse-grained tinted paper, although vellum, wood, and canvas have been also employed. These colours will not fade or darken, but, since they are not absorbed by the surface of the support, they lie as pigment powder and are easily smudged. Unfortunately, pastel colours lose their luminosity and tonality if fixed with a varnish and so are best preserved in deep mounts behind glass. Degas often overcame the fragile nature of true pastel painting by the unorthodox method of working on turpentine-soaked paper, which absorbed the powdery pigment.

Eighteenth-century portrait pastellists, such as Maurice-Quentin de La Tour, Jean-Baptiste Peronneau. Jean-Étienne Liotard, and Anton Raphael Mengs, blended the pigment with coiled paper stumps into the surface appearance of a smooth oil painting. Later pastel painters such as Degas, Toulouse-Lautrec, Mary Cassatt, Everett Shinn, Klee, and Arthur Dove, contrasted broad masses of granular colour, spread with the side of the stick, with broken contours and passages of loose cross-hatching. They often used the tinted ground as a half-tone, and, according to the amount of manual pressure exerted on the chalk, they varied the degree of pigment opacity to extract a wide range of tints and shades from each pastel colour.

Oil pastels. Oil pastels are pigments ground in mastic with oil of turpentine, spermaceti, and poppy oil. They are used in a similar way to that of French pastels but are already fixed and harder, producing a permanent, waxy finish. Oil-pastel paintings are generally executed on white paper, card, or canvas. The colours can be blended if the surface of the support is dampened with turpentine or if they are overworked with a brush and turpentine. They are popular for small preparatory studies for oil and acrylic paintings.

Folk art tradition

Glass paintings. Glass paintings are executed with oil and hard resin or with watercolour and gum on glass sheets. These have been a folk art tradition in Europe and North America and, from the 15th to the 18th centuries, were regarded as a fine art in northern Europe, where they have been more recently revived by such painters as Willi Dirx, Ida Kerkovius, Lily Hildebrandt, Klee, Oskar Schlemmer, and Heinrich Camperdonck. Colours are applied from the back in reverse order. Unpainted areas of glass are often coated with mercury, providing a mirror background to the coloured images; this creates the kind of illusionary, bizarre spatial relationship between the viewer and picture space sought by the modern artist Michelangelo Pistoletto with his use of photographic images fixed to a polished steel sheet. The colours seen through glass appear translucent, jewel-like, and, since they cannot be touched, even magical.

Ivory painting. Ivory painting was practiced in the 18th and 19th centuries in Europe and America for portrait miniatures. These were generally oval shaped and designed as keepsakes, lockets, and mantle pictures. They were painted under a magnifying glass in fairly dry watercolour or tempera stippling, with sable- or marten-hair brushes on thin, semitranslucent ivory pieces. Corrections were made with a needle. The velvet quality of their colours was enhanced, on the thinner ivories, by the glow produced by a gold leaf or tinted backing (see below *Forms of painting: Miniature painting*).

Lacquer. Lacquer has been a traditional Chinese medium for over 2,000 years. It combines painting with intaglio relief. Linen-covered wood panels are coated with chalk or clay, followed by many thin layers of black or red lacquer-tree resin. The surface is polished and a design engraved, which is then coloured and gilded or inset with mother-of-pearl. Layers of compressed paper or molded papier-mâché have also provided supports. In China and Japan, lacquer has been used principally for decorating shrine panels, screens, caskets, panniers (large baskets), and musical instruments.

Sand, or dry, painting. Sand, or dry, painting is a traditional magic art of the North American Indians; it is still practiced in healing ceremonies among the Navajos of New Mexico and Arizona. Ground sandstone, natural ochres, mineral earths, and powdered charcoal are sprinkled onto a pattern marked into an area covered with yellow-white sand. The patient sits in the centre of this vivid symbolic design of coloured figurative and geometrical shapes. Following the ritual, the painting is destroyed. These "floor" pictures influenced Jackson Pollock in his horizontally spread "action paintings."

Paper. From the end of the 18th century profiles and full-length group portraits were cut in black paper, mounted on white card, and often highlighted in gold or white. A silhouette ("shade") might be first outlined from the sitter's cast shadow with the aid of a physionotrace.

Collage. Collage was the Dada and Synthetic Cubist technique of combining labels, tickets, newspaper cuttings, wallpaper scraps, and other "found" surfaces with painted textures simulating wood graining and marbling. Frottage was Max Ernst's method of taking paper rubbings from surfaces, unrelated to one another in real life, and combining them to create fantasy landscapes. Cut paper shapes, hand coloured in gouache, were used by Matisse for his monumental last paintings; Piet Mondrian composed his famous "Victory Boogie Woogie" (1942–43) in coloured-paper cutouts.

Mechanical mediums. The use of mechanical mediums in painting has run parallel to similar developments in modern music and drama. In the field of cybernetics, painters have programmed computers to permutate drawings, photographs, diagrams, and symbols through sequences of progressive distortion; and light patterns are produced on television screens by deliberate magnetic interference and by sound-wave oscillations. Artists have also explored the expressive and aesthetic possibilities of linear holograms, in which all sides of an object can be shown by superimposed light images. Painters are among those who have extended the boundaries of film making as an art form. Following the Surrealist film fantasies created by Berthold Bartosch, Jean Cocteau, Hans Richter, and Salvador Dalí, by Schlemmer's filmed ballets and Norman McLaren's hand-painted abstract animations, some artists are now experimenting with video cassettes for television, intending to distribute them in the manner of limited or multiple editions of prints.

For some Conceptual artists the typewriter is the only equipment used when visual ideas are expressed in the form of instruction sheets alone. For example, typed proposals for defining the real space of an exhibition area with painted lines might invite the active participation of visitors (see below *Forms of painting: Modern forms*).

Mixed mediums. Some pictures are first painted in one medium and corrected or enriched with colour and tex-

ture in another. Examples of this kind of mixed mediums are the Renaissance tempera–oil technique, William Blake's relief etchings colour-printed in glue tempera and hand-finished in watercolour, and Degas's overpainted monotypes and his combinations of pastel, gouache, and oil. More recent examples are Richard Hamilton's photographs overpainted in oil colour, Dubuffet's patchwork assemblages of painted canvas and paper, and Klee's alchemy in mixing ingredients such as oil and distemper on chalk over jute and watercolour and wax on muslin stuck on wood.

FORMS OF PAINTING

Mural painting. Mural painting has its roots in the primeval instincts of people to decorate their surroundings and to use wall surfaces as a form for expressing ideas, emotions, and beliefs. In their universal manifestation in graffiti and in ancient murals, such as Ice Age cave paintings and protodynastic Egyptian frescoes, symbols and representational images have been spread freely and indiscriminately across walls, ceilings, and floors. But, in more disciplined attempts to symbolize the importance and function of particular buildings through their interior decoration, murals have been designed for the restricted framework of specific surface areas. They therefore have to be painted in close relationship to the scale, style, and mood of the interior and with regard to such siting considerations as light sources, eye levels, the spectators' lines of sight and means of approach, and the emotive scale relationship between spectators and the painted images.

Early mural decorations

Early mural decorations for tombs, temples, sanctuaries, and catacombs were generally designed in horizontal divisions and vertical axes. These grid patterns were in harmony with the austere character of the interiors, and their geometrical plan enabled the artist to depict clearly the various episodes and symbols of a narrative subject. In these early traditions of mural design, in China, India, Mexico, Egypt, Crete, and Byzantium, no illusionary devices were used to deny the true flatness of the wall surface; images were silhouetted against a flatly painted ground framed by decorative dadoes (the decoration adorning the lower part of an interior wall) of stylized motifs in repeat patterns. By the early Renaissance, however, innovators such as Giotto, Masaccio, and Fra Angelico were placing figures within architectural and landscape settings, painted as if extensions to the real dimensions of the interior. The peak of technical skill and artistic expression was reached in the 15th and 16th centuries with the frescoes of Piero della Francesca, Michelangelo, and Raphael. The irregular shapes of wall areas and the distortions produced by convex surfaces were inventively exploited in the design. Intruding doors and windows, for example, were skillfully circumvented by sweeping pattern rhythms or were incorporated as features in the painting, and figures were foreshortened so as to appear to float across or to rise into cupolas (rounded vaults that form ceilings), lunettes (rounded spaces over doors or windows), and apses (domed projections of a church, usually at the east end or altar), the curving surfaces of which might be painted to simulate celestial skies. Existing structural wall features provided the divisions between narrative episodes. These were often supplemented by trompe l'oeil ("deceive the eye") columns, pilasters, arcading, balustrading, steps, and other architectural forms that also served to fuse the painted setting with the real interior.

With the increasing dependence upon tapestry hangings and stained glass as primary forms of interior decoration, mural painting suffered a decline in the Western world. Except for those given to Rubens, Tiepolo, Delacroix, and Puvis de Chavannes, there were relatively few important mural commissions in the period following the High Renaissance. In the 20th century, however, enlightened patronage has occasionally enabled leading modern artists to execute paintings for specific sites: Monet's "Water-lilies" series for the Paris Orangerie, for example, and other murals in France by Vuillard, Matisse, Léger, Chagall, and Picasso; in Mexico and the United States by Orozco, Rivera, Tamayo, and David Siqueiros and also in the U.S. by Matisse, Shahn, and Willem de Kooning; in Britain by Sir Stanley Spencer and Bawden; in Norway by Edvard Munch; in Holland by Karel Appel; and in Italy by Afro Basaldella.

Easel and panel painting. The easel, or studio, picture was a form developed during the Renaissance with the establishment of the painter as an individual artist. Its scale and portability enabled European artists to extend the range of themes, previously restricted to those suitable to mural decoration. Easel and panel forms include still life, portraiture, landscape, and genre subjects and permit the representation of ephemeral effects of light and atmosphere that the more intimate forms of Oriental art had already allowed the painters of scrolls, screens, and fans to express. Although easel paintings are occasionally commissioned for a special purpose, they are generally bought as independent art objects and used as decorative focal features or illusionary window views in private homes. They are also collected as financial investments, for social prestige, for the therapeutic escapism their subject may provide, or purely for the aesthetic pleasure they afford.

Panel paintings

Panel paintings, by strict definition, are small pictures designed for specific sacred or secular purposes or as part of a functional object. Although these wooden boards are sometimes categorized as a form of "decorative" rather than "fine" art, the best examples justify their place in museums alongside great easel paintings. Among the functions they originally served were as predellas (the facings to altar-step risers); devotional and ceremonial icons; portable, folding diptych and triptych altarpieces; shop and tavern signboards; mummy cases; and for the panels of carriages, musical instruments, and cassoni. Many of them were painted by acknowledged masters, such as Fra Angelico, Paolo Uccello, and Antoine Watteau, as well as by anonymous folk artists.

Miniature painting. Miniature painting is a term applied both to Western portrait miniatures and to the Indian and Islāmic forms of manuscript painting discussed below. Portrait miniatures, or limnings, were originally painted in watercolour with body colour on vellum and card. They were often worn in jewelled, enamelled lockets. Sixteenth-century miniaturists, such as Hans Holbein, Jean Clouet, Nicholas Hilliard, and Isaac Oliver, painted them in the tradition of medieval illuminators. Their flat designs, richly textured and minutely detailed, often incorporated allegorical and gilded heraldic motifs. In 17th- and 18th-century Western portrait miniatures, the two-dimensional pattern of rich colours was developed by atmospheric tonal modelling into more naturalistic representations; these were sometimes in pastel and pencil or painted in oils on a metal base. Pantographs (reducing and enlarging copying instruments made on the lazy-tongs lever principle) might be used to transfer a drawing. Among the exponents of this naturalistic style were Francisco Goya, Fragonard, Samuel Cooper, John Hoskins, François Dumont, and the U.S. miniaturists Robert Field and Edward Green Malbone. The introduction of painted ivory miniatures was followed, in the 19th century, by a decline in aesthetic standards, although a classical simplicity was achieved by unsophisticated itinerant limners and by the German miniaturist Patricius Kittner. The painted miniature was eventually superseded by the small, hand-tinted photograph (see above *Mediums: Ivory painting*).

Manuscript illumination and related forms. Among the earliest surviving forms of manuscript painting are the papyrus rolls of the ancient Egyptian Book of the Dead, the scrolls of Classical Greece and Rome, Aztec pictorial maps, and Mayan and Chinese codices, or manuscript books. European illuminated manuscripts were painted in egg-white tempera on vellum and card. Their subjects included religious, historical, mythological, and allegorical narratives, medical treatises, psalters, and calendars depicting seasonal occupations. In contrast to the formalized imagery of Byzantine and early Gothic manuscript painters, Celtic illuminators developed a unique, abstract style of elaborate decoration, the written

Abstract style of Celtic illuminators

text being overwhelmed by intricate latticework borders, with full-page initial letters embraced by interlacing scrolls. The medieval Gothic style of illumination, in sinuous, linear patterns of flattened forms isolated against white or gilded grounds, had developed, by the end of the 15th century, into exquisitely detailed, jewel-like miniatures of shaded figures and spatial landscapes. These were often framed by gilded initial letters as vignettes or by margin borders in simulated half relief. With the advent of printing in the 15th century and a final, brilliant period of Flemish and Italian illumination, European manuscript painting survived only in official documents, maps, and in the form of hand-coloured, block-printed pages. Pennsylvanian-German birth and baptismal certificates in the U.S. and William Blake's hand-coloured engravings to the Bible and to his own poems were isolated revivals of those forms.

Indian and Islāmic miniature painting, however, was practiced into the 19th century; and 11th-century Oriental albums of poem paintings in ink, on leaves of silk or paper, represent a tradition that was continued into modern times. The subjects of Near Eastern miniatures included religious and historical narrative, cosmic maps, and medical, palmistry and astrological charts, as well as illustrations to poems, songs, and romantic epics. These were generally painted in gouache on paper, with occasional gold- or silver-leaf embellishment. The linear design was first drawn with a brush in delicate contours and soft shading. Landscape and architectural detail was as well observed as in that of the principal figures.

The rapprochement established between text, painted borders, margin spaces, and illustration is characteristic of both Eastern and Western manuscript paintings. In Indian and Islāmic miniatures, for example, the panels of decorative script are integrated within the overall pattern as areas of textural enrichment; and, with the margin and inset frames, these panels serve also as concrete screens and prosceniums to the action depicted, the participants in the narrative episode making their exits and entrances across or behind them.

Scroll painting. Hand scrolls, traditional to China and Japan, are ink paintings on continuous lengths of paper or silk. They are unrolled at arms' width and viewed from right to left. These generally represent panoramic views of rivers, mountain and urban landscapes and domestic interiors. They also illustrate romantic novels, Taoist and Buddhist themes, and historical and genre subjects. Narrative poetic commentaries were included as integral textures in the flowing design. The scrolls are remarkable for their vitality, the lyrical representation of atmospheric space, and for the rising and dipping viewpoints that anticipate the zooming motion-picture camera. The earliest surviving scrolls, such as Ku K'ai-chih's "Admonitions of the Imperial Perceptions," date from the 4th century AD. Oriental hanging scrolls and Indian and Tibetan temple banners are forms similar to those of Western easel and panel paintings. Their subjects range from the seasons (symbolized in bird, fruit, and tree motifs), domestic interiors, landscapes, and portraits to Vishnu epics, mandalas (a circle enclosing a square with a deity on each side), and temple icons. They are painted in ink or gouache on silk and paper and are usually mounted on embroidered or block-printed silk. The dramatic interplay of bold, flattened images against the open space of an unpainted or gilded ground influenced 19th-century Western Art Nouveau decoration.

Hanging scrolls and temple banners

Screen and fan painting. Folding screens and screen doors originated in China and Japan, probably during the 12th century, and continued as a traditional form into the 20th. They are in ink or gouache on plain or gilded paper and silk. Their vivid rendering of animals, birds, and flowers and their atmospheric landscapes brought nature indoors. In some screens each panel was designed as an individual painting, while in others a continuous pattern flowed freely across the divisions. Japanese screens were often painted in complementary Yin and Yang pairs. Large 12-panel Chinese coromandel lacquer screens were imported into Europe during the 17th and 18th centuries. French Rococo boudoir screens depicting *fêtes champêtres* (townspeople enjoying rural surroundings) and toile de Jouy (landscape or floral) pastoral themes were painted on silk or on wood panels in a flamboyantly scrolled, gilded framework. The designs of Art Nouveau screens were inspired by the Japanese tradition. Sidney Nolan's screens on Greek themes and the pastiches of Victorian paper-scrap screens by Pop Art painters are recent Western revivals. Traditional to the Greek and Russian Orthodox churches is the iconostasis screen, which stands between the nave (central part of the church) and sanctuary and displays icon panel paintings in rows representing the Virgin, the saints, and narrative subjects.

Rigid fans are depicted in the paintings and reliefs of ancient Egypt, Assyria, Greece, and Rome, but the oldest surviving specimens are the round and folding fans from Japan and China. These were painted in India ink and colour on paper, card, and silk, the ground often sprinkled with gold dust or laid with gold or silver leaf. Spread freely across the mount, a calligraphic design depicted seasonal landscapes, genre scenes, and bird, flower, and animal motifs, with accompanying poems and commentaries. Leading Oriental painters produced much of their finest work in this form. In Europe, however, where fan painting had been rarely practiced until the 17th century, it was considered a minor art, and designs were often based on frescoes and easel paintings. The richest and most elegant of these were painted in France and Italy during the 18th century. Watercolour and gouache paintings and hand-coloured engraved designs were made on paper, card, kid, and gauze. Allegories and romantic pastoral landscapes were frequently designed as separate vignettes, linked by floral swags and border scrolls. Both sides of the mount might be painted. The guards and sticks of the spoke framework were in delicately carved wood or ivory, inlaid with gold leaf or mother-of-pearl. Round hand-screens of parchment, mounted on handles like lollipops, were popular in early 19th-century English society. Charles Conder was a notable fin de siècle ("end of the century," characterized by effete sophistication) fan painter and, more recently, Oskar Kokoschka decorated a lively set of fans on an autobiographical narrative theme.

Round hand-screens

Panoramas. Panoramas were intended to simulate the sensation of scanning an extensive urban or country view or seascape. This form of painting was popular at the end of the 18th century. Notable examples are "The Battle of Agincourt" (1805), by R.K. Porter, and "Scheveningen," by Hendrik Willem Mesdag. Panoramas might be compared to cinerama films and enjoyed as a stimulating optical entertainment, along with cyclorama drums (large pictorial representations encircling the spectator), trompe l'oeil diorama peep shows, and the show box for which Thomas Gainsborough painted glass transparencies. More serious forms of panoramic painting are exemplified in Chinese Buddhist sanctuary frescoes, Oriental hand scrolls, Dürer's watercolour townscapes, Andrey Rublyov's 14th-century mural of Moscow, and Uccello's original sequence of three panels depicting the Battle of San Romano.

Modern forms. The concept of painting as a medium for creating illusions of space, volume, texture, light, and movement on a flat, stationary support has been challenged by many modern artists. Some recent forms, for example, have blurred the conventional distinctions between the mediums of sculpture and painting. Sculptors such as David Smith, Eduardo Paolozzi, and Philip Sutton have made multicoloured constructions; painters such as Jean Arp and Ben Nicholson have created abstract designs in painted wood relief, and Richard Smith has painted on three-dimensional canvas structures the surfaces of which curl and thrust towards the spectator. And, rather than deny the essential flatness of the painting support by using traditional methods of representing volume and texture, Robert Rauschenberg and Jim Dine have attached real objects and textures to the painted surface, and Frank Stella and Kenneth Noland have designed their irregularly shaped canvases to be seen as explicitly flat art objects. Rejecting earlier painting methods of reproducing effects of light with tonal con-

trasts and broken, pigment colour, some artists have made use of neon tubes and mirrors. Instead of simulating sensations of movement by optical illusion, others have designed kinetic panels and boxes in which coloured shapes revolve under electric power. The traditional definition of painting as a visual, concrete art form has been questioned by recent aspects of Conceptual art, in which the painter's idea might be expressed only in the form of documented proposals for unrealized and often unrealizable projects.

IMAGERY AND SUBJECT MATTER

The imagery and subject matter of paintings in early cultures were generally prescribed by tribal, religious, or dynastic authorities. In some Eastern countries, traditional models survived into the 18th century and even later. With the Renaissance, however, images and themes in Western painting, reflecting the new spirit of Humanistic, objective curiosity and scientific research, came to be decided by the artist and his patron and, in more recent periods, by the artist alone.

Kinds of imagery. Within the various cultures the art of representing things by painted images has rarely shown a continuously developing pattern toward greater realism. More often, religious and philosophical precepts have determined the degree of naturalism permitted. Rules governing portrayals of the human figure have been particularly stringent in certain traditions of representational painting, reflecting different attitudes to the cosmic significance of man. For example, a belief in the inferiority of man in relation to an almighty deity is expressed in the faceless figures in early Jewish painting and in the dehumanized stylizations of Byzantine imagery; and his insignificance against the dynamic forces of nature is symbolized in Chinese landscape paintings by his puny scale within a monumental setting. An earlier view, which instead sought to glorify the spiritual, intellectual, and physical attributes of mankind, is typified in the noble figures of Greco-Roman art and in the renewed celebration of human physical beauty in the Renaissance and subsequent Neoclassical styles. The uniqueness of man among living things and the expression of his individual physical and emotional characteristics is exemplified in Japanese and northern European narrative and genre painting. Concomitant with the antipathy toward figurative representation in some cultures was a general distaste for the portrayal of all things of the exterior world, animals, landscape features, and other natural forms rarely appearing except as stylized images signifying spiritual forces of good and evil. The representational imagery of modern painting borrows freely from ancient and contemporary sources such as primitive and child art, Classical mythology, commercial advertising, press photography, and the allegories and fantasies of the motion picture and the comic strip. Nonrepresentational imagery is not restricted to modern painting but appears also in earlier forms such as Aurignacian (Paleolithic) decorative meanders, the scrollwork of Celtic illuminations, and the patterns of Islāmic Kūfic calligraphy (an angular variety of the Arabic alphabet). And the abstraction of natural forms into rudimentary symbols, characteristic of modern painting, is echoed in the "pin-men" conventions of Magdalenian caves, in Aztec pictograms, and Indian and Tibetan cosmic-diagram paintings.

Attitudes toward the cosmic significance of man

Kinds of subject matter. *Devotional.* The range and interpretation of subjects in different forms of devotional painting express a particular attitude to the relationship between man and his deity. Early Christian and Buddhist murals, for example, portrayed an all-powerful, remote, and mysterious being, painted as a flat, formalized head or figure whose stern gaze dominated the interiors of temples, churches, and sanctuaries. Christian Last Judgments and Oriental hell paintings were intended to frighten the believer, while subjects such as the Virgin enthroned, the Assumption, and Buddha descending from Paradise sustained his faith with hopes for salvation and rewards of blissful immortality.

Narrative. When the autocratic ecclesiastical control over Western painting weakened under Renaissance Humanism, the religious narrative picture became a window onto a terrestial rather than a celestial world. Both emotional and physical relationships between the figures depicted were realistically expressed, and the spectator was able to identify himself with the lifelike representation of a worldly space inhabited by Christ, his disciples, and saints, wearing updated dress and moving naturally within contemporary settings. This kind of narrative interpretation persists in the modern religious paintings of Sir Stanley Spencer, where biblical environments are represented by the clipped hedgerows, the churchyards, and the front parlours of his neat, native English village of Cookham.

Allegorical narrative subjects

Allegorical narrative subjects might exalt the sensuous arts, as in the symbolic muses portrayed by Poussin and Luca Signorelli and the paradisiac gardens of 15th-century French illuminated manuscripts. But they might also carry warnings. In the 16th century, Pieter Bruegel, for example, combined overt and often grotesque symbols with subtle visual metaphors to point stern morals in such paintings as "The Triumph of Death" (alluding to the "wages of sin"), "The Land of Cockaigne" (attacking gluttony and sloth), and "Mad Meg" (ridiculing covetousness). Even Bruegel's apparently straightforward genre subjects, such as "The Peasant Dance" and the festival of "The Fight between Carnival and Lent," conceal parables on man's follies and sins, while Hieronymus Bosch introduced abstruse, allegorical phantasmagoria into such traditional narratives as "The Temptation of St. Antony" and "The Prodigal Son" and made his "Garden of Delights" an expression of disgust rather than of joy. Botticelli's late paintings, probably produced under the influence of the 15th-century Italian monk and reformer Girolamo Savonarola, are other savagely pessimistic allegories: "The Story of Virginia Romana" and "The Tragedy of Lucretia" representing virtue upheld only by death and "The Calumny of Apelles," in which envy, suspicion, deceit, guile, repentance, and truth are identified, like medieval mummers, by their costume, pose, and gesture. Rubens, however, found in allegorical symbolism a means of dramatizing mundane state commissions, such as "The Union of Scotland and Ireland" and "The Bounty of James I (Triumphing Over Avarice)." Among famous 19th-century allegories are Delacroix's "Liberty Leading the People" and Pierre-Paul Prud'hon's "Crime Pursued by Vengeance and Justice."

Possibly the highest achievements in narrative illustrations to poetry and literature are found in Eastern miniatures and Oriental scrolls, such as the Persian paintings of Ferdowsī's 11th-century national epic poem, the *Shāh-nāmeh*, and the 12th-century Japanese scrolls of the *Genji monogatari* and the *Story of Ben Dainagon*. An example of modern literary painting is Sidney Nolan's narrative series portraying the Australian folklore hero Ned Kelly.

Ancient Greek and Roman mythologies have provided Western artists with rich sources of imagery and subject matter and with opportunities for painting the nude. Historical narrative painting includes Classical mythology and heroic legend, as well as the representation of contemporary events; examples include Benjamin West's "Death of Wolfe," Théodore Géricault's "Raft of the Medusa," and Goya's "The 3rd of May in Madrid."

Portraiture. The earliest surviving portraits of particular persons are probably the serene, idealized faces painted on the front and inside surfaces of dynastic Egyptian sarcophagi. The human individuality of the Roman mummy portraits of the 1st and 2nd century AD, however, suggests more authentic likenesses. Although portraits are among the highest achievements in painting, the subject poses special problems for the artist commissioned to paint a notable contemporary. The portraits of patrons by artists such as Raphael, Rubens, Hyacinthe Rigaud, Antoine-Jean Gros, Jacques-Louis David, and Sir Thomas Lawrence were required to express nobility, grace, and authority, just as the sultans and rajahs portrayed on frontispieces to Persian and Indian illuminated books and albums had understandably to be flattered as benevolent despots. Such concessions to the sitter's vanity and social position seem to have been disregarded, how-

ever, in the convincing likenesses by more objective realists such as Robert Campin, Dürer, Jan van Eyck, Velázquez, Goya, and Gustave Courbet. Probably the finest are the self-portraits and studies of ordinary people by Rembrandt and Van Gogh, where psychological insight, emotional empathy, and aesthetic values are fused. A more decorative approach to the subject is seen in the flattened portraits by Holbein, the Elizabethan and itinerant naïve U.S. limners, and the Far Eastern paintings of ancestors, poets, priests, and emperors. Like these paintings, the full-length portraits by Boucher, Gainsborough, Kees van Dongen, and Matisse display as much regard for the texture and form of their sitters' dress as for their facial features.

Photographs and painted portraits

Since the development of photography, however, portraiture has been rarely practiced for its own sake as a serious art form, except where artists such as Cézanne and Braque have used it as a subject for structural research or—like Amedeo Modigliani, Chaim Soutine, and Francis Bacon—for the expression of a personal vision beyond the scope of the camera.

Genre. Genre subjects are scenes from everyday life. Hunting expeditions and tribal rituals figure in prehistoric rock paintings. Domestic and agricultural occupations, with banquet scenes of feasting, dancing, and music, were traditional subjects for ancient Egyptian tomb murals. Far Eastern hand scrolls, albums, and screens brilliantly describe court ceremonies, the bustle of towns, and the hardships of the countryside. The depiction of earthly pursuits was forbidden under the strict inconography prescribed by the early Christian Church, but the later illuminated Books of Hours provide enchanting records of the festivals and occupations of northern European communities. In Renaissance painting, genre subjects were generally restricted to background features of portraits and historical narratives. Domestic scenes, however, not only provided Bruegel with subjects for moral allegories but, as with Rembrandt, were used to counterpoint the emotional intensity of a dramatic religious theme. The withdrawal of religious patronage in northern Europe directed painters towards secular subjects. The rich period of genre painting in 17th-century Holland is represented by the interiors, conversation pieces, and scenes of work and play by David Teniers the Younger, Frans Hals, Jan Steen, Gerard Terborch, Pieter de Hooch, Adriaen van Ostade and, the finest, by Jan Vermeer. Pictures of rustic life had a special appeal for collectors in 18th-century France and England; these were the somewhat picturesque representations of peasant life painted by Jean-Baptiste Greuze, Boucher, George Morland, and Gainsborough. Jean-Baptiste-Siméon Chardin's paintings of servants and children, however, exhibit a timeless dignity and grandeur. The harsher realities of working life were depicted by Jean-François Millet, Daumier, Courbet, Van Gogh, and Degas; the robust gaiety of cafés and music halls was captured by Toulouse-Lautrec, John Sloan, Everett Shinn, and Walter Richard Sickert; and intimate domestic scenes were recorded by Bonnard and Vuillard. Modern genre movements have included the American Scene painters, the Ashcan and Kitchen Sink schools (represented by such painters as George Wesley Bellows, Jack Smith, and Derrick Greaves), the Camden Town and Euston Road groups (Frederick Spencer Gore, Sir William Coldstream, and Victor Pasmore), and the Social Realists in England and in the United States (Robert Henri, Stuart Davis, and Maurice Prendergast).

Landscape. Idealized landscapes were common subjects for fresco decoration in Roman villas. Landscape painting (as exemplified by a Chinese landscape scroll by Ku K'ai-chih dating from the 4th century) was an established tradition in the Far East, where themes such as the seasons and the elements held a spiritual significance. In Europe, imaginary landscapes decorated 15th-century Books of Hours. The first naturalistic landscapes were painted by Dürer and Bruegel. Landscapes appeared in most Renaissance paintings, however, only as settings to portraits and figure compositions. It was not until the 17th-century Dutch and Flemish schools—of Rembrandt, Jacob van Ruisdael, Meindert Hobbema, Aelbert Cuyp, Rubens, and Hercules Seghers—that they were accepted in the West as independent subjects. The most significant developments in 19th-century painting, however, were made through the landscapes of the Impressionists and the Neo-Impressionists and Postimpressionists. Styles in landscape painting range from the tranquil, classically idealized world of Poussin and Claude, the precise, canal topography of Francesco Guardi and Canaletto, and the structural analyses of Cézanne to the poetic romanticism of Samuel Palmer and the later Constables and Turners and the exultant pantheism of Rubens and Van Gogh. Modern landscapes vary in approach from the Expressionism of Oskar Kokoschka's cities and rivers, Maurice de Vlaminck's wintry countrysides, and John Marin's crystalline seascapes to the metaphysical country of Ernst, Dalí, and René Magritte and the semi-abstract coastlines of Nicolas de Staël, Marie Hélène Vieira da Silva, and Richard Diebenkorn.

Still life. The earliest European still-life painting is usually attributed to Jacopo de'Barbari (*i.e.*, "Dead Bird," 1504). In Western paintings, still life often appears as a minor feature of the design; but until the 17th century it was not generally painted for its own sake, although it was already traditional to Far Eastern art. The subject is particularly associated with northern European painting, and the choice of objects very often has a religious or literary significance: wine, water, and bread symbolizing the Passion; skulls, hourglasses, and candles, the transience of life; and selected flowers and fruits, the seasons. Flower painting, especially, held a spiritual and emotional meaning for Japanese artists and for 19th-century European painters, such as Odilon Redon, Paul Gauguin, and Van Gogh. Still life has been expressed in many different ways: Giuseppe Arcimboldo's witty arrangements of fruit, flowers, and vegetables made into fantastic allegorical heads and figures; the sensuous representation of food by Frans Snyders, Goya, and William Merritt Chase; the trompe l'oeil illusionism of François Desportes and William Harnett; the formal decoration of folk artists or primitives such as Henri Rousseau and Séraphine and of modern painters such as Matisse, Dufy, and Pat Caulfield; the semi-abstract designs of Picasso, Gris, and William Scott; and, probably at its highest level of expression, the majestic still lifes of Chardin, Cézanne, and Giorgio Morandi.

Subject matter of still lifes

Other subjects. Since ancient times, animals and birds have provided the primary subject matter of a painting or have been included in a design for their symbolic importance. In the paintings of prehistoric caves and dynastic Egyptian tombs, for example, animals are portrayed with a higher degree of naturalism than human figures. Their texture, movement, and structure have provided some artists with a primary source of inspiration: the classical, anatomical grace of a George Stubbs' racehorse and a more romantic interpretation in the ferocious energy of a Rubens and Géricault stallion; the vivid expression of rhythmically co-ordinated movements of deer by Tawaraya Sotatsu and Antonio Pisanello; the weight and volume of George Morland's pigs and Paul Potter's cows; the humanized creatures of Gothic bestiaries and of Edward Hicks' "Peaceable Kingdom"; and, finally, Dürer's "The Hare," which is possibly as famous as Leonardo's "Mona Lisa."

Increasing interest is shown in notable painters' versions of other artists' works. These are not academic copies (such as the study made by Matisse, when a student, of Chardin's "La Raie"), but creative transcriptions. Examples that can be appreciated as original paintings are those by Miró of Sorgh's "Lute Player"; by Watteau of Rubens' "Apotheosis of James I"; by Degas of Bellini's "Jealous Husband"; by Caulfield of Delacroix's "Greece Expiring on the Ruins of Missolonghi"; by Larry Rivers of Jean-Auguste-Dominique Ingres's "Mlle Rivière"; and by Picasso of Manet's "Déjeuner sur l'herbe," Velázquez's "Las Meñinas," and Delacroix's "Woman of Algiers" (which produced Roy Lichtenstein's "Femmes d'Alger, After Picasso, After Delacroix"). Picasso has also painted free versions of works by El Greco, Lucas

Cranach, Poussin, and Courbet, as Rubens had of Mantegna and Titian, Rembrandt of Persian and Indian miniatures, Cézanne of Rubens and El Greco, and Van Gogh of Millet, Gustave Doré, and Delacroix.

The subject of an abstract painting

In an abstract painting, ideas, emotions, and visual sensations are communicated solely through lines, shapes, colours, and textures that have no representational significance. The subject of an abstract painting may be therefore a proposition about the creative painting process itself or exclusively about the formal elements of painting, demonstrating the behaviour of juxtaposed colours and shapes and the movements and tensions between them, their optical metamorphosis and spatial ambiguities. Many abstracts, however, are more than visual formal exercises and produce physical and emotional reactions in the spectator to illusions of shapes and colours that appear to rise and fall, recede and advance, balance and float, disintegrate and re-form; or of moods created of joy, sadness, peace, or foreboding; or of effects produced by light or by flickering or throbbing movement. Some abstracts evoke the atmosphere of a particular time, place, or event; and then their titles may be significant: "Pancho Villa, Dead and Alive" (Robert Motherwell); "Late Morning" (Bridget Riley); "Broadway Boogie Woogie" (Mondrian); "Gold of Venice" (Lucio Fontana); "Capricious Forms" (Kandinsky).

SYMBOLISM

Most early cultures developed iconographic systems that included prescriptions for the site, design, function, form, medium, subject matter, and imagery of their painting. The siting of early Byzantine murals, for instance, echoed the symbolic, architectural planning of the basilica. Thus, a stylized, linear image of Christ, surrounded by heavenly hosts, occupied the central dome; the Virgin was represented in the apse; and stiff figures of apostles, prophets, martyrs, and patriarchs occupied the aisle walls. The format of early devotional paintings was also prescribed, Christian and Buddhist deities being placed in the focal centre of the design, above the eye level of the audience and larger than surrounding figures. And, in the conventional arrangement of a Christian subject such as the Holy Trinity, a central, bearded, patriarchal God, flanked by archangels, presented Christ on the cross; between them was a dove, representing the Holy Spirit. In a rendering of the risen Christ, the Son faced the audience, with the Virgin Mother on the left and St. John on the right of the design. In the Far East a traditional format depicted Buddha on a lotus throne or in a high chariot drawn by oxen across clouds, surrounded by figures representing the planets. Deities generally appear against undefined grounds of white (signifying eternity or nothingness), blue (the celestial vaults), or gold (representing heavenly light by radiating lines or the spiritual aura by a nimbus). The elaborate surface preparation of supports and the painstaking execution with the finest materials symbolized the intention that paintings dedicated to a deity should last forever. The imagery, subject matter, and form might also have a mystical function: the realistic rendering of animals in contrast to the perfunctory human representations in Ice Age rock paintings, thought to signify a wishful guarantee for success in hunting; the earthly pleasures depicted on ancient Egyptian tomb murals intended to secure their continuance for the deceased; and the North American Indian sand paintings designed for magic healing ceremonies and the Tantrik (relating to Tantrism, a school of Mahāyāna Buddhism) mandalas used for meditation and enlightenment.

Symbolism in Eastern painting—intended to deepen the experience of a picture's mood and spirituality—is more generalized and poetic than in Western art. Both the execution and the subject matter of Buddhist Chinese and Japanese painting have a religious or metaphysical significance: the artist's intuitive, calligraphic brush movements symbolizing his mystical empathy with nature and his cyclic landscape and flower subjects expressing his belief in the spiritual harmony of natural forms and forces. Much of Indian symbolism is visually emotive, images such as snakes, plantain leaves, twining creepers, and rippling water being overtly sexual. And, although symbolic attributes and colour codes identify Indian mythological characters (for example, the four arms of the terrible goddess Kālī and the blue skin of the divine lover Kṛṣṇa [Krishna]), the formal character and colour scheme of settings generally reflect the narrative's emotional mood (for example, vibrant, dark-blue, cloudy skies and embracing, purple-black glades evoking amourous anticipation and red grounds expressing the passions of love or war).

Western symbolic systems

Western symbolic systems, however, are more intellectually directed, their imagery having precise literary meanings and their colour codes intended primarily for narrative or devotional identification. The iconographic programs of the early Christian churches, for example, laid down complex formulas for the viewpoints, gestures, facial expressions, and positions of arms, hands, and feet for religious figures. An elaborate Ethiopian Christian iconographic system was followed until very recently, and elsewhere traditional methods survive of identifying archangels and saints by their attributes and by the symbols of martyrdom that they display: distinguishing white-bearded St. Peter from black-bearded St. Paul, for example, and portraying St. Catherine with a wheel and St. Bartholomew with a knife and skin. Christian iconography adopted and elaborated Greco-Roman and Jewish symbolic imagery: the pagan signs of the vine and the fish, for example, and the image of Christ as the Good Shepherd based on the Greek Hermes Kriophoros. Medieval and Renaissance writings define an immense vocabulary of symbolic images, such as the crescent, sea urchin, and owl signifying heresy, the toad and jug representing the devil, and the egg and bagpipes as erotic symbols (all of which appear in Hieronymus Bosch's 15th-century narrative moralities). Angels and devils, hell fire and golden paradise, heavenly skies and birds in flight representing spirituality and rebirth are examples of the similarity of symbolic meaning for many religious, mythological, and allegorical traditions. The significance of images common to several cultures, however, may also be very different: the dragon representing avarice in European medieval allegory symbolizes friendliness in Japanese Zen painting; and the snake, symbol of temptation and eroticism in the West, signifies, by its skin shedding, the renewal of life in Far Eastern iconography.

BIBLIOGRAPHY

General reference—dictionaries and encyclopaedias: (*On artists, movements, and techniques*): PETER and LINDA MURRAY, *A Dictionary of Art and Artists*, rev. ed. (1968), not including Eastern art; *The McGraw-Hill Encyclopedia of World Art*, 15 vol. (1959–68); and HAROLD OSBORNE (ed.), *The Oxford Companion to Art* (1970), both include extensive bibliographies; *Praeger Encyclopaedia of Art* (1971); CARLTON LAKE and ROBERT MAILLARD (eds.), *Dictionnaire de la peinture moderne* (1954; Eng. trans., *A Dictionary of Modern Painting*, 3rd ed. rev., 1964); and RALPH MAYER, *A Dictionary of Art Terms and Techniques* (1969), concise modern art references. (*On historical background of, and relationships between, the visual arts*): RENE HUYGHE (gen. ed.), *The Larousse Encyclopedias*, by periods: *Prehistoric and Ancient Art*, rev. ed. (1966); *Byzantine and Medieval Art*, rev. ed. (1968); *Renaissance and Baroque Art* (1958, 1968); *Modern Art* (1961, 1967).

Design: (*Elements of design in painting—form, space, movement*): JOHANNES ITTEN, *Mein Vorkurs am Bauhaus: Gestaltungs- und Formenlehre* (1963; Eng. trans., *Design and Form: The Basic Course at the Bauhaus*, 1965); HOWARD GARDNER, "From Mode to Symbol," *British Journal of Aesthetics*, 10:359–375 (1970); ROBERT J. GOLDWATER, *Primitivism in Modern Art* (1967); JOHN WHITE, *The Birth and Rebirth of Pictorial Space* (1957); GYORGY KEPES (ed.), *The Nature and Art of Motion* (1965). (*Colour theory*): R.M. EVANS, *An Introduction to Color* (1948); EGBERT JACOBSON, *Basic Color: An Interpretation of the Ostwald Color System* (1948); FABER BIRREN, *Creative Color* (1961) and *A Grammar of Color* (the Munsell System) (1969); ANDREAS KORNERUP and J.H. WANSCHER, *Farver i farver* (1961; Eng. trans., *Methuen Handbook of Colour*, 1963). (*Important modern standard works on colour*): JOHANNES ITTEN, *Kunst der Farbe* (1961; Eng. trans., *The Art of Color*, 1961); JOSEF ALBERS, *The Interaction of Color* (1963). (*On colour in particular periods and movements in painting*): L.G. DERUIS-

SEAU, "Colours in the Renaissance," *CIBA Review*, 17:601–603 (1939); BARBARA ROSE, "The Primacy of Color," *Art International*, 8:22–26 (1964). (*On modern painting*): WILLIAM INNES HOMER, *Seurat and the Science of Painting* (1964); ROBERT L. HERBERT, *Neo-Impressionism* (1968); R.W. PICKFORD, "The Influence of Colour Vision Defects on Painting," *British Journal of Aesthetics*, 5:211–226 (1965). (*Principles of design*): MICHAEL AYRTON, *Golden Sections* (1957); GYORGY KEPES (ed.), *Module, Symmetry, Proportion* (1966). (*Design relationships between painting and other visual art mediums—influence of photography on artists*): AARON SCHARF, *Art and Photography* (1968).

Mediums: Standard works on most painting materials, supports, surfaces, and techniques include: MAX DOERNER, *Malmaterial und seine Verwenderung im Bilde*, 4th ed. (1933; Eng. trans., *The Materials of the Artist and Their Use in Painting*, rev. ed., 1949, reprinted 1969); W.G. CONSTABLE, *The Painter's Workshop* (1954); RALPH MAYER, *The Artist's Handbook of Materials and Techniques*, 3rd ed. rev. (1970), with extensive bibliography; KURT HERBERTS, *The Complete Book of Artist's Techniques* (1958); MARIA BAZZI, *Abecedario pittorico* (1956; Eng. trans., *The Artist's Method and Materials*, 1960), with bibliography of important treatises on mediums and techniques; FREDERIC TAUBES, *A Guide to Traditional and Modern Painting Methods* (1963); HILAIRE HILER, *The Painter's Pocket Book of Methods and Materials*, 3rd ed. (1970). (*Tempera*): Included in D.V. THOMPSON, *The Materials of Medieval Painting* (1936; reprinted as *Materials and Techniques of Medieval Painting*, 1956) and RALPH MAYER, *The Painter's Craft* (1948). (*Fresco*): OLLE NORDMARK, *Fresco Painting* (1947). (*Oil*): FREDERIC SCHMID, *The Practice of Painting* (1948), in the 18th century; CHARLES JOHNSON, *The Language of Painting* (1949); FREDERIC TAUBES, *The Mastery of Oil Painting* (1953). (*Watercolour*): R.L. BINYON, *English Watercolors*, 2nd ed. (1944); I.A. WILLIAMS, *Early English Watercolours* (1952); GRAHAM REYNOLDS, *A Concise History of Watercolors* (1971). (*Ink*): SHIO SAKANISHI, *The Spirit of the Brush* (1939); FEI CH'ENG-WU, *Brush Drawing in the Chinese Manner* (1957); PHILIP S. RAWSON, "The Methods of Zen Painting," *British Journal of Aesthetics*, 7:315–338 (1967); and included in JOSHUA TAYLOR, *Learning to Look* (1957) and OSVALD SIREN, *Chinese Painting: Leading Masters and Principles* (1958). (*Gouache*): ARNOLD BLANCH, *Methods and Techniques for Gouache Painting* (1946); ADOLPH DEHN, *Watercolor, Gouache and Casein Painting* (1955). (*Encaustic*): FRANCES PRATT and BECCA FIZELL, *Encaustic Materials and Methods* (1949). (*On the new acrylic paints*): FRED GITTINGS, *Polymer Painting Manual* (1971), thorough and well illustrated. (*Other mediums*): H.G. CLARKE, *The Story of Old English Glass Pictures, 1690–1810* (1928); HARRIET JANIS and RUDI BLESH, *Collage* (1962).

Imagery: (*Wide survey of different forms of visual imagery*): E.H. GOMBRICH, *Art and Illusion* (1960); GYORGY KEPES (ed.), *Sign, Image, Symbol* (1966). (*Modern representational imagery*): CHRISTOPHER FINCH, *Pop Art: Object and Image* (1968). (*Abstract imagery*): THOMAS B. HESS, *Abstract Painting: Background and American Phase* (1951).

Subject matter: (*Narrative*): a concise survey of Western mythological and religious subjects in HOWARD DANIEL, *Encyclopedia of Themes and Subjects* (1971). (*Portraiture*): LUDWIG GOLDSCHEIDER (ed.), *500 Self Portraits from Antique Times to the Present Day, in Sculpture, Painting, Drawing, and Engraving* (1936); MONROE WHEELER, *Twentieth Century Portraits* (1942); JOHN POPE-HENNESSY, *The Portrait in the Renaissance* (1966), includes interpretive discussion of the works and extracts from the artists' letters. (*Landscape*): KENNETH CLARK, *Landscape into Art* (1949, reprinted 1961); A. RICHARD TURNER, *The Vision of Landscape in Renaissance Italy* (1966). (*Still life*): HERBERT FURST, *The Art of Still Life Painting* (1927); and included in M.J. FRIEDLANDER, *Landscape, Portrait, Still-Life* (1949). (*Figure painting*): KENNETH CLARK, *The Nude: A Study in Ideal Form* (1956).

Symbolism: (*General*): ERWIN PANOFSKY, *Studies in Iconology* (1939, reprinted 1962), and *Meaning in the Visual Arts* (1955); archetypal symbolisms in F.D.K. BOSCH, *The Golden Germ* (1960); language of archetypal symbolism as communicated in dreams: CARL JUNG *et al.*, *Man and His Symbols* (posthumous ed. 1964), with excellent illustrations. (*Christian iconography*): GEORGE KAFTAL, *Iconography of the Saints in Tuscan Painting* (1952); PAUL FRANKL, *The Gothic Literary Sources and Interpretation Through Eight Centuries* (1960); GEORGE FERGUSON, *Signs and Symbols in Christian Art* (1959); L.D. ETTLINGER, *The Sistine Chapel Before Michelangelo* (1965); JOAN EVANS, *Monastic Iconography in France: From the Renaissance to the Revolution* (1970). (*Indian*): B. BHATTACHARYYA, *The Indian Buddhist Iconography* (1924); J.N. BANERJEA, *The Development of Hindu Iconography* (1956). (*Indian and Tibetan Tantrik symbolism*): A.K. GORDON, *The Iconography of Tibetan Lamaism* (1939).

Writings: (*Classical treatises*): LEONARDO DA VINCI'S *Notebooks*, ed. by EDWARD MACCURDY, 2 vol. (1938); Dürer's books discussed in ERWIN PANOFSKY, *Albrecht Dürer*, 2 vol. (1943); LEONE BATTISTA ALBERTI, *On Painting*, trans. by JOHN R. SPENCER (1956); GIORGIO VASARI, *Vasari on Technique*, ed. by G. BALDWIN BROWN, trans. by LOUISA S. MACLEHOSE (1961); *The Strasburg Manuscript*, trans. by V. and R. BORRADAILE (1966); WILLIAM HOGARTH, *The Analysis of Beauty*, ed. by J. BURKE (1955); JOSHUA REYNOLDS, *Fifteen Discourses Delivered in the Royal Academy* (1928); reprint of ALEXANDER COZENS, *A New Method of Assisting the Invention in Drawing Original Compositions of Landscape* (1785) included in PAUL OPPE, *Alexander & John Robert Cozens* (1952). (*Chinese standard textbook*): *The Mustard Seed Garden Manual of Painting, 1679–1701*, included in MAI-MAI SZE, *The Tao of Painting*, 2nd ed., 2 vol. (1963). (*Anthology of artists' and philosophers' writings on art*): ELIZABETH G. HOLT (ed.), *A Documentary History of Art*, 2nd ed., 3 vol. (1957–65). (*Letters by painters*): by Constable in C.R. LESLIE, *Memoirs of the Life of John Constable* (1951); JOHN REWALD (ed.), *Cézanne's Letters* (1941) and *Camille Pissarro's Letters to His Son Lucien* (1944); *Van Gogh: Complete Letters*, 3 vol. (1958). (*Writings and statements by modern painters*): MYFANWY EVANS (ed.), *The Painter's Object* (1937) and by Cubist painters in EDWARD F. FRY, *Der Kubismus* (1966; Eng. trans., *Cubism*, 1966); WASSILY KANDINSKY, *On the Spiritual in Art*, trans. by GEORGE WITTENBORN (1947); PAUL KLEE, *Pädagogisches Skizzenbuch* (1925; Eng. trans., *Pedagogical Sketchbook*, 1944); *Über die moderne Kunst* (1945; Eng. trans., *On Modern Art*, introduction by H. READ, 1948); *The Thinking Eye: The Notebooks of Paul Klee*, ed. by JURG SPILLER (1961) and his *Diaries, 1898–1918*, ed. by FELIX KLEE (1964); HENRI MATISSE, *Notes of a Painter* (1908), included in A.H. BARR, *Matisse: His Art and His Public* (1951); *The Purist Manifesto: Ozenfant and Le Corbusier, After Cubism* (1918, trans. and ed. by ANTH EARDLEY, 1971). (*Statements by modern American abstract painters*): ROBERT MOTHERWELL and AD REINHARDT (eds.), *Modern Artists in America* (1951); SELDON RODMAN, *Conversations with Artists* (1957); KATHARINE KUH, *The Artist's Voice: Talks with Seventeen Artists* (1962); *The New York School*, foreword by MAURICE TUCHMAN (1970), by painters and critics, with extensive bibliography.

(P.D.O.)

Paints, Varnishes, and Allied Products

Paint, a fluid suspension spread in thin coats to decorate and protect surfaces, consists of pigment or colouring matter, and the vehicle in which the pigment is suspended. The function of the vehicle is to form a tough film when applied to a surface and to bind the pigment to the surface. Paint may be applied to metal, wood, stone, paper, leather, cloth, or other surfaces.

Varnish is a liquid coating material containing a resin that dries to a hard transparent film. Though usually clear, varnishes occasionally may contain pigment. Lacquer is a varnish that solidifies by evaporation of the solvents contained within it rather than by one of the more complicated processes of film formation described later in this article. Lacquers may be clear and tinted, or opaque and coloured.

Certain types of surface coatings not ordinarily considered paints, though they have some characteristics in common with paints, are described in the articles PRINTING; PAPER AND PAPER PRODUCTS; ADHESIVES; and TEXTILE INDUSTRY.

HISTORICAL DEVELOPMENT

Origins. Paints were in use for representational and decorative purposes for thousands of years before the idea of using them as protective coatings appeared. The earliest known paintings, found in the caves of Lascaux, France, and Altamira, Spain, made with iron oxide and applied without binder, date from as early as 15,000 BC. Early peoples of Africa, Oceania, and the Americas also used paints to decorate temples and dwellings. The Egyptians prepared colours from soil and by 1500 BC imported such dyes as indigo and madder to make blue and red pigments. By 1000 BC they had developed a varnish from

First known paintings

the gum of the acacia tree (gum arabic) that contributed to the permanence of their art.

Asian art appears to have begun independently, with coloured crayons as pigment and clay as a binder. Natural ores served as sources of the first pigment, but calcined (fired) mixtures and organic pigments were developed at least before 6000 BC. Vehicles were prepared from gum arabic, egg white, gelatin, and beeswax. The use of lacquer in China goes back to prehistoric times. During the Chou dynasty (*c.* 1122–221 BC), it served for the decoration of carriages, harnesses, and weapons. By the 2nd century BC, Chinese buildings were decorated with lacquer both inside and outside; lacquer was also extensively used in Korea and Japan.

The first use of protective coatings was by the Egyptians, who employed pitches and balsams to seal ships, a practice followed by other peoples of the ancient world.

Not until the Middle Ages did paint as a preservative for exposed wood surfaces have considerable use. The paints developed then were handmade, using such costly raw materials as egg white; the craftsmen kept their formulas secret and their products were expensive. Medieval city streets, nevertheless, bore considerable evidence of the painters' art, with colourful tradesmen's signs and shop facades painted red or blue. In the 17th century white lead paint became widely available; yet ordinary houses and even such important structures as bridges remained unpainted for the most part until the 18th century, when there was an increased availability of both vehicles and pigments. Extensive exploitation of linseed oil from the flax plant and pigment-grade zinc oxide produced a rapid expansion of the paint-manufacturing industry. In the 19th century for the first time the two ingredients, pigment and vehicle, were brought together before the paint was marketed.

Emergence of paint science and technology. The 20th century brought an enormous proliferation of articles requiring protective coating; a corresponding proliferation of paint products was achieved by intensive research efforts. The dye industry contributed the knowledge needed to upgrade and evaluate pigments, and the plastics industry contributed the development of polymers, the giant molecules containing many repeating basic molecules.

The study of heat energy, or thermodynamics, begun in the 19th century, proved a useful tool for understanding chemical reactions of paint components and the more subtle interactions leading to compatibility and pigment wetting. A newer discipline, rheology—the science of the deformation and flow of matter—had a great impact; chemists and engineers appeared in paint-industry laboratories.

New coating materials

Many new types of coating materials have evolved: resins, solvents, plasticizers, pigments, driers, foam-control agents, and adhesion promoters. Fire retardancy, corrosion resistance, and heat stability have been achieved with selected new materials and new methods of application. The art of paint making has become a science.

The most significant change in paint technology in this century is the return to aqueous systems (water-base paints) at a level of sophistication and complexity never conceived by the Egyptians. The industry is stable; the raw materials are virtually unlimited; and the future portends a small but steady growth rate, notwithstanding competition from plastics, decorator fabrics, and ceramic tile.

PAINT MANUFACTURE

Composition. Paint is composed of a pigment, that hides the substrate (base on which it is applied) and imparts colour, and a vehicle, or binder, that carries the pigment onto the substrate and binds it there. Inorganic pigments are insoluble, metallic compounds the colours of which depend mainly on the electronic energy levels of the metal. Organic pigments are compounds containing colour-producing (chromophoric) groups of atoms within their molecules. Manufactured in batches of up to 2,500 gallons (9,500 litres), binders are composed of polymeric resins, usually dissolved in a solvent. A resin is a high molecular weight organic compound, soluble in organic solvents but not in water.

Dispersion. Paint manufacture involves bringing pigment particles and vehicle into close contact and achieving a thorough dispersion of the particles throughout the binder. Dispersion is in fact the most important operation in a paint factory; the economics of plant operation are keyed to it. The design of dispersion mills depends on the type of paint manufactured and on the formulations and colours the equipment must handle. The most popular types are so-called ball mills, which grind by tumbling heavy metal or ceramic balls through the paint in a cylindrical container, and sand grinders, which circulate a suspension of sand in paint through a rotor assembly at high speed. A fine grind can be achieved with ball mills operating for eight or more hours on single batches or with sand grinders delivering a nearly continuous output.

Paints are no longer hand mixed. Most of the important properties of a paint depend on the nature and extent of the interfaces (boundary surfaces) between the pigment particles and the vehicle. If the particles are not completely dispersed, with all the pigment particles wetted by vehicle, they link together in a reticulated structure that leads to poor brushability, low gloss, and the appearance of oversize particles in the film.

Oil absorption of a pigment

Oil absorption of a pigment, the volume of oil per unit weight of dry pigment needed to form a stiff coherent paste, indicates how well the pigment will disperse. Oil absorption is determined experimentally.

For each paint, there is an optimum ratio between the amount of pigment and the amount of binder. In the industry, this ratio is called critical pigment volume concentration (CPVC). As this critical figure is approached by adding more binder, gloss and blistering tendency decrease markedly, whereas permeation (penetration) of water and oxygen increase.

Control. Paint is usually made in batches. But there is a trend toward automation, especially in plants that make one product on a given line, even though continuous manufacture of paint is difficult (see illustration). If a large variety of colours or types must be made, the scheduling and control of the various operations become more important than the convenience of automation.

Colour adjustment. Paint products must be uniform in colour, flow, and other physical properties. Driers, fungicides, and tint bases are added during the process of thinning, and the paint is checked against a previously established colour standard. Skilled colour matchers have never been replaced in this phase of manufacture, although instruments are used to quantify the colour values of hue, brightness, and chroma (purity). As a rule, white pigments and extenders are used to provide hiding power by scattering light (see below), and the hue is provided by coloured pigments and toners that absorb light of certain wavelengths. By contrast, tint base used in the adjustment stage is a neutral colour that possesses good hiding qualities.

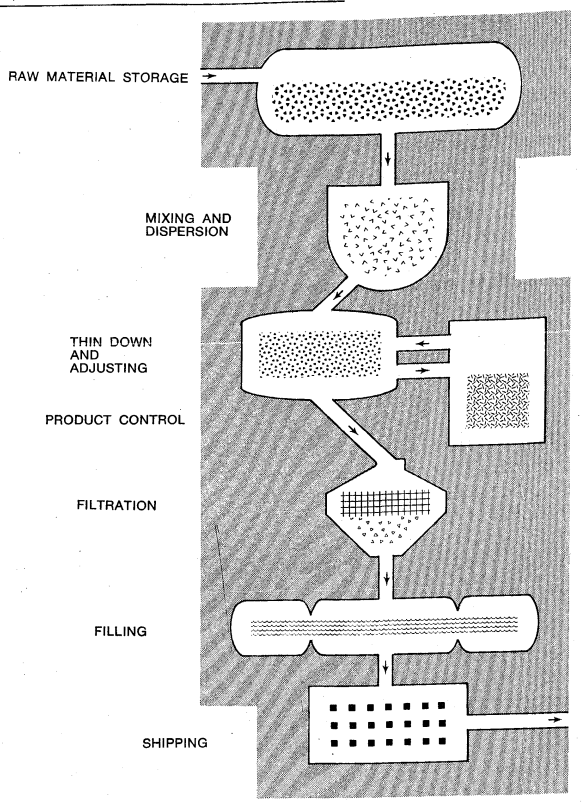
The purely technical aspects of the film are established instrumentally insofar as possible, and the observations are recorded instrumentally; but the final check is made by eye.

PIGMENTS AND OTHER COLORANTS

Pigment is the universally accepted term for the particulate matter dispersed in liquid or solid binders. Its two main functions are to impart colour and hiding and to improve hardness and durability.

Colour and hiding. The absorption of certain wavelengths of light gives rise to colour. The colour is the complement of the wavelengths absorbed.

When light strikes a pigment particle, the portions that are not absorbed are diverted in two ways, both of which result in hiding or obscuring the substrate. The light not absorbed may be reflected from the pigment surface or refracted as it passes through the pigment. Reflected light does not reach the substrate and hence cannot illuminate it at all. Refracted light, on the other



Stages in the paint manufacturing process.
Drawing by D. Meighan based on material courtesy of E.F. Wahl

hand, may reach the substrate, but at such an angle, and over such a tortuous path, that the diffuse reflection does not reach the eye. Reflection and refraction depend on a mismatch of refractive index between the pigment and the binder to scatter the light. Absorption is a characteristic that is present in dyes as well as pigments and that does not rely on light scattering. Hiding power has often been referred to the difference in refractive index between pigment and vehicle, but this difference plays a role in scattering.

Colour-defining systems

Colour is difficult to quantify. The best known colour-order system, the Munsell system, complete with ordered collections of colour chips, was originated in 1915 by an American artist, Albert Henry Munsell, and has since been refined. Another system in wide use that does not depend on collections of physical samples and is most often used in connection with instruments is the CIE (Commission internationale de l'éclairage) system, based on the matching of colours by mixing coloured lights from standard sources and providing a standard observer. In reducing the colour description to a set of quantitative terms, suitable for use in a computer, this system employs a set of standard coordinates, called tristimulus values. These values are used to match and guarantee the uniformity from batch to batch of one of the most important properties of a paint—its colour.

Pigments. Pigments are classified according to type or colour. Early classifications distinguished between mineral, or natural, pigments, which came from the earth and were either ground directly, or refined to a certain extent; and chemical pigments, which required a conversion of one sort or another. It is more convenient to classify pigments by colour, starting with the whites, then listing the blacks, and following with the more popular primary hues.

White pigments. White lead, the basic carbonate ($2PbCO_3 \cdot Pb[OH]_2$), has been manufactured for over 300 years by subjecting lead sheet to acetic acid fumes and carbon dioxide. The process takes several months.

Basic sulfate of lead ($PbSO_4 \cdot PbO$) is manufactured by heating lead in contact with sulfur. Reaction occurs when both the lead and the sulfur are vapourized. The basic silicate ($PbO \cdot SiO_2$) is formed as a coating on silica. Zinc oxide (ZnO) is the product of the reaction of zinc vapour with oxygen. Zinc sulfide (ZnS) is formed by precipitation out of aqueous solution and heating to a high temperature.

Lithopone ($ZnS + BaSO_4$) is a composite material, formed by precipitating zinc sulfate ($ZnSO_4$) and barium sulfide (BaS) together. Antimony trioxide (Sb_2O_3) is an exceptionally white pigment obtained by roasting antimony ore in air.

Titania, or titanium dioxide (TiO_2), is the superior pigment from the standpoint of hiding capability. Its preparation requires the removal of iron from the ilmenite ore $FeTiO_3$ (see also TITANIUM PRODUCTS AND PRODUCTION). Titanium dioxide is heated to a high temperature with small amounts of aluminum, antimony, silicon, or zinc oxide to produce white pigments with improved texture, resistance to light and moisture, and ease of dispersion in specific paint vehicles.

Black pigments. Most black pigments are made of elemental carbon. Carbon blacks are formed by allowing a smoking flame of natural gas to impinge onto a metal surface, usually iron. In the furnace process, the flame is not made to impinge on a surface but simply burns freely; the carbon is precipitated from the cooled gases. Lampblacks are obtained by burning oils in regulated amounts of air and allowing the carbon to fall by gravity from the combustion gases. Their blackness is inferior to carbon blacks, and they produce finishes of lower gloss. Bone blacks are made by charring bones. The char, which forms in the absence of air, consists of only about 10 to 20 percent carbon. Bone blacks are used principally in tinting when compatibility with other pigments is required. Graphite occurs naturally or can be prepared from coal in electric furnaces. It is an inferior pigment used only for specialty finishes. Amorphous graphite occurs in nature and is available at low cost. Mineral blacks derived from shale, peat, and coal dust are used as inexpensive pigments in undercoats.

Obtaining carbon black and lampblack

In addition to the carbon-containing black pigments, there exist iron oxide blacks found in nature or prepared. Blue basic lead sulfate, called blue lead, is a gray pigment that is used in the priming (*i.e.*, first coating) of structural steel.

Coloured pigments. It is convenient to group coloured pigments according to their position in the visible spectrum. Mineral varieties predominate at the red and blue ends and organic varieties in the centre.

Among the most common red pigments are red lead, iron oxide, cadmium reds, cuprous oxide, and various organics. Red lead (Pb_3O_4) is usually applied to steel for its protective value and is not used for tinting. This pigment is usually made by heating lead monoxide, or burning metallic lead. Iron oxide earth pigments (Fe_2O_3) are naturally occurring end products of the weathering of silicate rocks and were used in prehistoric times. The more important of these earth pigments are classified as ochres, the colorant of which is exclusively iron oxide and which tend toward the yellow part of the spectrum; siennas, which contain from 30 to 75 percent iron oxide and range from yellow-brown to reddish-orange depending on the variety and the degree of calcination; and umbers, which have a high manganese content in the form of manganese dioxide and tend toward brown coloration. Iron oxide ores, some of which can be ground directly to produce bright-coloured red pigments, include hematite ores, such as Spanish oxide and Persian Gulf oxide; magnetite (black); limonite (yellow); siderite (brown); and pyrites (dark brown). Synthetic iron oxide can be derived from ferrous sulfate ($FeSO_4$) by means of: calcination (heating to high temperatures) to produce grades ranging from light Turkey reds to dark Indian reds; calcination in the presence of lime (CaO), which produces a mixture of ferric oxide and calcium sulfate ($CaSO_4$) known as Venetian red; and precipitation from aqueous solution, which can produce any of three iron oxides of different hue.

Miscellaneous inorganic red pigments include cadmium reds, which rely on the mixture of cadmium sulfide (CdS) and selenide (CdSe) to produce colours ranging from

light red to maroon, and cuprous oxide (Cu_2O), which starts out as a bright red pigment but gradually oxidizes to the cupric form (CuO), which is black. Because cuprous oxide is a fungicide, this pigment is used in spite of its inherent chemical instability.

Organic red pigments include naturally occurring organic pigments, such as carmine and madder (mainly of historical interest, though still used as artists' colours), and synthetics. Synthetics are high-cost items, used because of the infinite variety of hues that can be achieved by minor changes in molecular structure.

Yellow and orange pigments

Yellow and orange pigments are obtained from metallic salts and from organic compounds. Chrome yellows and oranges are low-cost, bright, and stable pigments that display a considerable range of hues. Consisting of various proportions of lead chromate ($PbCrO_4$), lead sulfate ($PbSO_4$), and lead monoxide (PbO), all turn brown on exposure to sulfide gases in the air because of the formation of lead sulfide (PbS). Molybdate oranges, inclining toward the red, are made from lead chromate precipitated together with lead molybdate ($PbMoO_4$) and lead sulfate ($PbSO_4$). These also darken on exposure. Zinc yellow has a greenish cast because of a high content of chromic oxide. It is lightfast, rust inhibitive, and sulfide resistant. Basic zinc chromate ($5ZnO \cdot CrO_3$) is widely used as an undercoat. Cadmium yellow and oranges are formed by diluting yellow cadmium sulfide (CdS) with white zinc sulfide (ZnS) and barium sulfate ($BaSO_4$), using small amounts of red cadmium selenide (CdSe) to impart the orange coloration. Yellow-tinted titanium dioxide is produced by calcining it in the presence of nickel. The resulting product is very lightfast. Organic yellows and oranges are high-cost compounds (see DYES AND DYEING).

Blue and green pigments are mostly synthetic today. The natural blues and greens such as lapis lazuli have disappeared from coatings technology. In their place are synthetic ultramarines, produced by fusing china clay in a mixture of sulfur, sodium sulfate, and an organic reducing agent such as rosin. The glassy product is pulverized, leached free of sodium sulfate, and size-separated by sedimentation.

The most widely used blue pigments are the iron blues that contain various amounts of iron, ammonium and sodium ions (electrically charged particles), and water of crystallization (water trapped in the molecular structure as the pigment crystallizes). Two versions of iron blue are produced. Chinese blue has a greenish-blue tint; Milori blue tends toward the red. Hiding power of iron blue is good. Phthalocyanine pigments have a rather complicated organic structure; are reproducible, lightfast, brilliant; and make excellent tinting pigments but are expensive to produce.

The anthraquinone molecule ($C_{14}H_8O_2$), also used in dyes, can be made to form a permanent strong blue pigment that can be obtained in green hues by substituting chlorine atoms for most of the hydrogen atoms. It is an expensive pigment that withstands baking.

Iron blues mixed with chrome yellows in various proportions produce a popular class of chrome greens. Except for the darkening inherent in these lead-containing compounds (and the growing objections in many places to lead in coatings on grounds of health hazard), chrome greens offer the best combination of features for indoor and outdoor applications. Chromic oxide (Cr_2O_3) is a heat resistant, lightfast pigment.

Miscellaneous pigments. Finely divided aluminum, copper, bronze, and zinc are used to create a metallic or bronzing effect in coatings, particularly on metal panels, and also to impart corrosion resistance. The metals are produced in the form of powders or flakes that align themselves in the finished coating roughly parallel with the substrate. The flakes have good hiding power and are highly reflective, whereas spherical particles produce a less brilliant, sometimes milky, effect, with somewhat less protective value.

Fluorescent pigments include materials that absorb short wavelength radiation and re-emit light of longer wavelength. These are exceptionally brilliant if the emitted radiation falls in the same wavelength range as the reflected radiation. The molecular structure of the dyes responsible for this reinforcement must be such that electronic energy levels for the reflected and emitted light lie near each other. Phosphorescent pigments, which glow in the dark, are made from radioactive materials. Pearlescent pigments create a pearly, or nacreous, effect by diffraction of visible light.

Fluorescent pigments

Extenders. Certain inorganic salts that have little hiding power by themselves can be used to extend or conserve the more expensive hiding pigments; these are called extenders, or fillers. Extenders are also used to control the flow properties of paint, improving brushability and mechanical strength of the dried film. They sometimes decrease gloss and increase the resistance of a film to vapour transmission.

The main types of extenders are sulfates, which include natural or precipitated calcium or barium salts; carbonates, which comprise calcium or a mixture of calcium and magnesium salts; oxides, which include silica, either natural, precipitated, or hydrolyzed; and silicates, among which are clay, talc, mica, and calcium silicate. The types of coating that employ large volumes of extenders are architectural and maintenance finishes, particularly interior flat wall paints where hiding is not as critical as uniform coverage.

VARNISHES AND OTHER PAINT BINDERS

A surface coating protects the substrate against external influences such as moisture, light, and corrosive elements in the atmosphere. Thus, the binder for the pigment and whatever extenders are present are characterized by fluidity before application to the surface and rigidity soon afterward. This metamorphosis requires that a surface coating be either a solution from which the solids can be smoothly deposited on the substrate, a solution from which the binder will evaporate, or a liquid capable of solidifying or polymerizing to a rigid mass.

Natural binders. Originally, pigments were suspended in water, with and without binders. As binders the Persians used gum arabic; the Chinese, weak glue; the Asiatic Indians, boiled rice; the East Indians, shellac; and the Japanese, the sap from the wood-oil tree (which they called the varnish tree), possibly the first drying oil. The American Indians used a mixture of salmon eggs and fish oil.

Oil, wax, egg albumen, and glue were used in medieval Europe. During the Renaissance each artist developed his own formula for incorporating resins in drying oils. With the Industrial Revolution, copals and amber became the principal resins, and five classes of varnish were in use, only one of which contained linseed oil, which was to become the backbone of the industry in its formative stage.

Early varnishes

The first varnishes were solutions of natural resins, characterized by transparency, hardness, amorphous (noncrystalline) structure, and virtually no long-lasting integrity. They developed cracks on aging, even when used indoors; the utility of an outside coating relied largely on the natural tendency of pigments to adhere to the substrate by mechanical means.

Versatility is a characteristic of natural resins. They vary not only in solubility but also in transparency, brittleness, and crystallinity. Many are formed as natural secretions of plants or insects. Those resins that are compatible with oils (rosin, copals, amber) generally require heating. By contrast, the spirit-soluble resins (balsams, turpentine, lacs) usually dissolve without application of heat, but the danger of fire was a hazard in preparing their solutions with the highly volatile solvents.

Modern vehicles. Synthetic resins are now the heart of the paint industry. Not subject to variations in availability or quality as were their natural predecessors, these substances can be adapted with great chemical precision to various end uses and can be formed in many different suspending media.

There are two main types of synthetic resin: condensation polymers and addition polymers. Condensation polymers, formed by condensation of like or unlike molecules into a new more complex compound, include polyesters,

phenolics, amino resins, polyurethanes, and epoxies. Addition polymers, formed by direct chemical combination of substances into a single large molecule consisting of repeated units of the simpler original substance, include polyvinyl acetate, polyvinyl chloride, and the acrylates. (See also POLYMERS; PLASTICS AND RESINS.)

Condensation polymer resins. Alkyds (any of several common condensation polymer resins) are the most widely used protective-coating vehicles. Almost 1,000,000,000 pounds (450,000,000 kilograms) of these substances are consumed annually throughout the world. Brittle and insoluble in common solvents, they are always manufactured along with an agent (plasticizer) that makes them soluble and plastic. By manipulating their molecular structure in various ways, manufacturers can make alkyds more or less viscous, more or less hard as a coating, soluble in water or other substances, adherent, and capable of mixing successfully with various pigments.

Phenolic resins

Phenol (C_6H_5OH) reacts readily with aldehydes (any of various compounds containing the group —CHO, composed of carbon, hydrogen, and oxygen) to form oil-soluble compounds called phenolic resins, which are rapid drying, hard, chemically resistant, and abrasion resistant. The outdoor life of phenolic resins is not long, however, and they tend to yellow on aging.

Condensation of urea (H_2NCONH_2) or melamine ($C_3H_6N_6$) with formaldehyde (HCHO) produces compounds called amino resins, which are non-yellowing and alkali resistant and have good adhesive properties. When used in coatings, these amino resins contain an alcohol as a third component.

Both polyurethanes and epoxy resins are produced by means of complex organic reactions. Polyurethane coatings dry hard and inert on a substrate; epoxy resins have the qualities of toughness, adhesion, chemical resistance, rigidity, and thermal resistance.

Silicones are produced when various organic groups are attached to the silicon atom. They are hard and tough, with clarity and high gloss. They resist heat, weathering, and chemical attack and can be modified by blending or by chemical combination with other resins.

Addition polymers. Ethylene ($CH_2 = CH_2$), one of the simplest unsaturated organic molecules, can be made into translucent, inert, waxy polymers (called polyethylenes) at high pressures. Though the low-molecular-weight resins lack toughness and the higher are insoluble and unworkable, the inertness of the polymers gives them special interest in paint applications. When polyethylenes are applied to substrates, the resulting film is resilient and resistant to acids and oxidizers.

Polyvinyl acetate resin

One of the synthetic addition polymers that has largely replaced natural resins is polyvinyl acetate. This substance is a clear transparent resin that softens below the boiling point of water. It is sensitive to water and decomposes in acids and bases. Its solubility in a wide variety of organic solvents is good. It also has excellent adhesion properties.

Vinyl resins are marketed either dissolved or suspended in an organic medium like the natural resins or polymerized and marketed as aqueous emulsions called latexes.

Polyvinyl chlorides, though related to the compounds discussed above, lack desirable qualities as coatings. Acrylic compounds, on the other hand, which are clear, non-yellowing and heat stable, are used in the production of acrylic emulsion paints. Slight variation in the chemical composition of acrylic esters can produce wide variations in the physical properties of the polymer. Manufacture of acrylic emulsion paints consists of dispersing the pigment in water and adding the latex with only moderate agitation.

Solvents. Except for the solventless coatings (linseed oil and unsaturated polyesters) and the aqueous dispersions (vinyls and acrylics), it is necessary to dissolve solid resinous binders. Sometimes solvents are referred to as thinners because their inevitable effect is to reduce the viscosity of the resin solution. Plasticizers belong to the same category except that their volatility is not as pronounced as that expected of a solvent.

FILMS

Film formation. In the sense used by paint scientists, film formation involves conversion of a fluid coating, such as paint, into one with considerable rigidity.

Colloid chemical basis. Before the development of coatings science, little was known about the process of film formation except that a successful conversion required the film to maintain both integrity and adherence. Much of the modern thrust of paint science has been devoted to upgrading appearance and durability of materials that were known from the start to be film formers; little attention was paid to promising materials of good stability and desirable mechanical properties that did not form films.

The process of film formation is still not thoroughly understood. It is a far more complicated phenomenon than simple hardening or freezing. It involves transient and nonreversible conversions that are subject to many interdependent variables such as temperature, humidity, and the nature of the substrate. Film drying involves orientation of molecules at interfaces. Matter with thicknesses in the colloid range of dimensions is involved (see COLLOIDS).

Many of the test methods used by the paint industry are pragmatic: they are supposed to indicate when the coating is dry enough to be put in service. In effect they measure changes in the mechanical properties of the film and signify when they cross a particular threshold. All matter can be characterized by certain physical constants such as viscosity, modulus of rigidity, free volume, and density. Values of these constants, together with such basic chemical properties as bond strength, determine whether a film will form.

Types of film formation. There are two recognized ways to produce a rigid film on a substrate: continued growth of the polymer until it crosses the threshold between liquid and solid, and coalescence of finely dispersed polymer that has already terminated its growth as a molecule. The former method is chemical and the latter, physical. Combinations exist wherein the initial stage is physical and curing is effected chemically.

Physical and chemical film formation

The evolution of new concepts and new materials has broadened the number of variations of the film-forming process. Three strictly physical means of film-forming are: (1) from solution, in which evaporation of solvent leaves a deposit of polymer on the substrate; (2) from dispersions, in which evaporation of nonsolvent results in a deposit of particulate matter that coalesces into a continuous film; and (3) from melts, in which cooling of a supported fluid or plastic mass produces a rigid coating. Three strictly chemical means of film formation are: (1) by reaction with a component of the atmosphere in which the molecule combines with oxygen or water in the atmosphere; (2) by interaction of components already present in formulation, in which heat or radiation induce a reaction between species that coexist in the container without reacting; and (3) by reaction with a reagent or in the presence of a catalyst added just before application.

It is possible for solvent evaporation to accompany any of the three chemical means of film formation in a combination of physical and chemical drying. A combination of the two processes is provided by oil in the form of extended latexes, the initial rapid set of which is provided by evaporation and coalescence but the ultimate development of high rigidity of which is chemical.

Film formation by physical means. A solid resin or synthetic high polymer may be dissolved in a volatile solvent and converted by evaporation from a mobile fluid into a rigid film. Lacquer is an example of a coating that dries exclusively by this mechanism.

One of the standard polymerization techniques uses an emulsion directly as a film former, providing that the particles are deformable and hence are able to pack closely on drying. Only if the interstices and particle interfaces (points of contact) are obliterated will a clear, continuous film result. Acrylic ester emulsions may be formulated at extremely high molecular weights (up to 1,000,000), resulting in a polymer highly resistant to attacks by solvents.

Film formation by chemical means. Chemical reactions that occur during drying and hardening may lead to the formation of molecules of elongated or branched shape. When linseed oil dries and when epoxy resin hardens, the molecules link up and become elongated. In the case of some triglycerides and epoxies, elongation and cross-linking take place indefinitely, converting liquid films to a solid comprising one giant macromolecule. The greater the frequency of cross-links in the molecule, the more rigid the dried film.

Characteristics. Coatings are measured and characterized with instruments, ranging from the spatula of the craftsman and the penknife (or thumbnail) of the early development chemist to highly complex and expensive modern instruments.

Polymer properties. The properties of any substance depend on the size, shape, and composition of the molecules and on the magnitude and nature of the bonding forces within each molecule. Polymer structures are either open-chain or cross-linked; the former are generally thermoplastic and soluble, whereas the latter remain rigid at all temperatures below decomposition and merely swell in the presence of solvents.

Polymer properties and film formation

The properties of polymers that influence the quality of the ultimate film include the strength of bonding between atoms and between repeating groups of atoms in the chain of the polymer molecule; the crystallinity of the polymer—that is, the geometric arrangement of repeating atoms with the chain; and the manner in which polymers mix with various solvents. Measurement of these properties and determination of the exact means by which they influence film quality is complex and even controversial. Many difficult questions remain to be answered.

Paint properties. A paint should be applicable while liquid; on application all surface irregularities should smooth out, and paint film should flow into cracks and crevices before it develops appreciable rigidity. If a fluid condition is maintained too long, the coating will sag on vertical surfaces; consequently, most paints are so formulated that the time of hardening is kept short. This is done by adding what are called thixotropic agents to the paint that cause the binder to lose its structural rigidity as it is brushed onto the surface and to regain it shortly thereafter.

PAINT APPLICATION

Coatings have been applied for decades by brushing, rolling, spraying, and scraping. Techniques of deposition by means of an electrical charge and fluidized beds (finely divided pigments made to act as a fluid by passing air through them) recently have been introduced on a commercial scale.

Surface preparation. Wood, masonry, and other porous substrates must be treated to develop a surface capable of permanent and strong adhesion of the coating. Old loose paint and nonadherent weathering products, chalk, and dirt must be removed by abrasion. Metal surfaces require removal of scale and grease or pretreatment with phosphoric acid solutions.

Combating corrosion

The two main reasons for careful preparation of a surface are to maintain adhesion and combat corrosion. Corrosion of metals cannot be stopped simply by covering the surface with a thin film; corrosion is electrochemical in nature and can arise from irregularities in composition, aggravated by the slow diffusion of water and oxidants through the film. Thus, for complete protection, a first, or primer, coat is needed, followed by a second, or finishing, coat.

Application methods. Brushing and rolling are two very common methods of application. Brush marks are not simply indentations made by the bristles but arise from the unstable flow pattern of paint from brush to wall in addition to a profile produced by groups of bristles and substrate irregularities.

A common method for the application of primers, topcoat enamels, and lacquers is spraying, which minimizes the stress on the primer coat, formulated more for its chemical than for its mechanical properties. Whereas brushing creates a sizable stress on the undercoat, spraying largely avoids stresses at the interface.

In coating an object by spraying, much of the coating material may be lost by missing the target. One arrangement designed to overcome this waste is the provision of a high-voltage electrostatic field as the workpiece and then inducing a charge in the spray; its particles are then attracted to the workpiece.

Small objects are coated by a slow-dip process characterized by drying of the lacquer to a stable film as rapidly as the object is withdrawn. This procedure avoids the drainage problem that leads to sagging, and it is particularly suitable for the application of thick coatings. Curing may require subsequent baking.

Fast-dip application is used in coating large objects such as machinery. The piece is withdrawn and hung on a rack to drain and dry. In this version of dipping, the solvent is of a slow-evaporating nature, and the coating formulation has a low viscosity.

Electrophoretic coating

Electrophoretic coating, which involves movement of suspended particles through a fluid under the action of a voltage applied to electrodes in contact with the suspension, and electrocoating, which involves electrical attraction of the coating to the substrate, are modern developments carried out in fast-dip tanks. Charged resin particles are attracted only to bare, uninsulated metallic areas and therefore ensure complete, even coverage.

Flat stock ranging from metallic coils to porous paper can be coated at high speed by equipment that resembles a printing press. Printing of decorative patterns is performed in this manner.

Tumbling is employed in the coating of small objects such as hardware and buttons. The coating forms evenly if the objects are fairly smooth and have no sharp projections.

Fluidized-bed coating provides a solventless way of depositing an even coating of resin on surfaces that can be heated, thereby relying on fusion to provide the levelling and adhesion. A current of air is blown through a bed of resin beads, which impinge on the heated surface lowered into the heated fluid. The hot metal entraps particles that reach its surface. Flame spraying is employed in applying ceramic or refractory coatings to metallic surfaces.

Baking or stoving. The ultimate cure or final drying of many organic coatings is provided by baking. In general, the aim is to have this final step occur evenly throughout the thickness of the film in order to minimize residual stresses that would adversely affect the adhesion. Baking is used to level out buffed coatings to provide extremely high gloss.

Elevated curing temperatures are obtained by convective heating and infrared heat radiation. Some attempts have been made to use induction heating, and this method is likely to grow in popularity. Radiation is occasionally provided by a beam of electrons or by ultraviolet radiation.

ECONOMICS

The paint industry of the world enjoyed a steady growth through the 1960s and early 1970s, notably in western Europe, Japan, and the United States. The number of paint producers in any country is typically large because of the competitive advantage enjoyed by producers in the immediate geographic area in which the paint is marketed and used. Demand is met by many small producers as well as by branches of large companies; the consumers' requirements for small specialty batches and his call upon technical service staffs are based on many subjective factors requiring local contacts. It is estimated that 1,700 companies make paint in the United States; 480 in the United Kingdom; 350 each in Italy and France; 200 in West Germany; and 100 in The Netherlands. The bulk of the sales volume, however, is handled nearly everywhere by a few large manufacturers. Three manufacturers dominate the United Kingdom, and 15 firms in the United States account for 50 percent of the sales.

Half of the paint market involves industrial firms, that sell the coating as an integral part of the finished product, and only half involves over-the-counter sales to final users.

Nevertheless, the small paint manufacturer exerts a profound effect on the overall industry; it is common for the leading paint company in any given metropolitan area to be a local firm. To compete with the research efforts of the large companies, small companies often band together to form cooperative research groups.

Over one-third of the cost of paint is the cost of resins and other chemical raw materials; another third is for manufacturing costs and the remainder for packaging and marketing.

Technical societies

Technical developments are disseminated largely by federations of local paint societies made up of individual rather than corporate memberships. The Oil and Colour Chemist's Association in London, the Federation of Scandinavian Paint and Varnish Technicians, and the Federation of Societies for Paint Technology (FSPT) in Philadelphia are continuously active. The countries of western Europe and North America hold a biennial conference under the auspices of a loose international federation known as Fatipec (Federation of Associations of Technicians in the Paint, Varnish, and Printing Ink Industries of Continental Europe).

FUTURE OF THE COATINGS INDUSTRY

The surface-coating industry promises to continue to grow in the future as a result of overall world economic growth, technology that will widen the application of surface coatings, and finally, the important role foreseeable for paints, varnishes and allied products in the struggle to improve the environment.

Electrocoating methods and coil coating (described above) will grow, and curing methods requiring high energy radiation will be the preferred means of developing desirable mechanical properties within a brief time span after application. Powder coatings are on the threshold. New resin systems will be developed.

In order to minimize pollution and safety hazards, the trade will turn more to aqueous (water-base) systems. The chemistry of film formation from latexes will be thoroughly probed so that the limitations of latex application will be understood.

Paints will perform more varied functions than decoration and protection against degradation. Sacrificial coatings that protect surfaces from fire by swelling and bubbling are expected to double in volume; their use may be required for certain types of installations. Conductive coatings will serve a variety of uses ranging from panel heating to undercoats for electrodeposited finishes. Coatings that change hue on demand will provide a new dimension in decoration. Sound-absorbing coatings may be developed.

One of the most rapidly growing markets for coatings is floor coverings. Having lost much of the market for ceilings, the paint industry has turned to seamless flooring, a liquid system possessing decorative, high-build, and abrasion-resistant qualities. Because of the need for high-impact resistance, seamless flooring must be applied at least 35 mils (35 thousandths of an inch) thick.

Biodegradable plastics, which break down when exposed to sunlight, then decompose like natural substances, may serve as protective coatings for base metals or other degradable stock. After use, the article will be exposed to sunlight, which will destroy the integrity of the coating and allow degradation of the substrate to occur.

Pollution control

As a derivative industry dealing with conversions of raw materials into finished products, the paint industry is concerned more with pollution by its product and related hazards associated with coatings than with pollution during the process of paint manufacture. Research has already been brought to bear effectively on the solution of the production problem. Pollution by the product after application is more difficult to estimate or even to identify. Lead has been removed from coating formulations that are liable to come within the reach of children; it may come under more sweeping prohibition. Mercury salts were the best fungicides for preventing mildew growth on surfaces until that metal was ruled out of house paint.

Hiding without pigments is an attainable goal. The bloom on fruit, the opacity of flower petals, and blush films all result from the scattering of light at the interface of minute voids in the film. The low refractive index of such voids provides an optical mismatch with the film, as does high-index pigment, so that hiding is just as effective with voids as with solid pigments. Inferior mechanical properties of pigmentless paints, particularly as manifested by loss of hiding power when voids are crushed, must be overcome if these materials are to make deep inroads into paint markets.

Industry-wide basic research is supported by the Paint Research Institute of the FSPT. Through its program of awarding fellowships in chemistry, physics, and engineering the institute maintains contact with the latest developments in colloid and macromolecular science.

BIBLIOGRAPHY. R.R. MYERS and J.S. LONG, *Treatise on Coatings*, vol. 1, *Film-Forming Compositions*, 3 pt. (1967–71); successive volumes covering *Characterization*, *Vehicles*, and *Pigments*, offer the most comprehensive coverage of coatings since J.J. MATTIELLO (ed.), *Protective and Decorative Coatings*, 4 vol. (1941–42). The best monographs on coatings are D.H. PARKER, *Principles of Surface Coatings Technology* (1965); AND P. NYLEN and E. SUNDERLAND, *Modern Surface Coatings* (1965). See also the *Federation Series of Coatings Technology*, of the Federation of Societies of Paint Technology, Philadelphia, consisting of 20 pamphlets on basic subjects by various authors; and A.K. DOOLITTLE, *The Technology of Solvents and Plasticizers* (1954), for background. A *Raw Materials Index* is compiled by the National Paint, Varnish and Lacquer Association, Washington, D.C.; and *Physical and Chemical Examination of Paints, Varnishes, Lacquers and Colors* is produced by the H.A. Gardner Laboratory, Inc. in Bethesda, Maryland. Journals devoted to coatings science and technology include: *The Journal of Paint Technology*, Federation of Societies for Paint Technology, Philadelphia, Pennsylvania; JOCCA, the journal of the Oil and Colour Chemists' Association, London; *Farg och Lack*, Skandinavisk Tidskrift för Medlemsblad för Skandinaviska Lackteknikers Forbund (in Swedish), Copenhagen, Denmark; and an international journal entitled *Progress in Organic Coatings*, by Elsevier Sequoia S.A., containing review articles.

(R.R.M.)

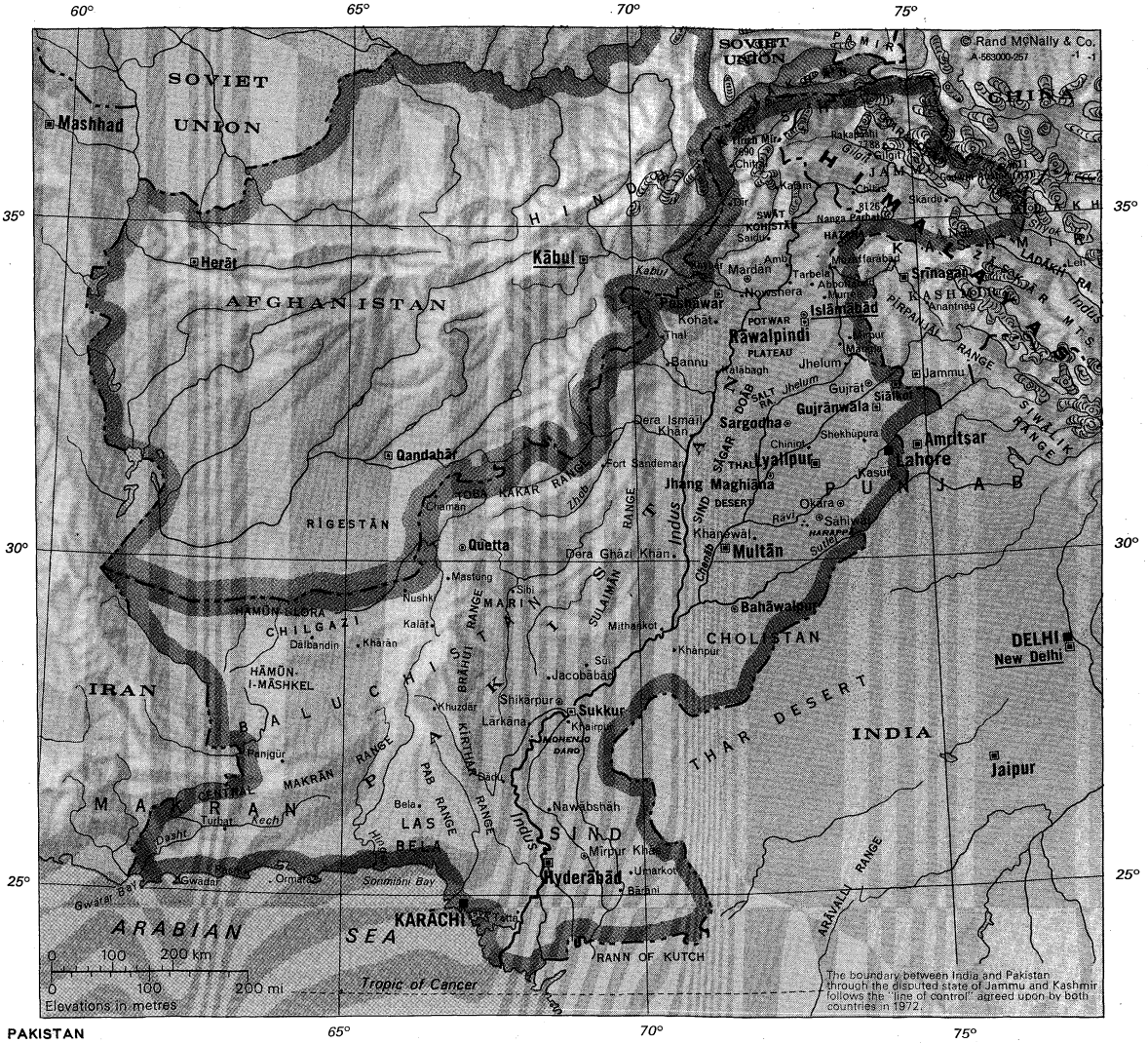
Pakistan

Pakistan (the Islāmic Republic of Pakistan) is a country in South Asia. It is bounded to the west by Iran, to the north by Afghanistan and the U.S.S.R., to the northeast by China, to the east and southeast by India, and to the south by the Arabian Sea. The territory has an area (excluding the Pakistani-held part of Jammu and Kashmir) of 307,374 square miles (796,095 square kilometres) and at the 1972 census had a population of 64,892,000. The capital of the country is Islāmābād.

Pakistan was brought into being at the time of the Partition of British India in 1947 in order to create a separate homeland for India's Muslims in response to the demands of Muslim nationalists. From independence in 1947 until 1971, Pakistan (both de facto and in law) consisted of two regions—West Pakistan in the Indus River Basin, and East Pakistan, located more than 1,000 miles away in the Ganges River Delta. In response to grave internal political problems, however, an independent state of Bangladesh was proclaimed in East Pakistan in 1971.

Economic development has been concentrated in two of Pakistan's four provinces—Punjab and Sind—which, with about 75 percent of the rural population, produce more than 90 percent of the country's wheat, more than 95 percent of its rice, and 100 percent of its cotton. The remaining two provinces—Baluchistan and the Northwest Frontier Province—are, by comparison, poor, for industrial development is also largely concentrated in Punjab and Sind.

Throughout the republic of Pakistan there is a high degree of concentration of landownership, although successive governments have pledged themselves to land reform. There is an even greater concentration of ownership of industry among a few. (See the city articles on KARACHI and LAHORE, as well as the articles on Pakistan's four provinces, BALUCHISTAN; NORTHWEST FRONTIER PROVINCE; PUNJAB [PAKISTAN]; and SIND; see also the articles on BANGLADESH [formerly East Pakistan] and on JAM-



PAKISTAN

MAP INDEX

MU AND KASHMIR. For associated physical features, see ARABIAN SEA; HIMALAYAN MOUNTAIN RANGES; HINDU KUSH [MOUNTAINS]; INDUS RIVER; KARAKORAM RANGE; SALT RANGE; and THAR DESERT. For historical aspects, see INDIAN SUBCONTINENT, HISTORY OF THE.)

This article is divided into the following sections:

I. Landscape and environment

RELIEF AND DRAINAGE

Pakistan is situated at the western end of the Indo-Gangetic Plain, which is bounded to the north by the mountain wall of the Great Himalayan mountain ranges and their offshoots. It is situated in the northwestern part of the southern Asian subcontinent and may be subdivided into six natural regions—the northern mountains, the submontane plateau, the Indus Plain, the Baluchistan Plateau, the western bordering mountains, and the desert areas.

The Himalayan sector

The northern mountains. The Himalayan and Trans-Himalayan mountains occupy the entire northern end of Pakistan, to a depth of about 200 miles; they are among the youngest mountains on Earth, having attained their present elevation only within the last 1,000,000 years. They rise to an average height of more than 20,000 feet and include such towering peaks as Nānga Parbat (26,660 feet, or 8,126 metres, high) and K2, also called Godwin Austen (28,251 feet, or 8,611 metres, high). Beyond the Karakoram Range in the extreme north lies the Chinese province of Sinkiang; to the northwest, beyond the Hindu Kush, are the Pamirs—the "Roof of the World"—where only a narrow strip of Afghan territory separates Pakistan from the Soviet Union. The Himalayan massif (mountainous mass) that isolates the South Asian subcontinent from Soviet and Chinese Central Asia was pierced in 1970 when a road was completed across the Karakoram Range, linking the town of Gilgit in Pakistan with Kashgar in Sinkiang, China.

The northern mountain barrier influences the rainfall pattern in Pakistan by intercepting monsoon (rain-bearing) winds from the south. Melting snow from the mountains also feeds rivers, including the mighty Indus, which cut across the east–west aligned mountain ranges to flow southward.

The population in this inhospitable region is generally sparse, although in a few favoured places it is dense. In most of the tiny settlements of this region, the usual crop is barley; fruit culture, especially of apricots, is of special importance. Timber, mainly pine, is found in some parts, but its occurrence varies with rainfall and altitude. Many slopes have been denuded of cover by excessive timber felling and overgrazing. The whole region in itself is of less significance than in its relation to the plains to the south.

The submontane plateau. Lying below the Himalayas, the submontane plateau has four distinct divisions—the Trans-Indus plains, the Potwar Plateau, the Salt Range, and Siālkot district.

The Trans-Indus plains. The Trans-Indus plains, west of the Indus, comprise the hill-girt plateaus of the Vale of Peshāwar, Kohāt, and Bannu, which are oases in the arid, scrub-covered landscape of the Northwest Frontier Province. Of these, the Vale of Peshāwar (consisting of Peshāwar and Mardān districts) is the most fertile. It was once a flourishing centre of Greco-Buddhist culture; with a territory of only 2,500 square miles, it has nearly half of the population of the entire Northwest Frontier Province, which has a total area of 39,000 square miles. Much of the area is covered by gravelly or clayey alluvial detritus. Rainfall is generally only between 10 and 15 inches but is higher in Mardān district. About 65 percent of the net sown area in Peshāwar Vale is irrigated from canals. Kohāt is, comparatively, little developed. Rainfall is about 16 inches. Only 12 percent of the sown area is canal irrigated. Nor is its groundwater adequately exploited, although the water table is generally high. Much of the area consists of scrub and poor grazing land. The region is much broken by limestone ridges, and the uneven limestone floor is variously filled with lacustrine clays, gravel, or boulders. In Bannu, soils are in general sandy or gravelly, except for rich silts that occur in places. About 23 percent of the net sown area is irrigated. Rainfall is low, amounting to about 11 inches. In both Kohāt and Bannu fat-tailed sheep, camels, and donkeys are extensively reared; wool is an important cash crop.

The Vale of Peshāwar

The Potwar Plateau. The Potwar Plateau lies at a height of 1,200 to 1,900 feet and covers an area of about 5,000 square miles east of the Indus, in the Punjab. It is an open, undulating country developed on the Siwālik Range, which is mainly of sandstone and is covered by varying thicknesses of loess (a loamy deposit formed by the wind), which erodes easily. Rainfall is between 15 and 20 inches (380 to 510 millimetres), being higher in the northwest; the southwest is very arid. The landscape is dissected and eroded by streams that, during the rains, cut into the land and wash away the soil. The streams are generally deep set and are of little or no use for irrigation. It is generally a poor agricultural area, with an excessive pressure of population upon its resources.

The Salt Range. The Salt Range lies at the southern edge of the Potwar Plateau and has an average height of 2,200 feet, the highest peak being at Sakesar (5,000 feet, or 1,500 metres); it is an extremely arid territory that sharply marks the boundary between the submontane region and the Indus Plain to the south. The Salt Range is of interest to geologists, as it contains the most complete

geologic sequence in the world, in which rocks from the Cambrian Period (from 500,000,000 to 570,000,000 years old) to the Pleistocene Epoch (from 10,000 to 2,500,000 years old) are represented.

Siālkot district. Siālkot district is a narrow submontane region in the northeast; unlike the Potwar Plateau, it is a rich agricultural region. Rainfall varies from 25 to 35 inches (650 to 900 millimetres), and the water table is high, facilitating well (and tube-well) irrigation; the soil is heavy and very fertile. Population is dense (about 1,100 per square mile); the land is divided into small farms on which intensive cultivation is practiced.

The Indus Plain. The Indus Plain, which covers an area of about 200,000 square miles, is the most prosperous agricultural region in Pakistan. It is an unrelieved featureless plain of fertile alluvium extending for 650 to 700 miles from the rim of the Potwar Plateau southward to the Arabian Sea. Its northern zone comprises the province of Punjab (literally, Five Waters) and is enclosed by the Indus and its five tributaries—the tributaries converging to their confluence with the Indus at Mithankot. It is divided into several interfluves (lands situated between streams) called *doābs*, the largest but the poorest of them being the Sind Sāgar Doāb, situated between the Indus and the Chenāb rivers, which is mostly a desert. The *doābs* that lie to the east of it, however, are the richest agricultural lands in the country. Until the advent of irrigation, at the end of the 19th century, much of the area was a desolate waste, for there is little rainfall. But irrigation has not been an unmixed blessing; it has also caused waterlogging and salinity in some places.

The *doābs*

In the southern zone of the Indus Plain lies the province of Sind, which is named after Sindhu (the Indus River). There, the size of the Indus—its waters augmented by the contribution of its five tributaries—is very great; its average annual discharge at Sukkur, in northern Sind, is more than 5,000,000,000,000 cubic feet (140,000,000,000 cubic metres), including nearly 10,000,000,000 cubic feet of water-borne silt. The land in Sind, consisting of alluvial sands and clays, tends to give way before the Indus flood, and the river frequently changes course. Manchhar, a marshy lake west of the Indus, has an area of 14 square miles at low water but extends for no less than 200 square miles when full; on such occasions, it is the largest freshwater lake in South Asia. The quality of groundwater in the Indus Plain varies, that in the southern zone (Sind) being mostly saline and unfit for agricultural use. Extensive areas in both the northern and southern zones of the Indus Plain have been affected by waterlogging and salinity. In the south the Indus Delta (in marked contrast to the Ganges Delta) is a wild waste. When high tides and Indus floods coincide, the littoral is flooded for 20 miles inland.

The Baluchistan Plateau. The Baluchistan Plateau extends westward, 1,000 feet high, with many ridges running across it from northeast to southwest; it is separated from the Indus Plain by the Sulaimān and Kīrthar ranges. It has a remarkable indigenous method of irrigation called the *kārez*, which consists of underground channels and galleries that collect subsoil water at the foot of hills and carry it to the fields and villages. The water is drawn from the channels through shafts that are sunk into the ground, at suitable intervals, in the fields. Because the channels are under ground, the loss of water by evaporation is minimized. The plateau is an extremely arid country and is the most sparsely populated region in Pakistan, with an average density of nine persons per square mile. Pastoral activity supplements a primitive form of agriculture. True pastoral nomadism survives in the northwest. Goats and fat-tailed sheep account for about 80 percent of the stock, and much of the local traffic consists of camels and donkeys, although trucks and buses are in use on the new roads.

The western bordering mountains. These mountains run south from the Hindu Kush, in several parallel ranges, outside the path of the monsoons. Rainfall is low, and there is little vegetation. Occasional heavy rain in the mountains results in destructive torrents, washing away the soil. Some cultivation is practiced beside streams and along riverbanks or (wherever conditions permit) by terracing hillsides. Three minor ranges run south from the Hindu Kush to the Kābul River, south of which lies the famous Khyber Pass, at the frontier with Afghanistan. Further south the Sulaimān Range runs southward for about 300 miles, after which the lower Kīrthar Range runs down to the coast. These hills separate the Indus Plain from Baluchistan. The population of these arid hills is tribal. The area in general is extremely poor; only a minimum of crops are grown in favoured valleys—which, however, are generally malarial. Large numbers of sheep and goats are kept. The mule is the main beast of burden in this broken country.

The desert areas. The desert areas include firstly the dreary steppes of the Sind Sāgar Doāb (at the centre of it the Thal, which has true desert conditions) and secondly Cholistan in Bahāwalpur (Punjab), which is known as the Nāra or Registān in Khairpur (Sind) and further south as the Thar Desert in the Thar Parkar District of Sind. These areas are extensions of the Thar Desert of western India.

SOILS AND CLIMATE

Soils. The soils of Pakistan are classified as pedocals, which comprise a dry soil group having a high content of calcium carbonate and a low content of organic matter; they are characteristic of a land with low and erratic rainfall. The soils of the Indus Plain consist of alluvium, old and new, whereas in the desert areas and most of Baluchistan they consist either of windblown deposits known as loess or else of loess mixed with alluvium. The soils of the Thar Desert area are, for the most part, sandy loams with patches of clay and larger areas of pure sand. Some of the virgin soils respond generously to irrigation, and an extensive reclamation scheme has made much progress in the Thal region.

Climate. As Pakistan is located on a great landmass, north of the Tropic of Cancer (between latitudes 24° and 37° N), it has a continental type of climate characterized by extreme variations of temperature, both seasonally and daily. Very high altitudes modify the climate in the cold snow-covered northern mountains; temperatures on the Baluchistan Plateau are somewhat higher. Along the coastal strip, the climate is modified by sea breezes. In the rest of the country, temperatures reach great heights in the summer; the highest temperature in the entire Indian subcontinent—126° F (52° C)—was recorded at Jacobābād, in Sind. In the summer, hot winds called *loo* blow across the plains during the day. Trees shed their leaves to avoid loss of moisture. The dry, hot weather is broken occasionally by dust storms and thunderstorms that temporarily lower the temperature. Evenings are cool; the diurnal variation in temperature may be as much as 20° to 30° F (11° to 17° C). Winters are cold, with minimum mean temperatures of about 40° F (4° C) in January.

Summer and winter temperatures

The characteristics of different regions are determined by variations in rainfall and irrigation rather than by temperature. Although the country is dominated by monsoon winds, it is extremely arid, except for the southern slopes of the Himalayas and the submontane tract, which have a rainfall of from 30 to 35 inches (750 to 900 millimetres). The 20-inch (500-centimetre) precipitation line, which runs northwest from near Lahore, marks off the Potwar Plateau and a part of the Indus Plain in the northeast; these areas receive enough rainfall for dry cropping (farming without irrigation by taking measures to conserve water). South of this region, cultivation was confined mainly to riverine strips until the advent of irrigation.

VEGETATION AND ANIMAL LIFE

Vegetation. Natural vegetation, except for wooded mountain slopes, is largely limited to tough wiry grass and stunted bushes, with only a few scattered trees, except for a few plantations in forest reservations and ubiquitous orchards of fruit trees. In Baluchistan the Salt Range and the western-bordering-mountains vegetation is mostly limited to xerophytic plants (plants adapted for growth under dry conditions). The deserts are virtually devoid of

vegetation. The Indus Plain when watered, however, is very fertile. Wheat is the principal winter crop, the main summer crops being cotton and rice. Widespread installation of tube wells, especially in the Punjab, is bringing about substantial changes in cropping patterns, encouraging substitution of wheat for less valuable drought-resisting crops, such as gram (chickpea).

Animal life. Wildlife abounds in the northern mountains and includes brown bear, black Himalayan bear, leopard, the rare snow leopard, Siberian ibex, and varieties of wild sheep. Buffalo rather than cows are kept as milch cattle, and bullocks are employed for plowing, although occasionally camels are also used for that purpose. Donkeys are the beasts of burden. The Manchhar Lake in Sind has an abundance of water birds, including mallards, teals, shovellers, spot bills, geese, pochards, and wood ducks; desert gazelles are also found in the region. Crocodiles, pythons, and wild boars are found in the delta area. Jackals, foxes, wild cats, and a variety of rodents and reptiles are found throughout the country.

TRADITIONAL REGIONS

The traditional regions of Pakistan, shaped by ecological factors and historical evolution, are reflected in the administrative division of country into the four provinces of Sind, Punjab, the Northwest Frontier Province, and Baluchistan. The encroachment of the Cholistan Desert and of the western bordering mountains narrows down the Indus Plain below Mithankot, thus separating the destinies of Sind and Punjab, each of which has a distinct language, culture, and history. Until 1937 Sind was a part of the Bombay Presidency, and it had no administrative links with the Punjab. Baluchistan and the Northwest Frontier Province are separate, both linguistically and culturally; they are distinguished by tribal societies subsisting on precarious agriculture, supplemented by pastoralism. In the Punjab, until the advent of irrigation, the bulk of the population was restricted to the *bārānī* ("land dependent on rainfall for cultivation") region—the Potwar Plateau and the Upper Indus Plain—which receives more than 20 inches of rainfall.

The Canal Colony districts of southern Punjab were large tracts of uncultivated land in the Indus Plain that were irrigated by canals and populated by colonists drawn from every part of the Punjab. They now form the richest agricultural region of the country. Agricultural wealth is concentrated in districts of Lahore Division (in the *bārānī* region) that have benefitted from irrigation, together with the Canal Colony districts and Sind. These areas contain 65 percent of the rural population of Pakistan and produce well over three-quarters of its wheat production and virtually all of its cotton and rice production. Landholdings are comparatively larger in the Canal Colony districts of the Punjab and in Sind.

Land-holding patterns

Elsewhere, in the overpopulated and poor districts of the *bārānī* region, which moreover do not benefit from irrigation, holdings are exceedingly small and fragmented. In these districts there is a great pressure to migrate from the villages to find employment in towns or in the armed forces. Consequently, unlike the richer agricultural areas, this impoverished and overpopulated region maintains a close symbiotic relationship with urban economy and society. Educational levels in these districts are higher than they are in the richer agricultural areas.

THE LANDSCAPE UNDER HUMAN SETTLEMENT

Rural settlement. The vast majority of the rural population of Pakistan lives in nucleated villages or hamlets (*i.e.*, in compact groups of dwellings). Sometimes, as is generally the case in the Northwest Frontier Province, the houses are placed in a ring with blank outer walls, so that each complex resembles a protected fortress with a few guarded entrances. Dispersed habitation patterns in the form of isolated single homesteads are rare, occurring only in a few mountainous areas. But it is not uncommon to find numerous satellite hamlets of varying sizes near larger villages; such hamlets are occupied either by a landlord, with his servants and sharecroppers, or else by members of a lineage living together in adjoining houses. The spread of tube wells in the Punjab has increased the tendency for such dispersal, for people often prefer to live near their tube wells in order to guard the valuable machinery. The concept of village, therefore, often tends to be equivalent to that of the *mawzaʿ* (an area of land that, together with a village and its satellite hamlets, forms a unit in land-revenue records). It is difficult to speak of an average size of village, for patterns of habitation are complex. Most groups of dwellings have a minimum of a dozen or a score of houses, and there are usually a few hundred dwellings in each "village." Large villages rarely have a population exceeding 2,500 persons.

Three basic types of village layout are to be found. Most of the older villages are of the "spider-web" form, having at least one focal point, such as the village mosque, some shops, or a well from which lanes radiate. A few villages follow the contours of hill slopes and other natural features. Finally, in the Canal Colony areas, villages are of a regular rectangular pattern, with a well, a mosque, and a school, as well as the house of the village headman, at the centre, and with the houses being arranged in a series of concentric rectangles. Houses are built from available local materials; the vast majority are of mud, a material that is not only cheap and reasonably durable in the dry climate but also provides better insulation from extremes of heat and cold than brick or stone. Houses usually have walled courtyards where animals are tethered and where people sleep in the open in the hot summer.

Urban settlement. The urban population of Pakistan represents about a quarter of the total. Two cities have a dominating position—Karāchi, which in 1961 had nearly a fifth of the urban population, and Lahore, which had 13 percent of it. By 1972 there were 18 cities with a population of more than 100,000, compared with seven such cities in 1941. During the 1960s government policy was directed toward the dispersal of industry, which had become heavily concentrated in Karāchi. As a consequence, since the 1961 census, urban growth has been more evenly distributed among several cities. Karāchi, the principal port and centre of commerce and industry, was also the national capital until 1966, when the new capital at Islāmābād, 750 miles away, began to function.

Rapid and unplanned urban expansion has been paralleled by a deterioration in living conditions, particularly in the housing conditions of lower income groups. According to one official estimate in 1969, about 72 percent of urban households are unable to pay rent for the cheapest form of available housing and live in makeshift shacks. Water supply and sewerage systems are inadequate, and in many areas residents have to share communal water taps. Inadequate urban transport is also a major problem. After Karāchi and Lahore, the principal cities (in order of size as of the 1972 census) were Lyallpur, Hyderābād, Rāwalpindi, Multān, Gujrānwāla, Peshāwar, Siālkot, Sargodha, Sukkur, Quetta, Jhang, Bāhāwalpur, Sahiwal, Mardān, Wah, and Kasur. Karāchi is the provincial capital of Sind, Lahore of Punjab, Peshāwar of the Northwest Frontier Province, and Quetta of Baluchistan. The new capital city of Islāmābād adjoins Rāwalpindi, where the army's general headquarters and the offices of the president of Pakistan are located.

Urban living conditions

II. People and population

ETHNIC GROUPS

Race as such plays little part in defining regional or group identity in Pakistan, and no ideal racial type is accepted by all Pakistanis. The population is a complex mixture of indigenous peoples, many racial types having been introduced by successive waves of migrations from the northwest, as well as by internal migrations across the subcontinent of India. Aryans, Persians, Greeks, Pashtuns (Pathans), and Mughals came from the northwest and spread across the Indo-Gangetic Plain, while the Arabs conquered Sind. All left their mark on the population and culture of the land. During the long period of Muslim rule, immigrants from the Middle East were brought in and installed as members of the ruling oligarchy. It became prestigious to claim descent from them, and many

members of the landed gentry and of upper class families are either actually or putatively descended from such immigrants. In 1947, when Pakistan and India became independent, there was another massive migration, of a different character, when millions of Muslim refugees were uprooted from different parts of India and settled in Pakistan; an equal number of Hindus were uprooted from Pakistan and driven across to India. This development further complicated the racial mixture of the population of the various regions of Pakistan.

Ethnic antecedents

Of the earlier people, almost the only known skeletal remains of much significance are those associated with the Indus civilization that flourished 5,000 years ago. Human skeletons and skulls found among the ruins at Mohenjo-daro in Sind reveal firstly that at least four racial types were present in that area in about 2500 BC—Mediterranean and Alpinoid (both Europoid), Proto-Australoid (Veddoid), and Mongoloid—and secondly that there was much racial mixture. The Indo-Aryans, who predominate in contemporary Pakistan, arrived between about 1500 and 1200 BC. In the 6th century BC the Persian Empire was extended to the Indus Valley, and in the 4th century BC Alexander the Great and his armies passed through the region. The region subsequently became a melting pot for diverse races and cultures that came not only from outside the Indian subcontinent but also as a result of internal movements, such as the expansion of the Mauryan Empire over northern India. These variegated influences and associated migrations render any attempt to identify racial origins of people in different parts of Pakistan somewhat arbitrary.

Racial types to be seen in Pakistan include the tall, fair-skinned, and blue-eyed type; the olive-skinned, fine-boned, hawk-nosed "Iranian" type; smaller, dark-skinned types of "Dravidian" and "Australoid" origin; wheaten-skinned, dark-eyed "Indo-Aryan" types; short-headed and long-headed Mongoloid peoples; and, in Baluchistan, the broad-headed Europoid type of stocky build, reminiscent of the Alpine type in Europe.

Regionally, the population is sometimes grouped into four types by racial origin as follows—(1) the "true" Mediterranean type, found in the Punjab; (2) the "Oriental" Mediterranean type, found in the Punjab and Sind, characterized by an unusually long and often convex nose and lighter skin colour but otherwise similar to the first type; (3) the Pashtuns of the Northwest Frontier Province; (4) the brachycephalic (short-headed) Baluchi type, which derives from Iranian stock.

Attitudes to racial types vary. In Baluchistan, the short Iranian head is admired, so that the three major tribes of the region sometimes bind their daughters' heads. In other areas, longheadedness is valued. Elsewhere, again, the shape of the head is a matter of indifference. There is, however, a general prejudice in favour of fair rather than dark skin colour.

LINGUISTIC GROUPS

Pakistan is in general linguistically heterogeneous, and no single language can be said to be common to the whole population. Each of its principal languages has a strong regional focus, although statistics show some languages to be distributed between various provinces because administrative boundaries cut across linguistic regions. The picture is also complicated by the fact that, especially in Sind, there are substantial numbers of Urdu- and Punjabi-speaking immigrants who have moved from the place of their origin and have settled on the land or taken up urban employment. The distribution of languages claimed as mother tongue at the time of the 1961 census was as follows: Punjabi, 66 percent; Sindhi, 13 percent; Pashto, 8 percent; Urdu, 8 percent; Baluchi, 2 percent; and Brahui, 1 percent.

The language distribution pattern

The provincial distribution of languages spoken (whether as mother tongue or as an additional language) was as follows: in Punjab, Punjabi is the principal language, although Urdu is also spoken by about 16 percent of the people, though of these only one-third speak it as their mother tongue; in Sind (excluding Karāchi), about 80 percent speak Sindhi, 14 percent Urdu, and 10 percent Baluchi; in the Northwest Frontier Province, about 84 percent speak Pashto and about 23 percent Punjabi; in Baluchistan, 41 percent speak Baluchi, 25 percent Pashto, and 20 percent Sindhi. In Karāchi, about 68 percent speak Urdu, 17 percent Sindhi, about 9 percent Punjabi, about 9 percent Baluchi, and about 9 percent English.

The languages of Pakistan are written in modified forms of the Arabic–Persian script, which is written from right to left. Punjabi has its own script called Gurmukhi, but this is mainly used in India. In Pakistan, Punjabi is mainly spoken rather than written; it is also a predominantly rural rather than an urban language. Urdu, rather than Punjabi, is the first language taught in schools in Punjab, so that every educated Punjabi reads and writes Urdu. There is, however, a movement for the promotion of the Punjabi language, and some Punjabi literature is being published using the Urdu script; among the works published are Punjabi classics that have hitherto been available in Gurmukhi script or preserved in oral tradition.

Sindhi is derived from the Virachada dialect of Prākrit; it has fewer dialects than Punjabi. It is written in a special variant of the Arabic script. Most of the educated middle class in Sind were Hindu, and their departure to India in 1947 had a traumatic effect on Sindhi culture. Vigorous efforts are therefore directed toward a recovery and preservation of the rich Sindhi literary and cultural heritage. Large numbers of Urdu-speaking refugees have settled down in Sind and constitute the majority of the population of its larger towns. As a consequence, the movement for promotion of Sindhi language and culture sometimes finds expression in hostility toward Urdu. Unlike Punjabis, who learn Urdu at school, in Sind, Sindhi children are taught Sindhi as their first language.

Pashto

Pashto, the language of the Pashtuns (Pakhtuns or Pathans) of the Northwest Frontier Province, has no written literary traditions although it has a rich oral tradition that is being put into a written form using a modified version of the Persian script. There are two major dialect patterns within which the various individual dialects may be classified; these are Pakhto, which is the northern (Peshāwar) variety, and the softer Pashto spoken in southern areas. As in the Punjab, Urdu is the language taught in schools, and educated Pashtuns read and write Urdu. Again, as in the case of Punjabi, there is a movement for developing the written language and increasing the usage of Pashto.

The two main spoken languages of Baluchistan are Baluchi and Brahui. Makrani is an important dialect of Baluchi; it is spoken in Makran, the southern region of Baluchistan, bordering on Iran.

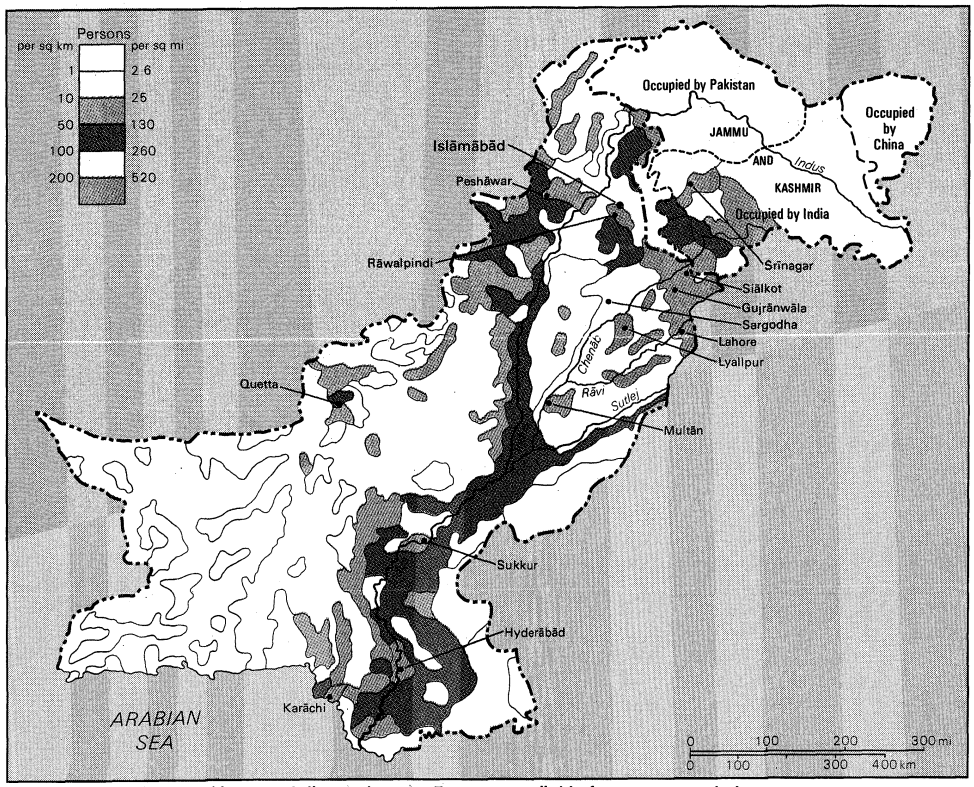
Urdu is the youngest of the nation's languages, and is not indigenous to Pakistan; it is very similar to Hindi, the official language of India. Although the two languages have a common base, in its literary form Urdu emphasizes words of Persian and Arabic origin, whereas Hindi emphasizes words of Sanskrit origin. Urdu is written in a modified version of the Persian script (written from right to left), whereas Hindi is written in Devānagari script from left to right. Because it is pre-eminently the language of the educated Muslims of northern India, including the Punjab, Urdu has strong associations with Muslim nationalism; hence the ideological significance of Urdu in Pakistan politics. Urdu was the mother tongue of only about 7.6 percent of the population of Pakistan in 1961; it was a minority language in every province.

English is used for official purposes except in local administration, where the local vernacular is used. English is the de facto official language; the 1956 Constitution prescribed its use for 20 years, and the 1962 Constitution made the period indefinite. It is spoken by only about 2 percent of the people of Pakistan.

RELIGIOUS GROUPS

About 97 percent of the people of Pakistan are Muslims. Most of them belong to the Sunnī, the major branch of Islām, with a significant representation among the small but equally important Shīʿī branch. There is also a very small, though influential, sect called Ahmadis, or Qadianis, which does not regard the prophet Muḥammad as

Muslim sects



Population density of Pakistan and disputed areas. Data not available for areas occupied by China and Pakistan.

being the final prophet, a basic tenet of Islām. The majority of Pakistani Sunnīs belong to the orthodox Ḥanafī school, which is one of four schools or subsects of Sunnīs. Shīʿites are also divided into numerous subsects; among them are Ismāʿīlīs (the followers of the Aga Khan), as well as The Twelvers (Ithnā ʿAsharīs) and Bohrās, which are prominent communities in commerce and industry. The principal business communities among Sunnīs are Gujarati Memons and Chiniotis from Punjab.

With the exception of some sects, such as Dawoodi Bohrās, there is generally no ordained priesthood among Pakistan's Muslims. Anyone may be appointed *imām* who leads prayers in mosques. Those who are trained in theology are given the title of mullah or *mawlānās* or, collectively, *ʿulamā'*. There are, however, powerful networks of "holy men" called *pīr*s, who receive great reverence, as well as gifts in cash or kind from a multitude of followers. They constitute a hierarchy in one of four Ṣūfī (Muslim mystic) orders. An established *pīr* may pass on his spiritual powers and sanctified authority to one or more of his *murīds* ("disciples"), who may then operate as a *pīr* in his own right. There are also many self-appointed *pīr*s who practice locally without being properly inducted into one of the four orders. *Pīr*s who occupy high positions in the *pīr* hierarchy wield great power and play an influential role in public affairs.

The number of Hindus in Pakistan was greatly reduced as a consequence of the exodus of refugees to India in 1947; in 1961 they formed only 0.5 percent of the population of present-day Pakistan (then West Pakistan). Christians constituted 1.4 percent.

DEMOGRAPHY

The population of Pakistan (then West Pakistan), excluding non-Pakistanis, at the 1961 census was 42,880,000, representing a density of 140 persons per square mile (54 per square kilometre). At the 1972 census the population had reached 64,892,000.

The rate of population growth is phenomenal. Between 1901 and 1961, the population of West Pakistan (as it then was) increased by about 158 percent. From 1951 to 1961, the population increased by 27 percent, and from 1961 to 1972 by 51 percent.

The growth of population was primarily caused by natural increase. As mentioned, the 1947 influx of Muslim refugees from India was balanced by an exodus of Hindu refugees to India. Refugees entering what was then West Pakistan numbered about 6,500,000, or about 20 percent of the population of the territory at that time.

Regional population distribution

Regional distribution of the population is uneven, being dense in the fertile Indus Valley, especially in the extreme northeast. In 1972, it was the highest in Lahore district, with 1,698 persons per square mile, the average for the whole of the Lahore division being 1,086. By contrast, in the vast expanse of Baluchistan, the population densities were only 26 in Quetta division and 15 in Kalāt division.

Estimates of fertility levels (and changes) vary widely and are unreliable. In the early 1970s it was generally believed, however, that the birth rate would remain almost constant for some time, whereas the death rate would decline because of improving health facilities, and that growth in population would persist unless the fertility rate could be brought down drastically. But the achievements of the family-planning program were not yet commensurate with the scale on which it was being operated; by 1968 the family-planning program was be-

Pakistan, Area and Population

	area		population	
	sq mi	sq km	1961 census*	1972 census*†
Provinces				
Baluchistan	134,050	347,188	1,353,000	2,409,000
Northwest Frontier, excluding: Centrally Administered Tribal Areas	28,773	74,522	5,731,000	8,402,000
	10,510	27,221	1,847,000	2,507,000
Punjab	79,284	205,345	25,488,000	37,374,000
Sind	54,407	140,913	8,367,000	13,965,000
Federal Capital Territory				
Islamabad	350	906	94,000	235,000
Total Pakistan‡	307,374	796,095	42,880,000	64,892,000

*De jure. †Preliminary. ‡Excluding portions of Jammu and Kashmir occupied by Pakistan.
Source: Official government figures.

lieved to be the third largest employer in the country, after the military and the railway system.

Urban population represented about 22 percent of the total in 1961; it grew by 60 percent during the decade 1951 to 1961. The trend toward greater urbanization has been accompanied by a relatively faster growth of the larger cities as compared to smaller ones; this has reflected the influence of industrial location as a new factor in the pattern of urban growth. In 1961 in what was then West Pakistan, there were 37 towns and cities with populations of over 25,000, having altogether a total population of 7,000,000. (H.A.A.)

III. The national economy

A COMPARATIVE VIEW

Pakistan, like neighbouring India and Afghanistan, is among the 20 poorest nations in the world. Output per head of population in the early 1970s was less than U.S. $140. The relative prosperity of the industrialized regions around Karāchi and Lahore contrasts sharply with the poverty of the semi-arid Baluchistan province and the Northwest Frontier. The economy still relies heavily on the agricultural sector, mainly cotton. The combination of crop failures because of unfavourable weather and weak trends in world commodity prices was a major problem in the early 1970s. At the same time, there has been a relentless increase in population, so that, notwithstanding a real growth in the economy, output per head rose slowly. Nevertheless, Pakistan's economic performance compares favourably with that of many other developing countries. Real progress has been made in reducing the dependence on agriculture and in the diversification of the manufacturing sector. In this, foreign aid has played an important part. During the 1960s, aid receipts per head of population were high compared with many other developing countries and double those received by India.

The gradual structural change in the economy is reflected in the composition of foreign trade. Although as a trading nation Pakistan is among the ten nations with the lowest imports and exports per head, it is among the ten principal raw cotton exporters. Pakistan also managed to become a leading world exporter of cotton yarn and cloth during the 1960s. Since the separation from India in 1947, Pakistan has evolved from an area supplying raw materials to the processing industries situated in what is now India into a more integrated economy processing and manufacturing its own raw materials both for export and for the home market. Meanwhile, the agreement on Regional Cooperation for Development (RCD) with Iran and Turkey and the trade from aid deals has strengthened ties with the Middle East, Europe, and the United States.

RESOURCES

Mineral resources. The exploration of Pakistan's mineral wealth is far from complete, but by the early 1970s some 20 different types of minerals had been located. Coal mining is one of the country's oldest industries. The quality of the coal is poor, and the mines are working below capacity because of the lack of demand. Estimated reserves are about 400,000,000 metric tons. Iron-ore deposits are also mostly of poor quality. The most extensive known reserves are situated in the Kālābāgh region in western Punjab, amounting to some 300,000,000 metric tons. There are plans to develop these in an effort to set up a local iron-and-steel industry. Other low-grade ore reserves have been found in Hazāra in the Northwest Frontier Province. Small reserves of high-grade iron ore have been identified in Chitrāl and in the Chilghāzi area (located in the Chagai district of Baluchistan province), also in the Northwest Frontier Province. Total known iron-ore reserves in Pakistan are estimated at about 520,000,000 metric tons. There are enormous reserves of easily exploited limestone that form the basis of a growing cement industry. Other minerals that are being exploited include chromite (mostly for export), barite (a white, yellow, or colourless mineral resembling marble), and celestite (strontium sulphate), antimony, aragonite (a mineral resembling calcite [calcium carbonate]), gypsum, rock salt, and marble. Radioactive minerals have been found in the Dera Ghāzi Khān district in the Punjab.

Oil and natural gas

Pakistan also has small quantities of oil and some very large natural-gas fields. The first oil discovery was made in 1915. By the early 1970s, several more fields had been found, none of them very important. The largest natural-gas deposits are at Sūi (on the border between Baluchistan and the Punjab), discovered in 1953. Reserves have been estimated at 6,300,000,000,000 cubic feet (180,000,000,000 cubic metres) and rank among the world's largest. A smaller field, at Māri, in the northwest of Sind province, was found in 1957 with estimated reserves of 3,900,000,000,000 cubic feet. Overall, the region's natural-gas reserves amounted to some 16,000,000,000,000 cubic feet. A network of gas pipelines links the fields with the main consumption areas: Karāchi, Lahore, Multān, Lyallpur, and Islāmābād.

Biological resources. The variety of climates and soils has given rise to a wide diversity in biological resources. As Baluchistan is mostly desert, only in the small areas of intensive cultivation do crops and orchards thrive. In Sind and Punjab, where the annual rainfall is also low, most of the vegetation is basically xerophilous, except for the riverine forests along the Indus and its tributaries. The coastal region has mangrove forests. Regular rain and snow in the Himalayan foothills of the north have given them a variety of vegetation and animal life ranging from the Mediterranean to the Alpine types. There is a fishing industry centred on Karāchi, and part of the lobster and other shellfish catch is exported.

Pakistan's wildlife, already described, is varied. There are no game reserves, and indiscriminate hunting is threatening the survival of several animal species.

Hydroelectric and other power resources. Although Pakistan is poorly endowed with water resources, great progress has been made with the development of its hydroelectric potential. The biggest hydroelectric plant in operation at the beginning of the 1970s was the Mangla Dam on the Jhelum River in Āzād Kashmir, which had a generating capacity of 650,000 kilowatts. This is to be raised first to 1,000,000 kilowatts and then to 3,000,000 kilowatts by 1980. Work on the giant Tarbela Dam on the Indus River started in 1968 and was expected to be completed in 1975; the power plant will have a final capacity of 2,100,000 kilowatts. Notwithstanding the heavy investment in hydropower, most of Pakistan's power needs are met by thermal plants, most of which are coal burning, although some of the newer ones use natural gas. A 137,000-kilowatt nuclear power plant has been built outside Karāchi. By the early 1970s, total electricity generating capacity amounted to about 1,900,000 kilowatts; the share of natural gas in total power consumption was rising steadily as new reserves were discovered and as the distribution system improved.

SOURCES OF NATIONAL INCOME

Agriculture, forestry, and fishing. In 1971–72, agriculture, forestry, and fishing accounted for 41 percent of the gross domestic product. Although the average annual growth rate during the second half of the 1960s was almost 6 percent, compared with 4 percent in 1960–65, its share of total output was declining by the early 1970s. This sector provides employment for at least two-thirds of the official labour force and a livelihood for more than 80 percent of the population. Of the total land area of 199,000,000 acres, only about 25 percent was under cultivation at the beginning of the 1970s. A land-reform program has dealt with the dual problem of large-scale, often absentee ownership and the excessive fragmentation of small holdings by introducing maximum and minimum area limits.

Changes in agricultural techniques

The priority given to the agricultural sector in the development plans during the 1960s brought about some radical changes in centuries-old farming techniques. The construction of tube wells for irrigation and salinity control, the use of chemical fertilizers and scientifically selected seeds, and the gradual introduction of farm machinery—all contributed to the notable increase in pro-

ductivity compared with the 1950s. One of the prime objectives of agricultural development programs was self-sufficiency in wheat; this was so successful that by the early 1970s Pakistan was disposing of surpluses.

During the year 1969 to 1970, agricultural output amounted to 11,000,000 metric tons of food crops, of which 7,000,000 tons were wheat and 2,000,000 tons rice; 26,000,000 tons of sugarcane made up the bulk of the 26,500,000 tons of cash crops produced.

Pakistan experienced a "green revolution" during the 1960s and early 1970s. In this period wheat production increased from 4,000,000 to 6,400,000 metric tons, leaving a surplus over domestic consumption that was partly shipped to East Pakistan (now Bangladesh) and partly exported. Cotton production rose from 2,000,000 to 4,000,000 bales in the course of the 1960s and early 1970s, but yields remain low by international standards. More and more of the crop is being locally processed, with a corresponding decline in exports, although foreign sales of manufactured cotton have been rising rapidly. Large domestic sugar subsidies were the main cause for the increase in sugarcane production from 15,000,000 to 23,000,000 tons in the second half of the 1960s and early 1970s.

Animal husbandry accounted for 12 percent of the gross domestic product in 1971 to 1972, or about 30 percent of the agricultural sector's total contribution. Apart from the supply of meat and dairy products for local consumption, it includes production of wool for the carpet industry and for export and of hides and skins for the leather industry.

The contribution of forestry to national income remains negligible, but that of fisheries has been rising. In 1969 the total catch of what was then West Pakistan amounted to 180,000 tons.

Mining and manufacturing. Mining and quarrying account for less than 1 percent of gross domestic product and of total employment. Production of natural gas has been by far the fastest growing sector of the industry: total output rose from 26,000,000 cubic feet (740,000 cubic metres) in 1959 to 1960 to 115,000,000 cubic feet (3,200,000 cubic metres) in 1970. Coal production amounted to just over 1,000,000 tons a year at the beginning of the 1970s, having grown at an average annual rate of 5 percent in the preceding decade. Output of limestone doubled during the 1960s to about 1,700,000 tons. It provides the basis for a growing cement industry.

By the early 1970s the manufacturing sector accounted for 17 percent of gross domestic product, compared with 12.5 percent ten years earlier. But it still employs only 10 percent of the official labour force. During the 1960s manufacturing output rose at an average annual rate of 9 percent. Most of the growth occurred in the large-scale industries, which represented 72 percent of total manufacturing output at the end of the 1960s.

Industrialization

The beginning of the main industrialization effort dates back to the Korean War boom and to the early 1950s. Initially, it was based on the processing of domestic agricultural raw materials for the home market and for export. This led to the setting up of cotton textile mills—an industry that accounts for 37 percent of the total employment in industry. Woollen textiles, sugar, paper, tobacco, and leather industries provided jobs for a further 20 percent of the industrial labour force. The growing trade deficit in the mid-1950s compelled the government to cut down on imports. This encouraged the implantation of a number of import-substitution industries. At first, they produced mainly consumer goods, but, gradually, they came to include intermediate goods and a range of capital goods, including chemicals, fertilizers, and light engineering products. Nevertheless, Pakistan still has to import a large proportion of the capital equipment and raw materials required by industry. Notwithstanding the diversification in the industrial sector during the 1960s, cotton textiles still account for a major share of total output.

The cotton mills account for about one-third of the manufacturing sector's contribution to Pakistan's total gross domestic product and with a labour force of 150,000 constitute the single largest sector of employment in the country. There has been a gradual diversification of the production mix, both in response to domestic demand and because of restrictions in export markets. Output of yarn amounted to nearly 670,000,000 pounds (300,000,000 kilograms) in 1970 to 1971, compared with 360,000,000 pounds (160,000,000 kilograms) ten years earlier; in the same period, output of cloth rose from 614,000,000 to 787,000,000 yards (560,000,000 to 720,000,000 metres).

Energy. Although energy production rose faster than the economy as a whole during the 1960s, it did not keep pace with demand, and consequent shortages of fuel and electric power inhibited industrial growth. The exploitation of domestic natural-gas resources was a major breakthrough, while heavy expenditure on thermal plants and hydroelectric resources increased electricity generating capacity more than fivefold during the 1960s.

Financial services. Pakistan has a fairly well developed system of financial services. The State Bank of Pakistan has overall control over the banking sector, which consists of a number of commercial banks and specialist credit institutions. The State Bank acts as banker to the central and provincial governments and administers official monetary and credit policies, including exchange controls. It has sole currency-issuing rights and has custody of the country's gold and foreign exchange reserves. In 1972 in Pakistan there were 25 scheduled banks subject to strict State Bank requirements as to paid-up capital and reserves. They accounted for the bulk of total deposits, collected through a network of about 2,400 branch offices. Four specialist financial institutions provided medium- and long-term credit for industrial, agricultural, and housing purposes. They were the Pakistan Industrial Credit and Investment Corporation (PICIC), the Industrial Development Bank of Pakistan (IDBP), the Agricultural Development Bank of Pakistan (ADBP), and the House Building Finance Corporation. The Karāchi stock exchange dealt in the stocks and shares of 220 registered companies in 1970. There is another stock exchange in Lahore, opened in 1971. The Investment Corporation of Pakistan (ICP), set up by the government in 1966, and the National Investment Trust have helped to channel domestic savings into the capital market. ICP mutual funds were heavily oversubscribed in the late 1960s.

Banking and credit

Foreign trade. Between 1961 and 1971 what was then West Pakistan's exports grew at an average annual rate of 5.2 percent to a total of $704,000,000. In the same period, imports increased by about 1.4 percent a year, reaching $730,000,000 in 1971–72. Although exports rose faster than imports in the 1960s, overall trade remained heavily in deficit, with total export earnings covering only about half of the import bill. The magnitude of these deficits began to decline during 1971, and for the first eight months of 1972, Pakistan actually showed a slight positive balance of trade. During the 1960s some important changes took place in the composition of foreign trade. In particular, while the proportion of total export from primary commodities, including raw cotton, fell, the share of manufactures almost doubled. But the bulk of the manufactured products coming into the export trade consisted of cotton goods, so that Pakistan remained as dependent as ever on its leading cash crop. The other manufactures exported came mostly from industries based on agriculture, such as leather and leather goods and carpets. The shift toward manufactured agricultural exports, which have a higher added value content than the primary commodities, has been strongly encouraged by the government. It is also because of government policies that the 1960s saw a gradual change in the composition of imports. The trade deficit and the shortage of foreign exchange made it necessary to cut down on all but the most essential purchases abroad. As a result, the proportion of the total import bill spent on consumer goods fell from 25 percent to 12 percent to about 23 percent in 1971–72. Capital goods accounted for 42 percent and industrial raw materials for the remaining 35 percent of total imports in 1971–72.

By the early 1970s Hong Kong was Pakistan's biggest single export market, followed by Japan and the United Kingdom. But the import markets were gradually being diversified, and, although the United States was still among the main suppliers, Japan and the European Economic Community (EEC) were gaining ground rapidly.

MANAGEMENT OF THE ECONOMY

The private sector. The government has traditionally supported a system of free enterprise and has encouraged private domestic and foreign capital investment within the framework of the development plans. The policy has been comparatively successful. At the time of partition, in 1947, numbers of Muslim businessmen from what was to become secular India moved to Lahore and Karāchi and set up the nucleus of a new industrial and financial society. They were helped by the strong bias in favour of the manufacturing sector maintained by the government in its foreign-exchange allocation policy, particularly in the early 1950s. The availability of imported machinery and raw materials helped to promote a process of fairly diversified industrialization. Overall, private enterprise in the manufacturing and financial sectors has been concentrated in the hands of a small number of family cartels. This was one of the big grievances nursed by the trade unions and left-wing political groups, which contributed to the social unrest of the late 1960s.

In an effort to remedy this situation, the government in 1972 introduced a program of economic reforms which placed control over a number of key industries, such as iron and steel, chemicals, and cement, as well as the insurance sector, in the hands of the state.

Foreign investment

Foreign private investors have been attracted to Pakistan by favourable investment incentive schemes, including tax holidays, long-term credit facilities from local industrial-financing institutions, and repatriation guarantees for capital and profits. By the end of the 1960s, before the secession of Bangladesh, foreign private investment in Pakistan was estimated at about $600,000,000. United Kingdom investors accounted for the largest share of the total as a result of their predominant pre-independence position; but United States, German, and Japanese investments were also gaining in importance.

The public sector and the role of government. Investment by the public sector has been concentrated on social programs (education, health, etc.) and on the infrastructure (transport, communications, and power), leaving agriculture and industry to the private sector. Even in these sectors, however, when private investment would not or could not fulfill the set development targets, the government has stepped in. The development targets themselves have been spelled out in a series of five-year plans that have given the economy its general direction, but the government has also made use of a number of other methods to make short-term changes. Thus, the biannual foreign-trade policy could liberalize or restrict imports at short notice, while manipulations of bonus rates proved to be a flexible way of promoting selected exports. Subsidies, price controls, and tariff protection have all been used with varying degrees of success.

Taxation. The overall tax burden was estimated at 12 percent of the gross national product in 1969 to 1970. Even though income tax rates rise to comparatively high levels—the top rate in 1972 was 70 percent—the tax base is so small that revenue from this source and from the 30 percent corporation tax has remained disappointingly low. The government has therefore had to rely on indirect taxes for most of its current income. It has not made much use of deficit financing, and the domestic public debt per head of population has been well below the world average. The government has been able to maintain heavy expenditure on development and defense because of the inflow of foreign aid. In doing this, however, the country has accumulated an enormous foreign debt, the financing of which was a major problem in the early 1970s.

The central-government budget is divided into a revenue budget, which deals with current expenditure, and a capital budget, covering development expenditure under the five-year plan. Taxation yields over 70 percent of revenue receipts. Excise duties alone accounted for more than 30 percent of revenue by the early 1970s. Customs duties provided a further 25 percent. Less than 10 percent of current revenue came from income and corporation taxes. Capital-budget receipts consisted of at least half of foreign aid. The provincial governments draw up their own budgets along lines similar to those adopted by the central government. Revenue consists of a share of the central government's tax receipts and of local taxes levied by the provincial governments themselves. Capital receipts are almost entirely made up of loans from the central government.

Sources of revenue

In the 1960s and early 1970s defense absorbed about half of the central government's current expenditure, but twice as much again was spent on development through the capital budget.

Trade unions and employer associations. The trade-union movement dates back to the late 19th century, but, because Pakistan's industrial sector (inherited at independence) was so small, organized labour as a proportion of total employment is still in a minority. This has not prevented it from becoming an important political force. The total union membership at the end of the 1960s (*i.e.*, before the secession of Bangladesh) was about 750,000, out of a labour force of some 30,000,000. There were well over 1,000 registered unions, most of them organized within individual establishments. Countrywide unions based on a common craft or industry are very few. Most of the unions are situated in the urban centres and are affiliated to one of three national labour confederations. These, in turn, are affiliated to the International Confederation of Free Trade Unions (ICFTU).

Many of the labour laws still in force in the early 1970s dated back to British colonial days. The extensive labour unrest in the late 1960s eventually led to a tripartite labour conference in 1969, attended by government, employer, and trade-union representatives. A new labour policy was agreed upon, dealing in particular with industrial relations, minimum-wage legislation, and workers' welfare. But, because of the high rates of unemployment, employers remained in a strong position, and many of them were able to bypass the new working conditions. Only the unions in the bigger industries (*e.g.*, cotton textiles) had the necessary coherence to fight back. The resulting wave of strikes and the rise in industrial costs slowed growth in the manufacturing sector.

Contemporary economic policies. The Pakistan government's economic strategy has been spelled out in the development plans. The first one was a six-year plan that came into operation in 1957 and was implemented through annual development programs. The projected expenditure targets were largely fulfilled, but production in agriculture particularly fell short of the expected levels. The first plan provided valuable experience for the formulation of the second plan, covering the period 1960 to 1965. This plan proved to be exceptionally successful: gross national product grew at an average annual rate of 5.5 percent, instead of the target rate of 4.7 percent, and agricultural output rose by 3.5 percent a year, compared with the 1.3 percent achieved during the first plan period. The manufacturing sector expanded by 10 percent a year, and the combined aims of import substitution and export promotion brought about a structural improvement in the foreign-trade sector.

The third plan (1965 to 1970) was introduced as the first five-year stage of a 20-year plan that would carry Pakistan through to 1985. The main objectives of this perspective plan were to quadruple the national product and to increase income per head from about $83 U.S. in 1965 to $200 U.S. in 1985, to achieve parity in incomes between East and West Pakistan, to eliminate the dependence on foreign aid, to provide full employment, and to achieve universal literacy. But the war with India in 1965, the falloff in foreign aid, setbacks in agriculture, the socio-political unrest in the late 1960s, and the secession of Bangladesh early in the 1970s all combined to make the third plan a period of disappointments. Exports grew by 7 percent a year compared with a 9.5 per-

cent target, imports of capital goods and raw materials declined, and foreign-aid receipts were lower than expected. The problems and imbalances that emerged during the third plan determined the guidelines for the fourth plan, which was to run from 1970 to 1975. The political disruption from 1970 to 1972, culminating in the secession of Bangladesh, left the plan in suspense, while annual development programs took its place.

The three major economic problems

Problems and prospects. Without entering into the political aspects of the secession of Bangladesh, it may be said that economic problems and prospects in Pakistan have not been essentially changed by this event. The three major economic problems that faced the government in the first half of the 1970s were those of slow overall growth, inequalities in income distribution, and the deterioration in the balance of payments (although much progress had been made with the latter by late 1972). The basic solution lies in a higher rate of investment. Although the planners had hoped that by 1970 investment expenditure would have risen to 20 percent of gross national product, it amounted to only 14 percent. This was partly because of the falloff in domestic savings and partly because of the reduction in foreign aid. The marginal rate of savings in the second half of the 1960s, instead of rising to the planned level of 20 percent, came down to less than 10 percent as a result of the political upheavals and declining profit levels in industry. Meanwhile, the contribution made to the economy by foreign aid had been smaller than expected, while at the same time foreign debt servicing was becoming more onerous, reducing the net inflow of new aid even further. The government tried to renegotiate its debt position, particularly with the Aid Pakistan Consortium (the group of countries and international organizations, including the World Bank, that underwrote the bulk of development spending in the 1960s). But debt servicing will remain a heavy burden on the economy and threatens to absorb up to one-fifth of Pakistan's export earnings during the 1970s.

A selective increase in investment expenditure would go a long way toward reducing the income inequalities that emerged so blatantly during the 1960s. Investment in the less developed regions of Baluchistan and the Northwest Frontier Province will also have to be raised if the income gap between them and the two better-off provinces, Sind and Punjab, is not to grow even wider. Finally, the government will have to take strong measures to combat the social concentration of wealth and the enormous income disparity between the rich and the poor. A start was made in the early 1970s with such a program, but the power of the big family cartels was difficult to break.

The third major problem facing Pakistan's economy during the 1970s concerned the balance of payments. Every effort will have to be made to increase earnings from traditional exports, such as cotton. At the same time, new exports will have to be developed if the country is to maintain its share of world trade. After export earnings, foreign aid is the most important balance-of-payments factor that will influence the economy during the 1970s. It is closely related to the problem of overall growth in that the availability of foreign credit will determine the level of new investment as well as the flow of imported raw materials and spare parts to maintain output. Even though Pakistan hopes to be able to do without foreign aid in the long term, during the 1970s external assistance will have to continue if only to make it possible to repay the outstanding debt without further reducing essential imports. (E.I.U.)

TRANSPORTATION AND COMMUNICATIONS

Component systems. *Roads and railways.* The dominant role of rail as the principal long-distance carrier has been displaced by the bus and truck; road transport now accounts for two-thirds of the total passenger miles and about one-half of the total freight. Motor trucks and tractor-drawn trailers are also displacing the traditional bullock cart for local transport of produce to markets. In 1970 there were about 12,700 miles (20,400 kilometres) of paved roads and, in 1969, 12,400 miles of unpaved roads. The main arterial road, which runs from Karāchi to Peshāwar via Lahore and Rāwalpindi, is 1,080 miles (1,740 kilometres) long. The rapid increase in road traffic had caused a rapid deterioration in road surfaces by the early 1970s. There are about 7,600 miles (12,000 kilometres) of railroad track for a total route mileage of 5,300 miles—a length that has changed little since independence. The main route runs more than 1,000 miles north from Karāchi to Peshāwar, via Lahore and Rāwalpindi. Another main line branches northwestward from Sukkur to Quetta.

Air transport. Pakistan International Airlines (PIA), established in 1955, is the sole carrier of internal air traffic. PIA also runs international flights to Europe, the Middle East, Africa, the Far East, and China, as well as to neighbouring Afghanistan. The principal airports are at Karāchi, Lahore, Rāwalpindi, and Peshāwar.

Shipping. Pakistan's merchant fleet in 1972 consisted of about 131 ships with a total tonnage of 739,000 deadweight tons. The semi-governmental National Shipping Corporation owns less than one-half of the ships. Karāchi is the principal port.

Postal services and telecommunications. There were over 8,000 post offices in 1970. The volume of mail carried has increased by 130 percent since independence. The service whereby remittance facilities are sent through the post office is of particular importance for the population in general. The total number of telephones in Pakistan was 207,000 in 1971; much of the equipment is manufactured locally. Teleprinter services connect Karāchi, Lahore, and Rāwalpindi. International communication is available by direct telephone and international radiotelegraph circuits, and a Telex service is available for many places abroad. A high-capacity microwave communications system, linking Pakistan, Iran, and Turkey, has been further extended to London, with links to the European network.

IV. Administration, social conditions, and cultural life

THE STRUCTURE OF GOVERNMENT

The political system of Pakistan has undergone several far-reaching changes since independence; in 1971, its turbulent politics culminated in the secession of its eastern region (having more than 54 percent of the total population at that time), which established itself as the independent state of Bangladesh. In the aftermath of that event, in 1972, Pakistan (now reduced to what was previously West Pakistan) was in the throes of political and economic crises and uncertainties about its future.

Political stresses

Three distinct sets of conflicts have left their mark on the politics of Pakistan. The first of these, initially obscured by the paraphernalia of a parliamentary form of government but later made manifest by overt seizure of power by men at the head of the military and bureaucratic establishment, is a continuing struggle between political leadership and a military–bureaucratic oligarchy for supremacy and authority in the state; ideologically, this struggle was expressed as a struggle for democracy. A second and a quite distinct conflict was a struggle between regional groups. Because it was directed against centralized authority, it merged with the democratic struggle. But its express aims were focussed on securing greater regional representation in the bureaucratic and military establishment, especially in the higher echelons, as well as achieving effective decentralization of powers within a federal governmental structure by emphasizing regional autonomy. A third set of conflicts concerned the allocation of economic resources and burdens and the distribution of a greater share of the benefits of development among the more deprived regions and strata of the population. In particular, a small number of business families control the bulk of the industrial wealth of the country. Agricultural wealth is concentrated in the hands of the landlords, especially of the Canal Colonies of the Punjab and of Sind. The incomes of the bulk of the rural as well as of the urban populations all over the country have been rapidly eroded by inflation and higher indirect taxes. This has resulted in spontaneous explosions of popular discontent in a leaderless form of revolt that has taken by

surprise not only the government but also the entire established political leadership ranging from the extreme right to the extreme left.

The constitutional framework. The task of framing a constitution for independent Pakistan was entrusted in 1947 to a Constituent Assembly that was also to function as the country's interim legislature under the Government of India Act (1935 [as adapted]), which was to be the interim constitution. It was federal in form, with the Constituent Assembly and a governor general at the centre and with provincial assemblies with governors of provinces on the regional level. Government was to be under Cabinets responsible to the central assembly and the provincial assemblies, respectively. Extraordinarily wide powers were, however, vested in the governor general, which were to prove decisive in establishing the power relationship between the bureaucratic and military establishment and the political leadership.

Pakistan's first constitution was enacted by the Constituent Assembly in 1956. It followed the form of the Government of India Act (1935), allowing the president far-reaching powers to suspend federal and provincial parliamentary government. The 1956 constitution also included the "parity formula," by which representation in the National Assembly for East and West Pakistan would be decided on a parity rather than population, basis. (A major factor in the political crisis of 1970–71 was abandonment of the "parity formula" and adoption of representation by population, giving East Pakistan an absolute majority in the new National Assembly.)

In 1958 the constitution was abrogated and martial law was instituted. A new constitution, promulgated in 1962 provided for the election of the president and National and provincial assemblies by an electoral college composed of 80,000 members of local councils.

Although a federal form of government was retained, the assemblies had little power, for, in effect, power was centralized through the authority of governors acting under the president. A massive revolt developed spontaneously in the winter of 1968–69, stimulated ironically by the regime's over-enthusiastic celebrations of its "Decade of Development." This led to the resignation of the President and the installation as president and chief martial-law administrator of the Commander in Chief of the army; the 1962 constitution was superseded by martial law. The government promised, however, to hold general elections on the basis of direct universal adult franchise and to convene a constitutional assembly.

Re-establishment of the four provinces

The popular demand to dissolve the One Unit—the single province of West Pakistan—and to reconstitute the provinces of the region was also conceded. The four existing provinces were, accordingly, re-established in 1970. Elections were held in the same year. In what was then West Pakistan, the Pakistan People's Party, which called for Islāmic socialism but enjoyed the support of powerful landowning groups, secured 60 percent of the West Pakistani seats, mainly from the two major provinces of the Punjab and Sind. The National Awami Party and the Jamīʿat-e ʿUlamā'-ye Islām (Hazarvi) secured strong positions in Baluchistan and the Northwest Frontier Province, where they drew support through regionalist sentiment. In East Pakistan, the electoral victory of the Awami League, which was committed to a program of regional autonomy, precipitated a political crisis that eventually led to the establishment of Bangladesh in 1971. As a consequence of the repercussions of the crisis. Zulfikar Ali Bhutto, the leader of Pakistan People's Party, was installed as president and chief martial-law administrator. He dismissed a large number of officers holding senior commands in the armed forces and also the military governors of the provinces, in whose place he installed his own party men as governors. The National Assembly, which was convened in April 1972, enacted an interim constitution and endorsed the presidential appointment. Martial law was then lifted. In April 1973 a new constitution, the third in Pakistan's 25-year history, was adopted by the National Assembly.

Political parties. The unity of Pakistan, it is sometimes argued, is premised on its Islāmic ideology, for that is the only cement that could bind together, it was supposed, its culturally diverse peoples. But it could be argued equally that it was the insistence on Islāmic ideology, in opposition to regional demands expressed in secular and cultural idiom, that progressively alienated regional groups and eroded national unity. The Muslim League, the party of Muslim nationalism in India, the aspirations of which were fulfilled by the creation of Pakistan, had by 1947 pushed all other political parties into insignificance. It had put forward the "two nation theory," which declared that Indian Muslims were a separate nation from other Indians. But its political idiom was not theological. It was the party of "Muslim modernists," the educated Muslim middle classes who aspired to more secure and more privileged positions in the worlds of government and public service and in business. The Muslim theocratic parties, however, such as the Jamīʿat-e Islāmī or the Jamīʿat-e ʿUlamā'e Hind, were opposed to the creation of Pakistan.

After independence the Muslim League was rapidly fragmented and the fragments developed into rival political parties. The rump of the Muslim League allied itself closely with the bureaucratic–military establishment, and, reluctant to face an electorate in elections, it was willing to submit to the bureaucratic–military establishment as a price for holding office. To oppose regionalist and democratic demands against centralized authority, the Muslim League, to an extent that it had never done before, began to stress Islāmic ideology, which it advocated as the only basis for the unity of Pakistan in opposition to regionalist secular politics, thus further alienating regional sentiment.

Comparative strength of political parties

At the time of the 1970 elections, the Pakistan People's Party was the largest, securing 81 out of West Pakistan's 144 seats in the National Assembly. The three factions of the Pakistan Muslim League—the Qayyum, the Convention, and the Council—won nine, one, and seven seats, respectively. Jamāʿat-e Islāmī won seven seats, Jamīʿat-e ʿUlamā'-ye Pakistan seven seats, and the National Awami Party (Wali Khan group), which was strong in Baluchistan and the Northwest Frontier Province, six seats.

Local government and administration. The basic administrative structure has remained virtually unchanged since the colonial period, despite all the constitutional upheavals and changes at the national and provincial levels. Provinces are subdivided into divisions, districts, and *tahsil*s (district subdivisions), which are run by a hierarchy of administrators, such as the divisional commissioner, the deputy commissioner (in the district), and the subdivisional magistrate, subdivisional officer, or *tahsildār* at the *tahsil* level. The key level is that of the district, where the deputy commissioner controls all branches of government, being directly in charge of the administration of revenue and having judicial functions as the district magistrate; he also controls the police.

District councils

Local self-government in the form of district boards has existed since colonial times. In 1959, however, these were transformed into district councils with the deputy commissioner as chairman instead of having an elected chairman; they were also made a part of a hierarchy of councils, which were set up at the divisional as well as at the *tahsil* level. The councils were subject to extremely close control and regulation by administrative officers. The system fell into disrepute because of corruption as well as because of authoritarian control, and there was a public demand to disband them; the system has been moribund since 1969 and was to be restructured drastically.

The administration itself is run by officials organized as a hierarchy at the top of which is the elite Civil Service of Pakistan (CSP). Below the CSP is the Provincial Civil Service (PCS) and several specialist departments, such as the Pakistan Police Service (PPS) and the departments of irrigation, agriculture, etc.

Justice. There is a division between the judiciary and the executive branches of government. The judiciary consists of the Supreme Court, the provincial high courts, and (under their jurisdiction and supervision) district courts that hear civil cases and sessions courts that hear criminal cases. There is also a magistracy that deals with

cases brought by the police. The district magistrate (who, as deputy commissioner, also controls the police) hears appeals from magistrates under him; appeals may go from him to the sessions judge. The Supreme Court is a court of record. It has original, appellate, and advisory jurisdictions and is the highest court in the land. At the time of independence, Pakistan inherited legal codes and acts that remain in force, subject to amendment.

Armed forces. The army is the senior service and had an estimated strength of 278,000 in the early 1970s; it consisted of one armoured brigade, with U.S. M-4 Sherman, M-47 Patton, and Chinese T-59 tanks; 10 infantry divisions; about 900 artillery pieces; and one air-defense brigade with anti-aircraft guns. At the same time, the total strength of the navy was 10,000 men; it had three submarines, two destroyers, three destroyer escorts, two frigates, six coastal minesweepers, and supporting units. The air force had a strength of 17,000 men, with another 2,000 as reservists. It was equipped with 200 combat aircraft, of which 140 were U.S. F-86 Sabre jet interceptor fighter bombers. The armed forces suffered heavy losses in East Pakistan during 1971, especially during the two-week war in December with Indian forces. The losses were estimated to be of the order of nearly half the navy, a quarter of the combat aircraft, and about 30 percent of the ground forces. One of the first acts of President Bhutto was to dismiss a large number of senior army generals and naval and air-force officers; this was not only because of their responsibility for the debacle in East Pakistan but also to reduce the top-heavy military establishment.

SOCIAL CONDITIONS

The literacy rate

Education. The literacy rate for persons over five years of age is over 16 percent. About 10 percent of the literates are without any formal education. Educational levels for women were much lower than those for men, the percentages of literates among the female population being about 7 percent. The share of females in educational levels progressively diminishes above the primary school level.

There are five multifaculty universities, as well as an agricultural university and an engineering university. There are also a number of colleges offering degrees, as well as some medical colleges.

Welfare. Welfare services are inadequately developed in relation to needs. In 1970, for a population of about 60,000,000, there were only 13,400 doctors and 4,700 nurses. Annual expenditure on education, health, and social welfare pales into insignificance beside the annual defense expenditure.

CULTURAL LIFE AND INSTITUTIONS

The cultural milieu. Pakistan shares influences that have shaped the cultures of the peoples of South Asia. There are thus wider regional similarities extending beyond the national boundaries. On the other hand, there are specific cultures of the various regions of Pakistan that present a picture of rich diversity. It would be difficult to speak of a culture of Pakistan in the singular.

The status of women

Family organization is strongly patriarchal, as in most agrarian societies, and most people live in large extended families. A woman's place in society is low, and she is restricted to the performance of domestic chores and to fulfilling the role of a dutiful wife and mother. Rich peasants and landowners and members of urban middle classes keep their women in seclusion (*pardah*); on the rare occasions on which they set foot outside their houses, they must be veiled. Among poor peasants, women have duties on the farm as well as in the house and do not observe *pardah*. In the Punjab, picking of cotton is exclusively a woman's job, and women keep the money thus earned for their own purposes. Houses of those who practice *pardah* have a men's section (*mardānah*) at the front of the house, so that visitors do not disturb the womenfolk, who are secluded in the *zanānah* (women's section).

Among the very rich, Western education and modes of living have eliminated *pardah*, but, in general, even among this group, attitudes toward women in society and the family are akin to those of Victorian England. Change is coming most rapidly among the urban lower middle income group, in which women are forced to seek employment under the pressure of economic necessity; *pardah* is then cast off and the education of women encouraged. In consequence, some women have gained distinction in the professions; significantly, some of the country's leading trade unionists are women.

Social organization revolves around kinship rather than caste. *Berādarī* (patrilineage) is the most important social institution. Marriage is preferably with the father's brother's daughter, and among many groups marriages are invariably within the *berādarī*. The lineage elders constitute a council that adjudicates disputes within the lineage and acts on behalf of the lineage with the outside world—for example, in determining electoral allegiances.

Pakistan claims a cultural heritage dating back more than 5,000 years, to the epoch of the Indus civilization. But the emphasis on Islāmic ideology has brought about a strong romantic identification with Islāmic culture—not only that of India but of the whole of the Islāmic world. *Qawwālī*, a form of devotional singing, is very popular. Poetry is also a popular rather than an esoteric art, and public poetry recitations called *mushā'irah*s are organized like musical concerts. Urdu, Sindhi, and Pashto poets are regional and national heroes. Literary tradition is the richest of all Pakistani art forms. Music and, especially, dancing are comparatively less developed arts. The visual arts, too, play little part in popular folk culture. In recent years, however, painting and sculpture have made considerable progress as expressions of a new sophisticated urban culture.

The cinema is now the most popular form of entertainment. About 50 feature films are produced each year, mostly in Urdu. The songs and music used in the films have a distinctive character and are often reproduced on phonograph records and broadcast on the radio.

Broadcasting and the press. During the 1960s radio and television attempted to harness folk cultural traditions (especially in song, music, and drama) for political and nonpolitical propaganda purposes. The ubiquitous transistor radio has brought broadcasts into every village. By 1970 it was expected that 65 percent of the area of West Pakistan and 85 percent of its population would be within the range of medium-wave broadcasting. Programs, including overseas broadcasts, are sent out in 17 languages. Television came to Pakistan in 1965, and its development is slow. There are three principal television broadcasting stations—at Karāchi, Lahore, and Rāwalpindi–Islāmābād.

Newspapers and periodicals

In 1965 there were about 450 English-language, 700 Urdu, 60 Sindhi, and two Bengali newspapers and periodicals in what was then West Pakistan. Among these, there were about 80 daily newspapers, of which only a handful had circulations between 20,000 and 160,000. In 1959 the government took over a major group of newspapers and in 1964 set up the Press Trust of Pakistan, which then took over most of the most important newspapers and periodicals. Censorship was subsequently imposed at various times, but it was replaced by what was euphemistically called a "press-advice" system by which the ministry of information gave guidance to the press.

PROSPECTS FOR THE FUTURE

The most significant fact about Pakistan is the rich diversity of its many regions and peoples. Its natural regions embrace a whole gamut of ecological variations, and its peoples are a mixture of many racial types speaking different languages and having different cultures.

But, throughout its short existence, Pakistan has been a nation in search of its identity and in search of national myths that could explain its existence and sustain its unity. Regionalist demands have been questions of a different order and related primarily to the allocation of resources and opportunities in the economy and the polity. But those who were concerned with the preservation of central authority and privilege resorted increasingly to opposing regionalist sentiment with an ideology of Islām-

ic unity. Paradoxically, instead of binding the nation more firmly together, the emphasis upon ideology appears to have convinced disgruntled regional groups that by that means their demands and needs were being ruled out of court. Ironically, therefore, the stress upon Islāmic unity and the ideological basis of Pakistan's existence strengthened centrifugal forces rather than cemented bonds. Both the emphasis upon ideology and the regionalist demand for autonomy made the constitution the central issue in political debate in Pakistan, virtually to the exclusion of most other issues. But, by the early 1970s, there had been a clear shift in emphasis to questions of equity and equality, both between regions and between different strata of the community. The secession of Bangladesh in 1971 was probably the culmination of that phase of Pakistan politics that was exclusively focussed on the regional issue and on its mirror image, the issue of a unifying Islāmic ideology. A new politics subsequently emerged in which economic, political, and social problems were projected in terms that transcended the regional perspective, while recognizing regional needs. Along with this shift of emphasis, there were indications of recognition that regional diversity of culture and language were not to be taken as signs of national weakness but as sources of national pride.

BIBLIOGRAPHY. M.L.P. PATTERSON and R.B. INDEN, *South Asia: An Introductory Bibliography* (1962), cites more than 4,000 entries on Pakistan with emphasis on history, culture, and politics. SHAUKAT ALI and GARTH N. JONES, *Pakistan Government and Administration: A Comprehensive Bibliography*, 2 vol. (1970–71), is exhaustive. O.H.K. SPATE and A.T.A. LEARMONTH, *India and Pakistan: A General and Regional Geography*, 3rd ed. (1967), is comprehensive and authoritative; KAZI S. AHMAD, *A Geography of Pakistan* (1964), is conveniently brief. ALOYS A. MICHEL, *The Indus Rivers* (1967), is concerned with development and with the dispute with India over allocation of river waters. W. CANTWELL SMITH, *Modern Islam in India* (1947), is as yet unsurpassed as an introductory account of the Muslim intellectual and political movements that culminated in the creation of Pakistan. KHALID B. SAYEED, *The Political System of Pakistan* (1967), is an excellent general work, and a worthy successor to his *Pakistan: The Formative Phase, 1857–1948*, 2nd ed. (1968). G.W. CHOUDHURY, *Constitutional Development in Pakistan* (1959) and *Democracy in Pakistan* (1963), is strongly oriented towards the "Muslim League" point of view, whereas TARIQ ALI, *Pakistan: Military Rule or People's Power?* (1970), offers a left-wing analysis. G.W. CHOUDHURY (ed.), *Documents and Speeches on the Constitution of Pakistan* (1967), is useful. RALPH BRAILBANTI, *Research on the Bureaucracy of Pakistan* (1966), is a comprehensive survey of the literature; and H.F. GOODNOW, *The Civil Service of Pakistan: Bureaucracy in a New Nation* (1964), is a valuable study of the structure of the bureaucracy and its role in public life. H.N. GARDEZI (comp.), *Sociology in Pakistan* (1966); and A.S. DIL (ed.), *Perspectives on Pakistan* (1965), contain essays by Pakistani scholars; STANLEY MARON (ed.), *Pakistan: Society and Culture* (1957), contains essays on history and rural social organization. FREDRIK BARTH, *Political Leadership Amongst Swat Pathans* (1959), is a classic study in political anthropology. S.K. CHATTERJI, *Languages and the Linguistic Problem*, 3rd ed. (1945), summarizes the *Linguistic Survey of India*, ed. by GEORGE A. GRIERSON, 11 vol. (1903–28); the *Census of Pakistan* (1951), but not that of 1961, briefly outlines the subject. B.S. GUHA, *Racial Elements in the Population* (1944), synthesizes the most widely accepted classifications of races of South Asia. V.G. KIERNAN, *Poems from Iqbal* (1955) and *Poems by Fraiz* (1971); H.T. SORLEY, *Shah Abdul Latif of Bhit* (1940, reprinted 1966); and G. ALLANA (ed.), *Presenting Pakistani Poetry* (1961), are some examples of poetry from Pakistan available in English translation.

(H.A.A.)

Palamas, Saint Gregory

Gregory Palamas, a 14th-century Greek Orthodox monk and theologian, is acknowledged as the intellectual leader and apologist for a monastic school of mysticism called Hesychasm, from the Greek word *hesychia*, or state of quiet. This Byzantine contemplative movement, at one time controversial but now sanctioned by the Orthodox Church as a legitimate form of prayer, is an ascetical method that integrates repetitive prayer formulas with bodily postures and controlled breathing for the purpose of experiencing a state of inner peace and mystical union. Its history includes involvement in Byzantine political and ecclesiastical upheavals and sharp polemics with Greek and Latin theologians.

Born in 1296 at Constantinople of a distinguished family with ties to the imperial court, Palamas mastered the classical philosophies of antiquity at the imperial university. In 1316, however, he renounced a political career in order to become a monk at Mt. Athos in northeastern Greece, the spiritual centre of Greek Orthodoxy. For 25 years he immersed himself in study and reflection on the sacred Scriptures and the writings of the Church Fathers; he was introduced to contemplative prayer by a spiritual master and in turn became a master for other initiates. Raids by the Turks in about 1325 forced him to interrupt his monastic life on Mt. Athos and to flee to Thessalonica and Macedonia. He was ordained a priest in 1326 and later, with ten companions, retired to a hermitage in Macedonia.

Early monastic training and responsibilities

He returned to Mt. Athos in 1331 to the community of St. Sabas and *c.* 1335 was chosen a religious superior (*hēgoumenos*) of a neighbouring convent. Because of differences with the monks who considered his spiritual regimen too strict, he resigned after a short term and returned to St. Sabas.

In 1332, Palamas entered into a theological dispute that lasted for a quarter of a century and involved polemics with a series of Greek and Latin Scholastic theologians and certain rationalistic Humanists. His first adversary was Barlaam the Calabrian, a Greek monk residing in Italy who visited Constantinople and other Orthodox monastic centres to engage in philosophical disputation for intellectual prestige. Expounding a mode of theological agnosticism, Barlaam denied that any rational concepts could express mystical prayer and its divine–human communication even metaphorically. Subsequently, he composed a satirical work defaming Hesychasm by referring to its adherents as "men with their souls in their navel" (Greek *omphalopsychoi*). The image derived from the Hesychast meditative posture of focussing the eyes on a spot below the chest in order to heighten the mystical experience. Palamas responded to this attack by composing his "Apology for the Holy Hesychasts" (1338), also called the "Triad" because of its division into three parts.

Defense of Hesychasm

The "Apology" established the theological basis for mystical experience that involves not only the human spirit but the entire human person, body and soul. This doctrine attempts to articulate a prayer experience that devotees call "the deification of the entire man," a reference to the Hesychasts' claim of an inner transformation effected by a mystical illumination uniting man with God in the depths of his spirit. Hesychast spirituality strove to bridge the gulf between human and divine existence. It held the necessity of an intermediary relationship between man's world (immanence) and God's eternity (transcendence). Hesychast prayer aspires to attain the most intense form of God–man communion in the form of a vision of the "divine light," or "uncreated energy," analogous to the Gospel account of Christ's transfiguration on Mt. Tabor, as noted in Matt. 16:17 and Mark 8:9. The corporeal disposition for this contemplative state involves intense concentration and a methodical invocation of the name of Jesus (the Hesychastic "Jesus prayer"). Palamas emphasized the nonmaterialistic nature of Hesychast spirituality by explaining that the experience of inner light was not available to all but only to the "pure of heart" empowered by grace to perceive it.

After a succession of public confrontations with critical theologians and Humanists, and a politically motivated excommunication in 1344, Palamas had his teaching systematized in the *Hagioritic Tome* ("The Book of Holiness"), which became the fundamental textbook for Byzantine mysticism. The Hesychast controversy became part of a larger Byzantine political struggle that erupted in civil war. At its conclusion in 1347, Palamas, with support from the conservative, anti-Zealot party, was appointed bishop of Thessalonica. His administrative duties,

Excommunication and return to church administration

together with continued writings against his Humanist critics, occupied him for the rest of his life. He died in 1359.

In his fusion of Platonic and Aristotelian philosophy, used as a vehicle to express his own spiritual experience, Palamas set a definitive standard for Orthodox theological acumen. At the provincial Council of Constantinople in 1368, nine years after his death, he was acclaimed a saint and titled "Father and Doctor of the Orthodox Church," thus placing him among the ranks of those who determined the ideological shape of the Eastern Church.

BIBLIOGRAPHY. An English translation of selected Hesychast texts is in *Writings from the Philokalia on Prayer of the Heart*, trans. by EUGENIE KADLOUBOVSKY and G.E.H. PALMER (1951). See also the critical Greek text of Palamas' works, *Συγγραμματα*, ed. by B. BOBRINSKY *et al.* (1962–). JEAN MEYENDORFF, *Introduction à l'étude de Grégoire Palamas* (1959; Eng. trans., *A Study of Gregory Palamas*, 1964), is the principal biographical study in a Western language, with extensive bibliography. BASIL KRIVOSHEIN, "The Ascetic and Theological Teaching of Gregory Palamas," *The Eastern Churches Quarterly*, 3:26–33, 71–84, 138–156, 193–214 (1938), is the fullest exposition of Palamas' thought. Excellent surveys of Orthodox spirituality are VLADIMIR LOSSKY, *Essai sur la théologie mystique de l'Église d'Orient* (1944; Eng. trans., *The Mystical Theology of the Eastern Church*, 1957); and ERNST BENZ, *Geist und Leben der Ostkirche* (1957; Eng. trans., *The Eastern Orthodox Church: Its Thought and Life*, 1963). For further insights consult IRENEE HAUSHERR, *La Méthode d'oraison hésychaste* (1927); C. KERN, "Les Eléments de la théologie de Gregoire Palamas," *Irénikon*, 20:6–33, 164–193 (1947); M.J. LE GUILLOU, *L'Esprit de l'orthodoxie grecque et russe* (1961; Eng. trans., *The Spirit of Eastern Orthodoxy*, 1962); A.J. PHILIPPOU (ed.), *The Orthodox Ethos* (1964).

(A.H.Ao.)

Paleogeography

Paleogeography is the geography of selected portions of the earth's surface at specific times in the geologic past. This geography consists of interpreted reconstructions, produced in map form, which represent visual summaries of a wide variety of complex geological information. The maps provide a series of relatively instantaneous views of the earth through geological time. Paleogeographic maps may be as simple as those that merely give the former distribution of ancient lands and seas. They may be extremely comprehensive, however, showing the occurrence and distribution of fossil, plant, and animal communities, environments of sedimentation (*e.g.*, deltas, reefs, deserts, or deep-sea basins), areas undergoing uplift and erosion or subsidence and deposition, and major climatic zones. The value and accuracy of such maps depend, of course, upon the extent and reliability of the basic data available for a given geographic area and time interval. Even when such data are available, various interpretations can often be made and these may lead to quite different paleogeographic syntheses. Despite the many difficulties and uncertainties inherent in paleogeographic reconstructions, they are viewed basically as useful summaries of the existing knowledge of earth history.

Virtually all the scientific specialties within the geological sciences contribute to paleogeography. The disciplines of stratigraphy and sedimentology, paleontology and paleoecology, structural geology and geophysics, and petrology and geochemistry are particularly important. In this article the discussion will focus on the interrelations of these fields and of the data that they supply for paleogeographic reconstruction. Illustrations are drawn from selected paleogeographic syntheses of the Middle Silurian of North America, the Jurassic of the world, and the Middle Tertiary of Europe. For additional information on the subject matter of these disciplines and on relevant aspects of earth history, see SEDIMENTARY FACIES; SEDIMENTARY ROCKS; STRATIGRAPHIC BOUNDARIES; EARTH, GEOLOGICAL HISTORY OF; MOUNTAIN-BUILDING PROCESSES; CONTINENTAL DRIFT; CLIMATIC CHANGE; OCEANS, DEVELOPMENT OF; FOSSIL RECORD; and the several pertinent geological period articles—*e.g.*, TERTIARY PERIOD.

FACTORS INVOLVED IN PALEOGEOGRAPHIC RECONSTRUCTIONS

Interpretations of the distant geologic past are based largely on observations of the rocks of the earth's crust in light of knowledge of present geological processes. In the late 18th and early 19th centuries, the explanation of the earth's history in terms of observable geologic processes was known as uniformitarianism (*q.v.*) and was expressed by the phrase, "the present is the key to the past." The intent of this phrase was not only to direct attention to existing geologic processes but to discourage appeals to mystical causes or unique special events to explain the geologic history of the earth. Since the latter half of the 19th century, geologists have sought to understand the past through knowledge of the present, and uniformitarianism is viewed as a guiding principle for interpretation of the past. Such interpretation, in the case of paleogeography, requires recognition of the earth's constantly changing character as well as the invariance of certain kinds of earth processes.

A fundamental consideration in any paleogeographic reconstruction is the length of time involved and the precision with which it is known. Ideally, a paleogeographic reconstruction should be developed for an instant in time, but this is impossible to achieve. Usually, a paleogeographic map covers a certain interval of time and thus represents the average geologic situation during that interval.

Should the time interval be too large, the average will be meaningless or misleading. For example, a paleogeographic map for the last 1,000,000 years, which approximates the second half of the Pleistocene Epoch, would have to span several glacial and interglacial stages of earth history. Ice sheets, glacial sediments, fluctuation of shorelines with sea level, and migrating animal and plant communities that parallel shifting climatic belts would all have to be superimposed on one map. In that case, two maps, one for a "typical" glacial stage and one depicting an interglacial stage, although not precisely accurate for each of the Pleistocene stages, would be infinitely more informative. Time intervals for any particular paleogeographic reconstruction should, therefore, be short, relative to the rates of geological change in the area. Thus, two reconstructions are the minimum requirement for the last 1,000,000 years of earth history, but a single map spanning a few tens of millions of years might reasonably portray North American paleogeography during the Early Silurian Period when large parts of the continent were covered by shallow seas with high-land areas and associated deltas located along the present Appalachian Mountain trend (Figure 1).

Relative and absolute age dating

Methods for determining geological time intervals are not sufficiently refined to permit resolution of instants in time in any case. The technique of dating rocks by measuring ratios of radioactive parent-to-daughter isotopes always involves certain intrinsic sources of error and the time of formation of a rock is given as a range of years rather than as a unique point in time. For example, the age of the Palisades Basalt, which crops out along the Hudson River opposite New York City, is given as 195,000,000, plus or minus 5,000,000 years.

The second technique for dating rocks is by correlating the similarity of fossil remains. Differing evolutionary stages of organisms provide a useful chronology for establishing the time equivalence of rocks in separate parts of the world. Rates of evolutionary change, however, usually are not sufficiently great to discriminate events of less than several million years' duration. Moreover, fossils from different geographic areas may be similar not because they occupy the same time interval of evolutionary history but because they prefer the same life habitats. In fact, it is sometimes difficult to determine if two similar but geographically separate fossil assemblages represent the same time period or whether they are derived from the same environment (which may have existed at two different times). Are the oysters in sedimentary rocks of the Gulf Coast the same age as the oysters in sedimentary rocks of the Atlantic Coast, for example, or do they simply record similar nearshore, shallow, brackish ma-

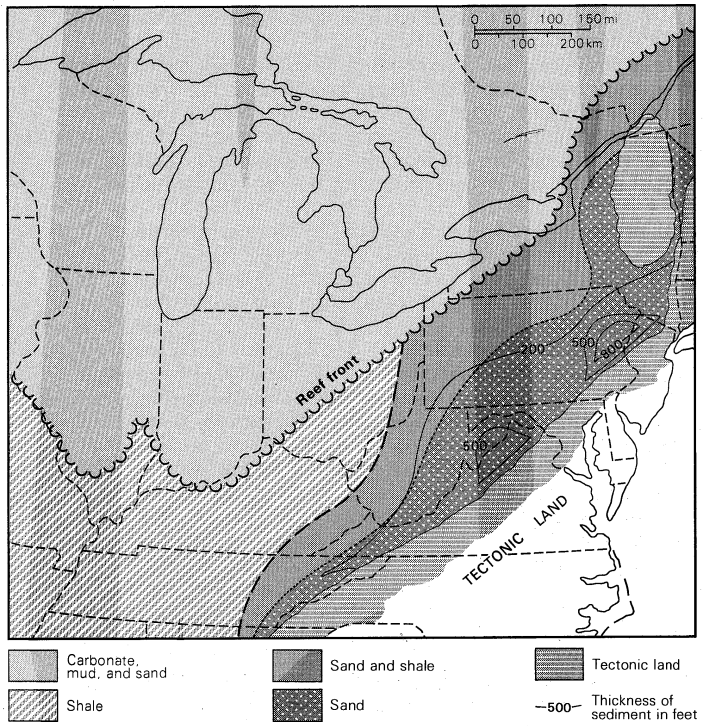


Figure 1: Paleogeographic structure of east central United States during Early Silurian time. Tectonic land raised during the Taconic orogeny provided a source for nonmarine and marine sandstones and shales, which were deposited along a broad shelf. Coral reefs formed along the shelf edge.

rine environments that existed at various times during the Tertiary Period in these two areas?

Paleomagnetic reversals

A relatively new technique for establishing instants in time is that provided by study of paleomagnetic reversals. At specific times in the past the polarity of the earth's magnetic field reversed and each reversal was essentially instantaneous. A sequence of reversals has been established for the last 4,000,000 years from the polarity and orientation of fine-grained iron oxides found in deep-sea sediments and terrestrial lava flows. There is some evidence that such reversals also occurred farther back in the geologic past. If used in conjunction with geochemical and paleontologic dating techniques, paleomagnetic reversals may ultimately provide the instant-time signatures so necessary for accurate reconstruction of the earth's geologic history.

Besides the importance of narrowing the time interval of a paleogeographic reconstruction to avoid blurring the results, it is also necessary that the geological data on which the reconstruction is based are, in fact, contemporaneous. Again, as previously mentioned, radiometric dating of rocks and paleontologic correlations cannot always guarantee the requisite time equivalency.

PALEOGEOGRAPHIC DATA AND THEIR APPLICATION

Biological information. The simplest kind of paleogeography is that showing the location of ancient lands and seas. The data which bear on such a reconstruction include the distribution of marine and nonmarine sedimentary rocks as defined by their fossil flora or fauna. Strata containing fossil corals or starfish, for example, would be interpreted as obviously marine, whereas rocks with the bones or teeth of hoofed herbivores would be considered definitely nonmarine or terrestrial. It should be pointed out that the present distribution of these rocks will only indicate the minimum limits of the former lands or seas, for it might be expected that erosion after the formation of the rocks would diminish their original areal extent.

Fossil organisms, however, can provide much more detailed information than whether their enclosing sedimentary rocks were deposited on land or in the sea. Terrestrial and marine environments include a wide range of habitats (*e.g.*, upland plains or plateaus, deserts, forests, river valleys, swamps and lakes, tidal flats, beaches, lagoons, reefs, deltas, or open ocean) in each of which there lives an assemblage of animals and plants especially adapted to the ecology of the particular habitat. Knowledge of the ecologic requirements of fossil organisms thus allows more refined interpretation of the paleogeography than simply designating "land" or "sea."

Fossil plants and their significance

Fossil land plants have long been used as climatic indicators in paleogeographic studies. In cases where the internal cellular structure is still preserved, it is possible to determine if the plant lived in an arid or wet climate and if there were marked seasonal changes in temperature or rainfall. In geologically younger rocks containing fossil plants that have living descendants one can define the past climate from the environmental preferences of the living form. For example, fossils of the genus *Artocarpus* (which includes the modern breadfruit plant, a native of the tropical Pacific) are found in Cretaceous rocks of Greenland; this clearly indicates the occurrence of a much warmer climate for that glacial land some 100,000,000 years ago.

More generally, however, it is the total assemblage of plant species that characterizes the local climate rather than any one single species. Thus, the assemblage of deciduous hardwoods, which includes maples, oaks, beeches, elms, and birches, occurs in temperate climates having warm summers, cold winters, and an average yearly rainfall of 92 centimetres (40 inches). Other assemblages occur from the northern tundra to the temperate grasslands and from the tropical desert to the rain forests. These individual assemblages parallel latitudinal gradients in temperature and precipitation; they also can be locally duplicated vertically as a consequence of variations in climate with altitude. Several fossil-plant assemblages, therefore, may suggest regional climate and local topographic elevation, thereby providing useful paleogeographic information.

Plants provide other fossil traces besides bits and pieces of leaves, stems, or fruits. In fact, most paleoclimatological interpretation is based upon the study of fossil spores and pollen, the microscopic reproductive bodies of plants that have extremely resistant outer coverings. Because of the great durability of spores and pollen, and their very great initial abundance (one flowering plant will produce more than 10^9 pollen grains), fossil spores and pollen can be obtained from many terrestrial or even marginal marine sedimentary deposits that may otherwise lack larger plant remains. Tracing spore and pollen assemblages across a broad geographic area or through a vertical sequence of rock strata will permit interpretation of areal or secular climatic variations.

Marine plants, especially the calcareous algae, are also found and recognized in the fossil state. Not only will they indicate the marine origin of the rocks in which they occur but because of the need for light for photosynthesis, the marine environment will be identified as being relatively shallow (solar radiation does not penetrate much beyond several hundred feet of seawater). One special group of marine algae, the soft-bodied, filamentous blue-green and green algae, thrives in tidal flat and very shallow marine waters. They form thin mats that trap and bind fine-grained sedimentary particles. Successive algal mats build up laminated sedimentary accumulations called stromatolites. Although the algae themselves are rarely preserved as fossils, their stromatolitic deposits are well-known. Because of their close association with the strandline environment, fossil stromatolites are extremely useful in delineating ancient shorelines.

Fossil vertebrates and land connections

Fossil vertebrates, whether fish, amphibians, reptiles, birds, or mammals, are also useful in paleogeography. Like plants, individual vertebrate species have characteristic adaptations to their environment so that their distribution in ancient rocks is a guide to the occurrence and extent of those environments. The presence of freshwater fish in the terrestrial rocks of Triassic age in eastern North America and those of Tertiary age in western North America provides evidence of inland freshwater

bodies in both of these regions during these geological periods.

One special value of fossil land vertebrates in paleogeography is their usefulness in defining migration routes across land bridges between continental landmasses. Fossil vertebrates from North and South America, for example, demonstrate that these two continents were connected by the Isthmus of Panama during early Tertiary time. During the Middle Tertiary the Panamanian land bridge became submerged and the two continents were separated for some 40,000,000 years. During this time the land mammals of South America evolved in isolation from the rest of the world, developing a number of types that paralleled those elsewhere. Toward the end of the Tertiary Period the vertebrate fossil record shows that the land bridge was again established and the faunas of the two Americas migrated northward and southward into the adjacent continental areas.

Similar evidence from land mammal fossils of the Pleistocene Epoch indicates periodic land connections and subsequent vertebrate migrations (including early man) between northeastern Asia and northwestern North America, across the Bering Strait. The presence of this land connection at different times also corroborates other independent evidence that the level of the world oceans periodically fluctuated several hundreds of feet with the waxing and waning of Pleistocene glaciers.

Some extremely interesting and significant finds of fossil terrestrial reptiles from Lower Triassic rocks of Antarctica that are practically identical to forms from South Africa have been used to substantiate the theory of drifting continents; in this instance, the paleogeography includes the former connection of Antarctica with Africa some 220,000,000 years ago. Furthermore, the climatic requirements of reptiles are such that the climate of Antarctica must surely have been much warmer then than now. Thus, the discovery of these reptile fossils indicates a very different paleogeography from that of today.

Marine invertebrates as environmental indicators

Shelly marine invertebrates, including microscopic calcareous and siliceous protists, sponges, corals, ectoproct bryozoans, brachiopods, mollusks, echinoderms, and arthropods, are used regularly by paleontologists to define environments within ancient seas. Foraminifera, for example, which are unicellular animals that secrete microscopic shells of calcium carbonate, include some forms that float in the upper surface waters of the sea. Upon death, these microscopic shells rain down on the sea floor where they are buried within the accumulating sediments. Certain varieties of the foraminiferans are restricted to water masses having a particular range of temperature and salinity. Examination of the foraminiferan assemblages in deep-sea cores raised from the oceans reveals significant water mass changes during the last 2,000,000 years. The changes correlate with continental evidence of glacial and interglacial stages during this time.

Other foraminiferans live directly on the sea floor rather than in the overlying water column, and different species are adapted to the varying conditions of temperature, salinity, sediment texture, and food content existing there. Since these ecologic conditions tend to vary more or less with changes in water depth, individual bottom-dwelling foraminiferan assemblages are used as indicators of depth when found in marine sedimentary rocks.

The larger shelly invertebrates are also bottom-dwelling forms for the most part. They, too, have certain environmental preferences with respect to water turbulence, grain size of the substrate, nutrients, temperature, salinity, and oxygen of the surrounding water. These characteristics of the sea floor and the immediately overlying waters are not, of course, everywhere the same but vary locally (from delta front and coastal marshes to the current-swept open shelf), as well as regionally (from the warm, shallow seas of tropical lagoons and reefs to the cold, iceberg-laden waters of the Arctic). Consequently, the composition of shelly marine assemblages varies systematically with changes in the marine environment.

Although not all assemblages are perfectly diagnostic of each different marine environment, certain marine assemblages are especially helpful in recognizing specific habitats. For example, coral-reef environments are characterized by a great diversity and abundance of shelly organisms, including massive, frame-building corals, encrusting calcareous algae, relatively large, bottom-dwelling foraminiferans, clams, snails, and sea urchins. Such reef environments are typical of warm, shallow waters protected from large run-off from the adjacent lands (silt and freshwater influx being unfavorable for good coral development). Usually the reefs lie next to low-lying land, or offshore, toward the edge of the continental shelf and marginal to the nutrient-rich waters of the open ocean. Clearly, recognition of an ancient reef tract would be extremely useful in reconstructing the paleogeography of an area.

Indirect organic information. Not all past evidence of life found in rocks is so direct as the presence of the hard skeletal remains of shelly invertebrates, calcareous algae, or plants. Indeed, much of the evidence is quite indirect. The occurrence of oil and gas whose organic constituents are derived from fossil marine life usually defines the origin of the containing strata as marine although, after formation in the source rock, oil and gas can migrate to the reservoir rocks whose origin might actually be nonmarine. Coprolites, which are the fossilized fecal remains of animals, can occasionally provide some paleogeographical data. For example, many marine sediments are pelleted by mud-ingesting bottom feeders, most of which are soft-bodied and therefore unpreserved as fossils. But the pellets may be preserved and there are many ancient limestones that are composed entirely of these pellets; such limestones are assumed to be marine. Human coprolites have been found in caves and their contents analyzed. Many ingested seeds, hair, feathers, and small animal bones within the coprolites may provide a reliable estimate of the paleoclimate of the time.

Other interesting indirect evidence of life is that of tracks, trails, and burrows made by ancient animals. For example, bottom-dwelling marine invertebrates living in the intertidal zone often make vertically directed burrows whereas those living in the subtidal parts of the sea make horizontally oriented ones. The purpose of burrowing vertically is to allow the organism to penetrate deeply enough in the sediment to escape the harmful effects of desiccation during exposure at low tide. The subtidal organisms, however, are continuously submerged and move parallel to the substrate, either on it or just below it, feeding on the accumulated organic matter. Geologists have recently come to realize the value of looking for vertical burrow traces in sedimentary rocks in order to define the position of the ancient strandline.

Biogeography and paleogeographic interpretations

From what has been discussed so far, it is clear that biogeography—that is, the geography of life—has great potential in interpreting and reconstructing the geography of the geologic past. Biogeography, however, can provide additional insights. For example, it is observed today that terrestrial and marine animal and plant communities are more diversified in the lower latitudes than toward the poles. This diversity, measured by the variety of species, is usually a maximum in the equatorial regions and decreases progressively poleward. Accompanying this decrease in variety is an increase in the abundance of individuals in the remaining species. This biogeographic trend has been studied in Permian marine rocks that occur on the present continents. The marine assemblages in these rocks have diversity trends that apparently parallel those of today, a conclusion that is contrary to current theories of continental drift, which claim significant changes in latitudinal position before drift (in the Permian) and after drift (today).

There are other biogeographic trends that have not, as yet, been as regularly applied to the fossil record for paleogeographic interpretation. They are the observations that the average size of a species tends to be smaller in warmer climates and larger in colder climates (Bergmann's rule, so called). A related observation is that the appendages or other protruding parts of a species are longer in warmer climates than in colder ones (Allen's rule). The biological significance of these two trends is

that heat is more easily conserved in organisms having a relatively higher body–mass to body–surface ratio.

Implicit throughout this discussion is the principle that the geography of life reflects the different distribution of organisms according to their adaptations to the various environments found across the Earth. Besides distributing themselves along ecological boundaries, organisms also are restricted in distribution by physical barriers, such as mountains, deserts, and seas. If the organisms were able to breach these barriers they often would adapt to the environmental conditions on the other side (consider the rapid spread of the English starling in North America after its transportation across the Atlantic Ocean). The barrier which separated South American mammals from the rest of the world during the Middle Tertiary Period is one example. Another barrier is that of the deep eastern Pacific Ocean, which prevents migration of shallow-water invertebrates from the Indo-Pacific to the shallow shores of the western Americas. The depths are too great for the adults, and the distances too large for the free-floating larval stages, for successful migration. In paleogeographic reconstructions, therefore, discontinuities among fossil assemblages may allow the inference of either ecological boundaries (regional climate or local environment) or physical barriers (water bodies, mountain ranges, deserts, etc.)

One final limitation of biological data used in paleogeography is that past biogeography cannot automatically be interpreted from the present distribution of life, for life has continuously evolved and changing adaptations may invalidate seemingly logical conclusions.

Petrological information. Paleogeographic data are not confined, of course, to fossil remains. Rocks of all sorts, fossiliferous or not, are used for paleogeographic reconstruction. Igneous, sedimentary, and metamorphic rocks are the products of a variety of processes occurring at the Earth's surface or within its crust and include phenomena as diverse as volcanism, erosion and deposition by streams and rivers, folding and faulting of strata, or the heating and recrystallization of earlier formed rocks. The environments in which rocks are made are distributed along geographic trends whether mountain belts, coastlines, earthquake zones, or volcanic-island arcs, and the composition, texture, and internal structures characteristic of these environments are often recorded within the individual rock types. Analysis of the rock record at a specific point in the geologic past allows definition and interpretation of these major rock-forming environments and indicates their geographic distribution.

Significance of sedimentary rocks

Sedimentary rocks are compacted and cemented mineral grains and rock fragments (with or without fossil debris) that have been eroded from older rocks and deposited by wind, water, or ice. The composition of the constituent grains varies as a function of the composition of the rocks in the source area from which the particles are initially eroded. The mineralogy of the sediments may thus reflect the mineralogy of the source. If the rates of erosion and sediment deposition are slow enough, however, the minerals may be broken down chemically and mechanically and only the most resistant (*e.g.*, quartz) will be left. Although the resulting mineralogy will differ considerably from the source rock, new paleogeographical clues may be obtained as to climatic conditions and relative rates of uplift in the source area and subsidence in the sedimentary basin. Sedimentary rocks may also contain grains formed directly within the area of sediment accumulation. Skeletal debris from organisms living in the depositional environment, for example, or chemical precipitates such as salt deposits or iron ores, may be incorporated with the transported sediments or even form entire strata by themselves. Such locally produced sediments will have their own paleogeographic implications.

The size of transported grains provides evidence of the strength of the transporting medium. The coarser the grains, the stronger the wind or water currents. Water-transported sediments will show decreasing grain size downstream, away from the source area, as the currents wane in strength near the area of deposition.

The roundness of individual grains as well as the size grades present provide a measure of erosional and depositional rates. Sediments with grains that are poorly sorted and angular in shape, with little or no rounding, are typical of areas undergoing relatively rapid erosion and sediment transport and deposition. There is little opportunity for size reduction through chemical or mechanical processes (solution or abrasion of grains); nor is there sufficient reworking by currents to size-sort the sediments.

The internal fabric, or geometry, of the grains within a sedimentary rock often contains environmental or geographic significance. The stratification, or layering, of the grains within the rock may be horizontal, inclined, or rippled, depending upon the character of the transporting currents. The direction of the inclination of stratification and the shape of the ripples will indicate current direction: the inclined layers and steep sides of the ripples point down current. Fine-grained sedimentary rocks that are water-laid will desiccate and shrink upon exposure to the air; the presence of mud-cracks in certain sedimentary rocks provides evidence, therefore, of periodic subaerial exposure during, perhaps, a drought (for terrestrial rocks) or a low tide (for marine rocks).

The analysis of individual sedimentary rocks, therefore, can be useful in paleogeography because it is often possible to identify the particular source area and its direction and distance from the depositional site, the nature of the transporting medium, the location of shorelines, and, perhaps, the general character of the climate. When these data are combined with knowledge of fossil assemblages, important and significant paleogeographic conclusions can be drawn (Figure 2).

By courtesy of the *Geological Society of America*

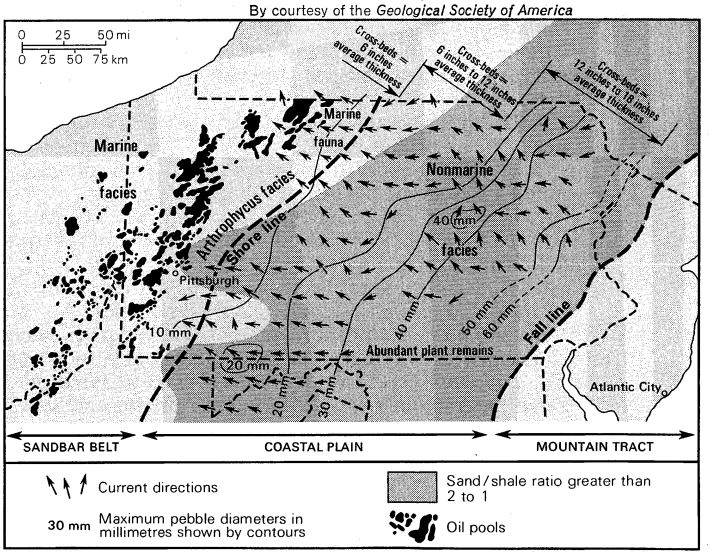


Figure 2: Paleogeography of the central Appalachians during the early Mississippian Period. Interpretation is based upon detailed analysis of the Pocono Sandstone Formation including grain size and composition, orientation of cross-stratification and plant remains, maximum size of quartz pebbles, sand/shale ratios, and fossil content. Sediment transport was to the west and northwest and the ancient shoreline trended northeast. The offshore parts of the Pocono are marine sandstones and shales with abundant burrows (*Arthrophycus*), and occasional marine fossils. Note relation of oil pools (black) to sediment type.

Igneous and metamorphic rocks

Igneous rocks are the cooled crystalline products of once-molten masses of viscous magma or lava that are derived from deep within the Earth's crust. These rocks may intrude overlying rocks higher in the crust or totally penetrate them and flow out upon the earth's surface. Igneous activity is usually associated with those parts of the Earth's crust that are subjected to stress. Around the rim of the Pacific today, for example, there are active zones of earthquakes and volcanoes within, or marginal to, geologically young mountain belts; the mid-oceanic ridges are interpreted as piles of submarine volcanic rocks formed during sea floor spreading and mid-oceanic rifting; and the exposed granitic cores of older moun-

tains were formerly buried deep within the crust during the subsidence and compression accompanying mountain building. Areas of the earth's crust that are structurally or tectonically quiet lack significant igneous activity.

The composition and grain size of igneous rocks are diverse and reflect many variables, including original composition of the molten material, rate of cooling, amount of water vapour and other dissolved gases, and depth of intrusion below the surface. Consequently, the paleogeographic significance of any specific igneous rock, beyond its recognition as a lava flow, granitic dike, or volcanic ash fall, is somewhat limited. This information alone, however, can indicate the presence of former mountain ranges, zones of tectonic activity, or areas of volcanism. The postulated existence of a mountain belt, for instance, will have paleogeographic implications, particularly because mountains influence climate. Climatic factors, in turn, affect the regional animal and plant communities and the volume of streamflow, rates of erosion, and other important variables. High mountain ranges provide large sources of erosional debris that build outward from the mountains as alluvial deposits and are carried down to the sea, where river deltas form. Offshore, the sediments will be distributed by longshore currents or slowly transported across the continental shelf into the deeper ocean basins.

Metamorphic rocks are the products of intense heat and pressure exerted upon pre-existing rocks or sediments so that new mineral assemblages recrystallize and new internal structures develop. Because the earth's crust is constantly undergoing stresses, the origins of which are still not entirely clear, earlier formed rocks are subjected to lateral or vertical increases in pressure and temperature. The minerals that initially had crystallized in equilibrium with the lower pressures and temperatures now become unstable, and new mineral suites recrystallize. Material also may be added or lost during metamorphism, which will alter the mineral assemblages still more. Like igneous rocks, metamorphic rocks can indicate former areas of regional deformation that are associated with mountain building or other movements of the earth's crust.

Because the focus of much modern effort is associated with the theory of continental drift, the value of igneous and metamorphic rock data in paleogeographic reconstruction has increased. Tectonic trends, linear-fold belts, metallogenic provinces, and exotic lithologies can be traced from one continent to another. These data provide evidence of the geologic and geographic configuration of the continents before drift occurred. The metamorphic trends and radiometric dates of South America and South Africa, for example, support a former joining of the two continental masses that accords to a fit suggested by other lines of evidence.

The dynamic activity of the earth's crust, as recorded by igneous and metamorphic rocks, is the fundamental mechanism underlying temporal changes in paleogeography. The uplift of mountains and the subsidence of sedimentary basins provide the topographic relief (and potential energy) by which other geologic phenomena occur; namely, changing patterns of sedimentation, shifting shorelines, varying climates, and migrating faunas and floras. It is these temporal variations in the distribution of land and sea, mountains and lowlands, and animals and plants that paleogeography attempts to depict (Figures 3 and 4).

Geophysical and geochemical information. Some of the most interesting problems in paleogeography relate to the question of whether the present distribution of the continents and ocean basins has remained more or less the same throughout geologic time. Opposed to the stability of continental position is the theory of continental drift, that the existing continents are fragmented portions of a once larger continent that have parted from each other and slowly moved into their present configuration. Recent revival of interest in continental drift was generated by the startling paleogeographic implications of geophysical measurements of paleomagnetism.

It has been observed that igneous and sedimentary rocks contain certain fine-grained minerals that become magnetized and aligned with the earth's magnetic field. When igneous rocks cool and crystallize, or sediments are compacted and lithified, the magnetic orientation of the minerals is "frozen" within the rocks. The individual grains act as small compass needles that not only point toward the poles but also have a dip from the horizontal that increases with latitude. Geophysicists have developed techniques for determining this remanent magnetic orientation in ancient rocks and this permits determination of the former geographic pole position and the latitude at the time a given rock formed. The interpretation is based on the assumption that the magnetic-pole positions approximate the position of the rotational, or geographic, poles and that the earth's magnetic field has been dipolar throughout geological time.

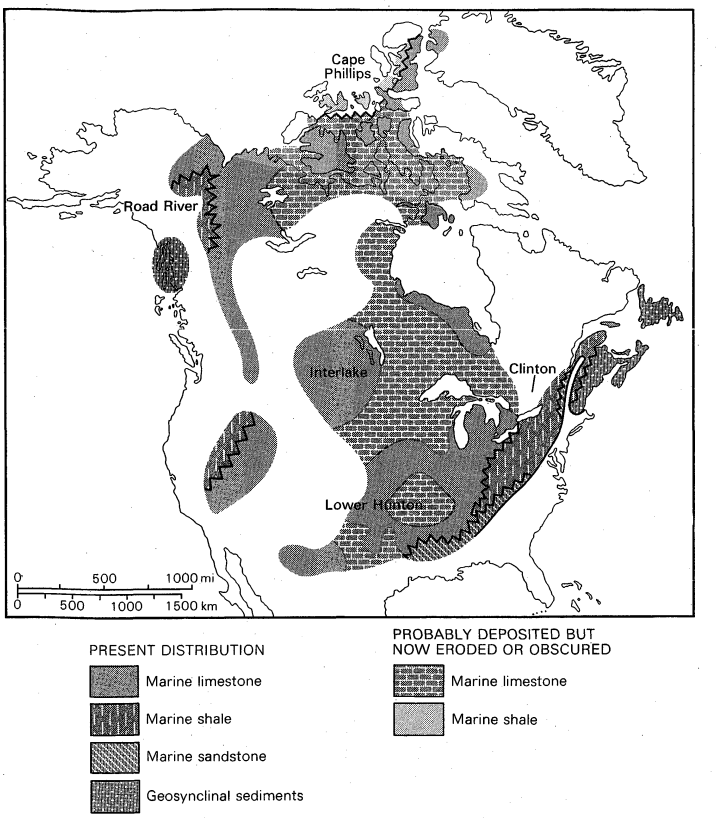


Figure 3: Paleogeography of North America during the Middle Silurian Period. Areas without rock pattern were either above sea level (as in southeastern U.S.) or lack definite paleogeographic indicators.
Adapted from T.H. Clark and C.W. Stearn, *Geological Evolution of North America*, 2nd ed. (© 1968); Ronald Press Company, New York

Polar wandering and continental drift

Paleomagnetic measurements have been made on many rocks of diverse ages from all the continents. It has been found that the magnetic and the geographic poles have wandered during the course of earth history, and that the observed path of polar migration, as recorded on an individual continent, becomes increasingly dissimilar with those of other continents going backward in time. Because there is only one pair of magnetic or rotational poles, the continents must have changed their spatial arrangement with respect to each other over the course of time. Paleogeographic reconstructions consistent with the paleomagnetic data, therefore, require some degree of continental drift. Although not all paleomagnetic results are fully consistent, this geophysical technique, together with measurements of paleomagnetic reversals, is extremely valuable for paleogeography.

Geochemical data are used to investigate paleogeographic interpretations based on the older and more traditional disciplines of paleontology, stratigraphy, and sedimentation. The chemistry of sediments, fossils, and rocks may record particular conditions of temperature, salinity, availability of water or carbon dioxide, or certain trace element abundances. These factors, in turn, may reflect past climates and environments. Although

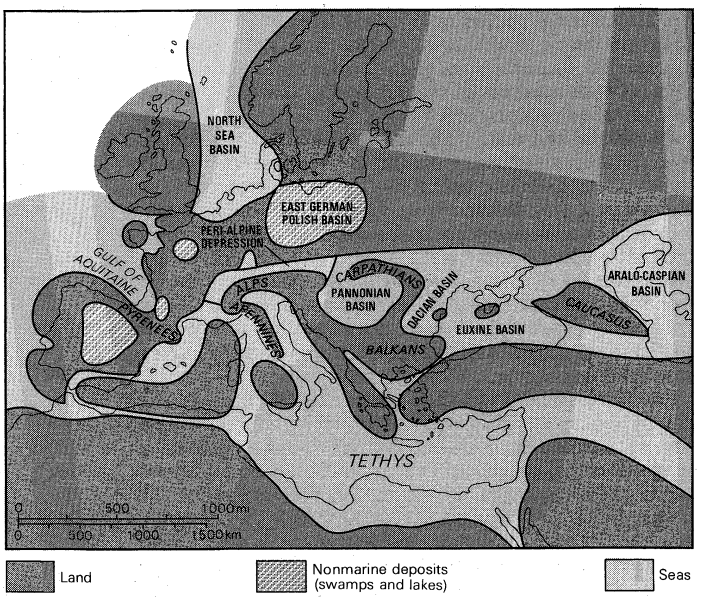


Figure 4: Distribution of land areas, nonmarine aquatic deposits, and marine sediments in Europe during the Miocene Epoch of Cenozoic Era. Complex arrangement of lands and surrounding basins in southern Europe reflects general tectonic instability of the area at that time.

Adapted from B. Kummel, *History of the Earth*, second edition, W.H. Freeman and Company; © 1970

sediments, fossils, and rocks may originally have been in delicate chemical equilibrium with their environments, their susceptibility to post-depositional alteration is very great and a large degree of uncertainty surrounds most paleogeographic conclusions that are based on geochemistry. Nevertheless, when used in conjunction with other independent evidence, geochemical techniques can make important contributions to paleogeography. The ratio of the stable oxygen isotopes O_{16} and O_{18} that occur in the shells of marine fossils, for example, is related to the temperature conditions at the time of their formation. This relationship has been used to document the cooling and warming of North Atlantic water masses during glacial and interglacial stages of the Pleistocene and the associated changes in oceanic circulation.

Oxygen isotope ratios

Trace element geochemistry has been used to determine the marine or freshwater origin of shales based on the relative abundances of gallium, rubidium, and boron. Empirically, it has been observed that freshwater shales have more gallium whereas marine shales contain relatively more rubidium and boron. Boron itself seems to have a special affinity for the clay mineral illite, and because boron in natural waters increases with overall salinity, some geologists have used boron-to-illite ratios as a measure of ancient salinities.

Other trace elements, particularly magnesium and strontium, have been observed in marine shells to vary directly in abundance with water temperature. Low-latitude, shallow-water, calcareous sediments, rich in the remains of shelly invertebrates, have relatively higher magnesium and strontium content than colder water sediments from higher latitudes or from deeper (and therefore colder) water at the same latitude. But this technique has proved difficult to apply to ancient shelly sediments because the mineral and chemical instability of most shell remains containing significant amounts of magnesium or strontium results in their very early post-depositional removal.

Red beds and weathering

The presence of "red bed" sedimentary rocks (sandstones and shales stained red by the iron-oxide mineral hematite) has been given paleoclimatological significance by a number of geologists. Although the details of formation may be disputed, there is a body of evidence suggesting that red beds accumulate in low latitudes (between 30° N and 30° S) and in relatively moist climates. Some red beds also may form in hot, dry climates as in the Recent alluvium of the Sonoran Desert of northwestern Mexico and southwestern United States. Thus, the ancient distribution of red beds, used in conjunction with other paleoclimatological data, may provide significant paleogeographic information.

A final example of how geochemical evidence can contribute to the interpretation of former climates and geography is that provided by the presence of bauxite in the geologic record. Bauxite is a rock rich in aluminum hydroxide that forms as a residual deposit under intense chemical weathering in tropical climates. Heavy rainfall, combined with the acidic conditions of tropical soils, causes the leaching of water-soluble ions and silica from the soil and weathered rock. The residual product, called bauxite, is an important aluminum-bearing ore and so has been extensively sought around the world. The occurrence of bauxite deposits helps to delineate the presence of tropical weathering conditions in the geologic past, and although these deposits are relatively thin and thus subject to removal by subsequent erosion, they are known to have formed as long ago as Cambrian time (500,000,000 to 570,000,000 years ago).

BIBLIOGRAPHY. T.H. CLARK and C.W. STEARN, *Geological Evolution of North America*, 2nd ed. (1968), a regional approach to the historical development of the major geologic structures of the North American continent; M. KAY and E.H. COLBERT, *Stratigraphy and Life History* (1965), on the principles of chronologically ordering and interpreting the physical and biological history of the earth, with emphasis on North America; B. KUMMEL, *History of the Earth: An Introduction to Historical Geology*, 2nd ed. (1970), similar to the previous works but also includes a discussion of other continents as well as North America; all three books have excellent paleographic maps and illustrations of past animal and plant habitats; L.F. LAPORTE, *Ancient Environments* (1968), a short introduction to paleoecology, or the environmental relationships of fossil organisms; A.G. SMITH and A. HALLAM, "The Fit of the Southern Continents," *Nature*, 225: 139–144 (1970), an interesting example of paleogeographic reconstruction using a variety of geologic data that supports current theories of continental drift; J.M. WELLER, *Stratigraphic Principles and Practices* (1960), a review of how earth history is interpreted from the geologic record of rocks and fossils, with many references.

(L.F.L.)

Paleography

Paleography (from Greek *palaios*, "old," and *graphein*, "to write") is the study of ancient and medieval handwriting. Precise boundaries for the study are hard to define. For example, epigraphy, the study of inscriptions cut on immovable objects for permanent public inspection, is related to paleography. Casual graffiti, sale or election notices as found on the walls of Pompeii, and Christian inscriptions in the Roman catacombs are likewise part of paleographical knowledge. In general, however, paleography embraces writing found principally on papyrus, parchment (vellum), and paper. Today, paleography is regarded as relating to Greek and Latin scripts with their derivatives, thus, as a rule, excluding Egyptian, Hebrew, and Middle and Far Eastern scripts. It is closely linked with diplomatic (*q.v.*), the study of forms in which official and private documents are drawn up.

The scientific study of Latin paleography (and diplomatic) dates from 1681, when the French monk Jean Mabillon published *De re diplomatica*, the first textbook on the subject, while his compatriot Bernard de Montfaucon performed a parallel service for Greek paleography in his *Palaeographia Graeca* in 1708.

The primary task of the paleographer is to read the writings of the past correctly and to assign a date and place of origin. Close acquaintance with the language of the text is a prerequisite. Help in dating is offered by changes in styles of handwriting and variations from area to area. Abbreviations in texts likewise help in dating and localization.

Writing materials. A paleographer must be familiar with writing materials. Any smooth surface able to accept writing has served in the past, notably, pottery fragments, animals' shoulder blades, slabs of wood, bark, cloth, and metal.

Papyrus

The great writing material of the ancient world was papyrus, in use by 3500 BC. In preparing the surface,

strips taken from the papyrus reed (*byblos*) growing in the Nile Delta were laid side by side, while other strips were laid across at right angles, and the whole impregnated with paste. After treatment a fine smooth surface was obtained. Much of the administration of the Roman Empire depended upon papyrus, in the same way that modern bureaucracies depend upon paper. Warfare and a damp climate resulted in an almost total disappearance of papyrus from Europe, though the dry (if war-scarred) sands of Egypt have preserved vast numbers of documents. Papyrus was imported into Europe from Egypt even after the fall of Rome. Chance survivals include charters of Merovingian kings in France (7th century) and business documents (5th–10th centuries) at Ravenna, the old administrative capital of the late Roman Empire.

The other great ancient writing material, still in occasional use today, is parchment, or vellum, the terms being often used interchangeably. Vellum is a term usually applied to skin from a calf (*cf.* veal, *veau*), while parchment is an expression often applied to sheepskin or goatskin. The word parchment is derived from Pergamum in Asia Minor, the ancient centre of its manufacture.

Both papyrus and parchment were expensive and were replaced for everyday use by wax tablets corresponding to today's notebook. Tablets made of wooden blocks were hollowed out and filled with melted, often black wax. Notes were made in the hardened surface. Even documents of permanent significance, such as property conveyances, were made on wax tablets.

Because ancient writing materials were expensive, they were often reused. Papyrus presented difficulties, for ink soon bonded itself firmly into the surface. Parchment could be more readily reused, because it is tougher and can be washed or scraped clean. Many medieval monks, when short of writing materials, took ancient books to pieces, cleaned off the leaves and used them again. The original script can often be brought out under ultraviolet light. Parchments thus cleaned and freshly inscribed are called palimpsests (Greek *palin*, "again"; *psēstos*, "scraped").

Paper is the third great writing material. In use in China at a remote period, it was employed extensively in the Arab world by the 9th century. Not in common use in Europe until the 14th century, it took over the name of the half-forgotten papyrus.

Form and binding of texts

In the early classical world the standard form of book was the papyrus roll, commonly called *biblion*, taking its name from the material of which it was made. It consisted of papyrus sheets pasted edge to edge with a slight overlap. The text was set out in columns, drawn up at right angles to the edge of the rolls, and started at the left. The reader unrolled as he went along and at the conclusion was obliged to reroll the book. The roll was an inconvenient form of book, difficult to consult, which probably accounts for the inaccurate quotations found in early literature, caused by an author relying on his memory rather than troubling to unwind a long roll. By the time of Christ, a new form of book was coming into fashion, the codex, or book in the shape in which it is known today. The codex is almost always of parchment, since papyrus cracks when folded. The codex seems first to have been used for notebooks or account books, the conservatism of booksellers and readers ensuring the survival of the roll for centuries. The Christians popularized the codex, using it for the Gospels.

Writing implements

Various instruments have been used for writing. The early Egyptians used a slender rush. From about 300 BC the thicker reed pen was used. The reed was in general use in the Romano-Greek world. Metal pens, copied from the reed, were also employed. For wax tablets a stylus was used, made of wood, bone, ivory, iron, or bronze. In many northern European areas, where reeds suitable for writing purposes are not indigenous, the feather (*penna*) became the main writing instrument. It was usually stripped of its vanes and the quill alone used.

Lead, used in classical times for ruling guidelines in manuscripts, was used extensively in the Middle Ages for rough notes and annotations in the margins of books.

Ink has been prepared in a variety of ways. In classical times the black discharge of cuttlefish was used, as well as concoctions of soot and gum. In the Middle Ages oak-apples were steeped in water with "vitriol" (ferrous sulphate) to produce ink.

Analysis of the text. The essential skill of a paleographer is the ability to recognize the numerous styles of handwriting prevalent in different ages and places. Most European scripts descend from Greek and Roman capital letters, but variations are enormous. It is a European convention that writing starts on the left at the top and works line by line down the page. An eccentricity known as boustrophedon (from Greek *boustrophēdon*, "following the ox furrow"), whereby alternate lines are written backward in mirror writing, occurs chiefly in very ancient inscriptions.

The Greek and Latin alphabets existed originally as capital, or majuscule, letters. The ancient Greek alphabet, as developed in chiselled inscriptions on stone or marble, served without much modification as the alphabet used in literary works written on papyrus rolls. This script, found in the oldest surviving Greek literary papyri of *c.* 300 BC or earlier, gave way to more rounded and elegant forms, probably developed in the Greek literary circles of Alexandria. Cursive scripts that were easier to write were developed for everyday use, for business, and to record the acts of the great bureaucracy of Egypt, where the Greeks settled in large numbers. The Greek cursive script and the formal book script greatly influenced each other, as can be seen from a vast series of cursive documents dating from the 4th century BC for about 1,000 years. Because so much material survived, early Greek cursive can be better studied than its Latin counterpart. In Greek cursive manuscripts the everyday life of ordinary people becomes a reality: they pay or fail to pay taxes, buy or sell houses, and harass civil servants with awkward demands.

A very rough division in Greek paleography may be made at around AD 300. The earlier age is called the papyrus period; and the later, the parchment or Byzantine (or Christian) period. The division, however, is imprecise, for parchment was used well before and papyrus long after this date. The change from papyrus to parchment is signalized by three great monuments of paleographical studies, the Vatican, Alexandrine, and Sinai Bibles, all on parchment and in codex form.

An alphabet of small, or minuscule, letters developed gradually and was in use by the 8th century. Numerous abbreviations exist in Greek manuscripts, though never so many as in Latin. Accents, an additional complexity, were not systematically applied before the 7th century AD.

Styles of writing

The ancient Latin alphabet of capitals (*quadrata*) is found in numberless inscriptions in stone and marble all over the Roman world. How far this alphabet was used for writing books is uncertain, because, though excellently adapted for incision, it is difficult to write. Some specimens of handwriting in *quadrata* do exist, such as 4th- or 5th-century copies of Virgil, but scholarly opinion largely regards these as abnormal productions. By the 1st century a handsome Latin alphabet existed, called rustic, based on the use of a broad pen or brush. Rustic was used for public inscriptions on walls, as in the sale and election notices found at Pompeii. Although specimens are scarce, it is likely that books were extensively written in this hand in classical times. By the 4th century another Latin alphabet existed, the script known as uncial, in the nature of a rounded form of *quadrata*. Uncial survived the fall of Rome and from it developed half-uncial, the ancestor of the small letters in use today.

The stately Roman scripts, *quadrata*, rustic, or uncial, were not used for everyday purposes, and, as in the case of Greek, a cursive, rapidly written hand arose in which letters and business documents were inscribed. This hand is found in graffiti on Pompeian walls and in wax tablets. After the disintegration of the empire, Roman cursive became the ancestor of regional hands in what are now Spain, France, and Italy.

During the flowering of Christianity and art in Ireland

(*c.* 500–*c.* 1000) a beautiful "insular" script developed, which found its way into England. There, two streams of influence commingled, for from 597 Christian missionaries arrived from Rome and brought in books in uncial script. Both scripts prospered in England, though insular gradually superseded uncial.

The most successful of all scripts proved to be Caroline minuscule, which takes its name from the emperor Charlemagne (died 814), patron of scholars and scribes, under whom the script was developed. Despite its inherent superiority and clarity, it did not predominate over regional scripts until the mid-12th century, and the local hand of southern Italy (Beneventan) maintained itself for much longer.

In the 12th century, Caroline minuscule, which had undergone moderate developments, started to display more obvious changes. It compressed laterally, while its rounded strokes became stiffer and straighter as it was converted into the so-called Gothic hands—very angular in northern Europe and more rounded in Italy. A revulsion against Gothic took place in scholarly circles in Italy in the 14th and 15th centuries, and a return to models based on Caroline minuscule took place. This revived hand, called Humanistic because humanist scholars used it, was adopted by 15th-century Italian printers, whose type faces ultimately triumphed over the Gothic. (This encyclopaedia is printed in a type scarcely modified from Caroline minuscule.) Meanwhile, Caroline and Gothic scripts had produced cursive hands for quick everyday use, as in the case of the ancient Greek and Latin alphabets. These cursive scripts were used for the vast mass of business documents written in the Middle Ages.

Cursive scripts

Abbreviations are the principal problem confronting paleographers. They were extensively used in Roman times by lawyers to avoid repetition of technical terms and formulas. Abbreviations fall into two classes, suspension and contraction. Suspension, omission of the end of a word and indication by a point or sign, was used extensively in Roman public inscriptions—*e.g.*, IMP.(ERATOR), CAES.(AR). Contraction, the omission of letters from the middle of a word and replacement by a sign or some other device, was common among Greek-speaking Jews, who contracted certain sacred or revered names, such as God, Lord, Israel, or David, as a mark of veneration. The Christians followed the practice by contracting their sacred names, such as Jesus and Christos. The great increase in use of abbreviation as a means of saving time and material dates from the 12th century, but some contraction signs are of high antiquity, such as the sign 7 for *et* (Latin, "and"). Some are quickly written versions of letter groups, such as "÷" for *est* (Latin, "is"), the top dot standing for e, the bottom dot for t, and the stroke being a long or f-shaped s, fallen on its side. The letter r is often omitted, the adjacent vowel being written above the line, as c^{a}ta for *carta* ("charter"). In works for semilearned readers, such as romances, abbreviations are often few, but books produced for the learned, such as university textbooks, are heavily loaded with abbreviations. The number of signs and devices in use by the end of the Middle Ages was enormous. More than 13,000 are listed in the standard work, Adriano Cappelli's *Lexicon Abbreviaturarum* (1912).

Abbreviations

Dating of books and documents also offers problems. Even when a precise date is given, the dating system of a given time and a given area must be checked because the year began at different times in different territories, and there were even variations in the same country. Calendar reforms initiated in 1582 by Pope Gregory XIII, for example, were not adopted in Protestant England until 1752. If a year of a monarch's reign is given as a date, it is necessary to determine whether the reign is counted from his accession or coronation. Moreover, few books were dated, the dated title page being nonexistent in medieval works, though sometimes a final paragraph, the colophon, supplies a date with the scribe's name and place of work. Important documents, such as English 12th-century royal charters, are undated.

In the absence of dates, inferences are drawn from handwriting, use of abbreviations, and internal evidence. Caution must be used, however; for an elderly scribe may be using a hand learned over half a century before, and work in the same scriptorium or office as a young clerk anxious to show off all the latest tricks and flourishes. Some styles lasted a long time: Caroline minuscule lasted for more than three centuries. Certain kinds of books, such as liturgical volumes, were produced in a highly stylized form for generations, and thus it is often difficult to provide a close date for a late medieval missal (with standard illustrations and marginal decorations) in a mechanical Gothic hand.

Internal evidence must be weighed carefully. A given historical event noted in a chronicle will provide an earliest possible date, unless the entry is an interpolation. Evidence for some legal practice or liturgical usage is no safe guide, for legislation on the subject by a king or a pope may merely be ratification of a long-standing practice.

A paleographer must get to know his scribes, for their mannerisms can be highly informative. Nearly 50 different scribes have been distinguished in the English royal chancery in the period 1100–89. First-rate scribes, such as notaries public, provide much information about themselves, giving their names, notarial signs, and information on their authority to act. Even anonymous clerks, who drew up innumerable property conveyances, can be identified by their script over a period of years, and their career can be traced through developing, mature, and deteriorating handwriting, thereby offering dating evidence.

Internal evidence

The provenance (origin) of many manuscripts can be immediately recognized because certain centres developed individual styles. Papal and royal chanceries issued documents of easily identifiable origin, while many monastic scriptoria—for example, that of Canterbury Cathedral in the earlier 12th century—had a virtually private handwriting.

Textual corruptions are another obstacle to correct elucidation. A legal document is certain to have been checked at the time of writing, but one cannot be sure in the case of a literary, philosophical, or theological text. Scribes were fallible, and, if there are no signs of any corrections in a text, then it probably embodies inaccuracies. A popular book, such as Chaucer's works, exists in large numbers of manuscripts, and many manuscripts produce variant readings. If a scribe made a mistake in copying, future scribes using his version are likely to reproduce the error and add others. Sometimes the same muddled passage in a group of manuscripts of a given author can be traced back to damage in an earlier copy, say a section eaten by rodents or impenetrably stained. Whenever copyists worked from different and faulty originals, various copies tend to fall into families. A paleographer must bring together various readings in families and decide which is the best reading.

Sometimes a scribe, set to work because he could write a fine hand, did not necessarily possess much knowledge of the language. Such a scribe faced with a text heavily loaded with abbreviations would usually make nonsense of it. Occasionally, a particularly stupid copyist, faced with a master copy in two columns of writing, would copy straight across the top line, then across the second. When he used a different number of words per line, the text was reduced to unintelligibility. In Greek and Roman times there was the difficulty that texts were written continuously, without space between the words. Copyists misread passages. For instance the historian Tacitus reported that some tribesmen went off to guard their own property: ADSVATVTANDA (*ad sua tutanda*). Some copyist thought "Suatutanda" was a place and this ghost name was perpetuated in geographical works. Later medieval Gothic hands presented a forest of vertical strokes called minims. The letter *v* rendered as *u* made two strokes, while *i* was often left without a dot or at best with a faint hairline, often misplaced. The group of letters *ium* could be read, as *uim, uiui, niui, mui, miu,* with many other variations. Accordingly, minim corruption, confusion of vertical strokes, is a term constantly heard in paleographical circles.

Latin and Greek are inflected languages in which the

same case and tense endings constantly occur, offering scope for error. Moreover, in biblical, theological, or philosophical texts, the same words abound. For example, in one place in the Gospel According to John there occurs the passage:

> Verba quae ego loquor
> vobis a me ipso non loquor
> pater autem in me manens . . .

("The words that I speak . . ."). The eye of a sleepy scribe might slip from the first *loquor* to the second, whereupon he would go on copying at *pater autem*, leaving out the second line altogether, a common type of error known as *homoioteleuton* ("like ending").

Because of the lack of surviving specimens, it is difficult to assess book decoration in classical times, but apparently it was very limited. In the later centuries of the Roman Empire, however, book illustrations were not infrequent.

Decorations

The narrative material in the Bible encouraged illustration. The Irish were foremost in applying decoration to the text in the form of elaboration of capital letters, producing such masterpieces as the Book of Kells (late 7th century), in which Celtic imagination and artistic sense ran riot in elevating the book to an object of outstanding beauty. Some of the greatest creative talent of the Middle Ages was lavished upon books, especially upon those used in worship, such as Bibles, psalters, and missals. When a book cannot be assigned either a date or provenance upon the appearance of the text alone, its style of illumination will often direct the paleographer to a certain monastery in which the carving on capitals or wall paintings may contain the same motifs.

Because of the immensely high prices of manuscripts, the question of forgery naturally arises, but it is safe to say that no modern forgery could survive for a moment. A convincing imitation of ancient script is virtually impossible, while the papyrus, parchment, or paper on which it would be written could not stand up to modern scientific inspection. Anything of recent vegetable or animal origin fluoresces brightly under ultraviolet light, to name but one test. William Henry Ireland (died 1835), the Shakespeare forger, used flyleaves from 16th-century books, but his handwriting and non-Shakespearean language gave him away. A modern would-be forger must either copy an existing work, which, in the present state of art history and paleographical study, would be immediately recognized, or be prepared to invent medieval subject matter.

Forgeries

There was fabrication of documents in medieval times on a considerable scale. A monastery might find itself in possession of estates held since remote antiquity but without any title deeds. When some powerful monarch made difficulties, there was a strong inducement to produce the required ancient-looking documents. The borderline between justifying legitimate possession and culpable attempts to gain extra territory or privileges, however, is ill-defined. Monks occasionally descended to falsifications of title deeds and charters of exemption. About 1125 a monk of Soissons on his deathbed confessed to a career of professional forgery for gain and admitted fabricating charters for various monasteries, including Westminster Abbey. Early forgeries, however, give themselves away through such inconsistencies as mentioning bishops of nonexistent sees or embodying legal phrases that came into use generations later or bearing seals when seals were not yet appended to documents.

The modern paleographer has great technical aids: photography since the 19th century and colour photography in the 20th. Ultraviolet light brings out faded handwriting. Microfilm makes the contents of a whole volume in a far-distant repository available quickly and cheaply.

BIBLIOGRAPHY. The most comprehensive work, certainly in English, is still E. MAUNDE THOMPSON, *An Introduction to Greek and Latin Palaeography* (1912). Also valuable, and in print, is his shorter version, *Handbook of Greek and Latin Palaeography* (1893, reprinted 1966). For Greek paleography, see B.A. VAN GRONINGEN, *Short Manual of Greek Palaeography*, 3rd rev. ed. (1963); and C.H. ROBERTS, *Greek Literary Hands, 350 B.C.–A.D. 400* (1956). The article "Handwriting" in C.G. CRUMP and E.F. JACOB (eds.), *The Legacy of the Middle Ages* (1926), is an important account of Latin book hands. Essential for the study of abbreviations in medieval Europe is ADRIANO CAPPELLI, *Lexicon abbreviaturarum: dizionario di abbreviature latine ed italiane*, 6th ed. (1961). A standard textbook for medieval English cursive hands is CHARLES JOHNSON and HILARY JENKINSON, *English Court Hand A.D. 1066 to 1500* (1915), continued in Jenkinson's *Later Court Hands in England from the Fifteenth to the Seventeenth Century* (1927). L.C. HECTOR, *The Handwriting of English Documents*, 2nd ed. (1966), contains a valuable introduction to the paleography of English administrative manuscripts. Very many classical, biblical, and liturgical texts have now been published in facsimile, often with colour plates. Large collections in facsimile comprise E.A. LOWE, *Codices latini antiquiores*, 12 vol. (1934–71); CHARLES SAMARAN and ROBERT MARICHAL, *Catalogue des manuscrits en écriture latine, portant des indications de date* (1959–); BERTRAM COLGRAVE (ed.), *Early English Manuscripts in Facsimile* (1951–); and the "Oxford Palaeographical Handbooks," a series designed to deal with various aspects of the subject, such as C.H. ROBERTS (see above); C.E. WRIGHT, *English Vernacular Hands from the Twelfth to the Fifteenth Centuries* (1960); M.B. PARKES, *English Cursive Book Hands, 1250–1500* (1969); and T.A.M. BISHOP, *English Caroline Minuscule* (1971).

(W.G.U.)

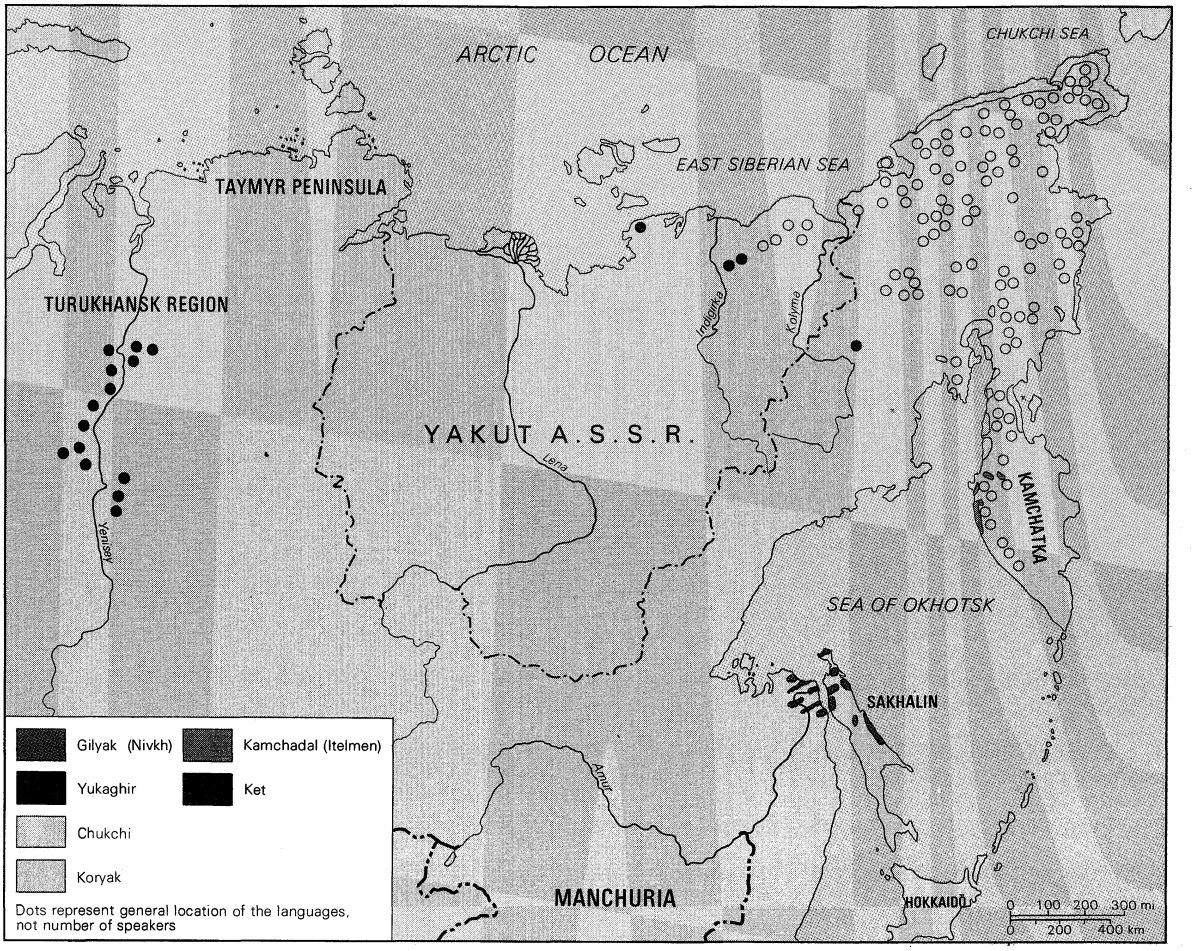
Paleosiberian Languages

The collective term Paleosiberian is applied to four genetically unrelated language groups situated in northern Asia—Yeniseian, Luorawetlan (Luoravetlan), Yukaghir (Yukagir), and Gilyak. The Yeniseian group, whose only living member is Ket (or Yenisey-Ostyak), is spoken by about 1,000 persons in the Turukhansk region along the Yenisey River. Kott (Kot), Arin, and Assan (Asan), now extinct members of this group, were spoken to the south of the present-day locus of Ket. The Luorawetlan family consists of (1) Chukchi, spoken by 12,000 people in the northeasternmost parts of Siberia, west of the small enclave of Siberian Eskimo; (2) Koryak, also called Nymylan, with approximately 8,000 speakers, found to the south of Chukchi; (3) the more remotely related Kamchadal (or Itelmen), with a bare remnant of 350 speakers in southern Kamchatka; (4) Aliutor, perhaps a Koryak dialect, with a small and unknown number of speakers; and (5) Kerek, with about 100 speakers. Some Soviet scholars list Aliutor and Kerek as being more closely related to Chukchi and Koryak than to Kamchadal.

The Yukaghir group, whose only living member is Yukaghir proper (or Odul), is spoken by about 450 persons in two enclaves in the Yakut A.S.S.R. (Yukutskaya A.S.S.R.), near the estuary of the Indigirka and along the bend of the Kolyma River. Extinct languages belonging to the Yukaghir group (or perhaps dialects of an earlier form of Yukaghir proper) are Omok and Chuvan (Chuvantsy); these were spoken south and southwest of Yukaghir proper. Gilyak (or Nivkh) has about 4,000 speakers, 2,500 of whom live in the estuary of the river Amur and 1,500 on the island of Sakhalin.

Classification of the languages. These four groups are *not* related to each other. They have been subsumed under the names Paleosiberian, Paleo-Asiatic, or, more rarely, Hyperborean, ever since the Baltic German zoologist and explorer Leopold von Schrenck surmised, in the middle of the 19th century, that they constituted the remnants of a formerly more widely dispersed language family that had been encroached upon by invading groups of Uralic, Turkic, and Tungus speakers. Schrenck's hypothesis is quite correct to the extent that as recently as the 17th century Yeniseian, Luorawetlan, and Yukaghir languages were spoken over much wider territories than they are today. For example, it is known that Samoyed languages (of the Uralic family) at one time in the past absorbed the languages of now extinct Yeniseian tribes, that Yukaghir was spoken as far west as the Taymyr Peninsula in the 17th century, and that the former domains of Chukchi and Koryak extended much further to the west. Little is known about the prehistory of Gilyak but it may be assumed that this language was also originally centred further to the west, perhaps in Manchuria. As far as can be determined with the help of methods of comparative linguistics, however, the four

Paleosiberian groups not genetically related



Distribution of Paleosiberian languages.

present-day Paleosiberian groups never formed a single family of languages in the accepted sense of that term. In fact, they may represent only a fragment of a possibly greater diversity of language families in prehistoric Siberia. Many of the languages spoken in the area during earlier periods may have been swallowed up by the more recent and culturally vigorous intruders in Siberia that are now the neighbours of the Paleosiberian enclaves: mainly the Yakut (whose domains stretch as far as the Chukchi and Yukaghir areas) and various Tungus tribes (one or another of which borders on each Paleosiberian language).

Comparisons with other language families

Nevertheless, many attempts have been made to show that the four Paleosiberian families are either related to each other or to adjacent (or more distant) language families. Thus, Ket has been compared with the Sino-Tibetan family (Chinese, Tibetan) and with some of the languages of the Caucasus, and Yukaghir has been compared with Uralic. Some of these comparisons are fanciful experiments or completely unfounded (*e.g.*, the comparisons of Ket with Caucasian languages). Others are more reasonable but not compelling (*e.g.*, the Yukaghir-Uralic hypothesis) because the evidence adduced so far has been too unsystematic and fragmentary. It is therefore safest, at present, to consider Ket, Yukaghir, and Gilyak as language isolates that are unrelated to any known language or language family and to regard Luorawetlan as a family in its own right that is also unrelated to any other family or isolated language. Further research on the internal history of these four groups will eventually enable comparatists to reconstruct hypothetical earlier stages of the languages and to make comparisons with non-Paleosiberian languages more cogent, if at all feasible. Mere resemblances in grammatical or phonological traits between Paleosiberian and adjacent languages (such as between Chukchi and Eskimo, or betweeen Gilyak and Korean or Japanese) are not indexes of genetic affinity, but are often the result of the diffusion of linguistic traits over large geographical areas. They may, however, provide clues to the linguistic prehistory of Siberia.

The cultures of the four Paleosiberian groups are, in general, similar in that they are all Arctic or subarctic. In detail, however, each group of speakers of a Paleosiberian language has its own characteristic cultural profile. These characteristics may on occasion even resemble the cultural profile of a non-Paleosiberian group very closely; *e.g.*, Ket culture resembles Selkup (Ostyak-Samoyed) culture more closely than it resembles that of any Paleosiberian group. (Selkup- and Ket-speaking groups are located in contiguous areas.)

Linguistic characteristics. *Grammar.* The grammatical structures of the four Paleosiberian groups differ considerably from each other. In a broad sense, Gilyak resembles Japanese in its grammatical categories and processes (in the word order, heavy inflection of verbs, and use of enclitics—words closely connected with the word that precedes), whereas Yukaghir shares certain grammatical categories with some Uralic languages (the objective conjugation—*e.g.*, "he shot it" versus "he shot"—and the negative conjugation—*e.g.*, Yukaghir *tet mer-ai-mek* "you shot" versus *tet el-ai-yek* "you did not shoot"). A typical feature of Luorawetlan is its strong tendency toward complex compounding, also called incorporation, and circumfixation; *e.g.*, in Chukchi *ga+mor-ïk+tor+orw-ïma* "in our new sleigh," the entire unit is surrounded by the circumfix *ga- . . . ïma* "in" (compare *ga+mor-ïk+orw-ïma* "in our sleigh," without *tor* "new," and *ga+tor+orw-ïma* "in the new sleigh," without *mor-ïk* "our"). A characteristic feature of the Ket verb is its succinct com-

Typological resemblances between Gilyak and Japanese

plexity, involving such categories as gender, animateness, and type of event; *e.g.*, *t-k-it-n-a* "I carved it up," which consists of *t-* "I," the verbal complex *k- . . . -a* "cut up (carve, split) into pieces once," *-it-* (feminine object marker "her, it"), and *-n-* (past-completed tense). All of the Paleosiberian languages are rich in devices for compounding words. In syntax, Luorawetlan favours ergative constructions in which markers indicate the agent or instrument of the action; *e.g.*, *Father* + agent marker, *bear* (subject), *shoot* (main verb), "Father is shooting a bear."

Phonology. Typical phonological features of the Paleosiberian languages are post-velar consonants (*i.e.*, sounds formed further back in the mouth than *k*, usually represented as *q*), vowel harmony of various kinds (*e.g.*, the alternation of *e* and *i* in the form for "my" in Gilyak *ñe-řla* "my harpoon" and *ñi-řly* "my sky"), consonant alternations (*e.g.*, the alternation between *b*, *v*, and *f* in Gilyak *bal* "mountain," *ñ-val* "my mountain," *c-fal* "thy mountain"), and rich consonant clusters in all but Yukaghir.

Vocabulary. In addition to the stock of native words inherited from its ancestral language, each Paleosiberian language also has numerous loanwords, some of which are recent and from adjacent or recently adjacent languages, and others of which are ancient, from languages with which it no longer has contact. Some of the loanwords from ancient times are, of course, more difficult to identify. In general, Tungus, a branch of the Altaic family, is the source of most loanwords, but the Turkic languages (including Yakut) have also served as the sources of loans, and Ket has some words from Selkup. There are also more complicated loan relationships, such as are found in the reindeer terminology of Gilyak, which is borrowed from a Tungus language but seemingly not from any of the Tungus languages with which Gilyak is now in contact. South Sakhalin Gilyak also contains a considerable number of loanwords from Ainu (a language of northern Japan) and was, during 1905–45, hospitable to potential loans from Japanese; the Japanese loanwords never became acculturated because the Japanese hegemony over South Sakhalin ceased after World War II. Chukchi has some Eskimo loans.

Loanwords in Paleosiberian languages

The most viable source of technical and all of the other modern vocabulary has been the Russian language, the influence of which began with the first contact and continues to be strong. Each Paleosiberian language adjusts the Russian loans according to the dictates of its phonology and grammar but the most recent borrowed words tend to retain their original Russian form or one closely resembling it.

Writing. The Yukaghir had a tradition of pictographic writing (incisions on fresh birch bark) used by men for route maps and by young women for a type of love letter. Limited use of such a system among the Koryak speakers has also been reported.

Since the 1920s and 1930s each Paleosiberian language has had a literary language and a script now based on the Cyrillic alphabet (and formerly based on the Latin script). Because at one time these native languages were used in part in elementary education, primers and arithmetic books for the lowest grades were available. Some natives continue their education and acquire a good knowledge of Russian and of Soviet culture. This has led to the rise of bilingualism but has also contributed to the growth of modern literatures in the native languages, based on Russian models, especially among the Koryak and the Chukchi.

The native traditions and folklore of the Paleosiberian peoples have been collected since the last century, mainly by Russians and Westerners. Work in these fields is still continuing and is attracting a slowly emerging corps of trained native specialists. Such trained natives are also beginning to collaborate in the compilation of dictionaries.

BIBLIOGRAPHY. ROMAN JAKOBSON, GERTA HUTTL-WORTH, and JOHN FRED BEEBE, *Paleosiberian Peoples and Languages: A Bibliographical Guide* (1957), very useful, with informative appendix; DEAN WORTH, "Paleosiberian," in T.A. SEBEOK (ed.), *Current Trends in Linguistics*, vol. 1, *Soviet and East European Linguistics*, pp. 345–373 (1963), good coverage of Soviet work since World War II. The best and most recent source on Paleosiberian languages is in Russian: П.Я. Скорик *et al.* (eds.), *Языки народов С.С.С.Р.*, vol. 5, *Монгольские, тунгусоманьчжурские и палеоазиатские языки* (1968).

(R.Au.)

Paleozoic Era, Lower

The Lower Paleozoic Era refers to the segment of Earth history from 395,000,000 to 570,000,000 years ago and includes the Cambrian, Ordovician, and Silurian periods. It is characterized by the evolution and marked preponderance of marine invertebrate faunas during this time interval, whereas terrestrial forms of life flourished in the Upper Paleozoic. This giant step from the seas to the land areas is the greatest biological revolution of the Paleozoic Era and is a reasonable basis for its division.

This article treats the rocks, life-forms, and environments of the Lower Paleozoic Era. The historical subdivision of Lower Paleozoic rock systems, with attendant nomenclature, is covered in the separate articles on the geological periods involved—*i.e.*, CAMBRIAN PERIOD; ORDOVICIAN PERIOD; and SILURIAN PERIOD. The other portion of this era is treated in PALEOZOIC ERA, UPPER. The interested reader should also consult STRATIGRAPHIC BOUNDARIES; GEOLOGICAL TIME SCALE; and, for an overview of all of geological time, EARTH, GEOLOGICAL HISTORY OF. The life forms of the Lower Paleozoic are given additional detail in the article FOSSIL RECORD.

LOWER PALEOZOIC ROCKS

World distribution. The trilobite *Asaphus boliviensis* and other fossils were reported from South America as early as 1842, and Cambrian fossils were described from eastern Asia and from South Australia in 1883 and 1884, respectively. Lower Paleozoic rocks are now known to be distributed extensively in all continents, although their extent is rather limited in the Southern Hemisphere.

In the Caledonian (Scotland), Appalachian (United States), Sayan-Altai (Soviet Union), Tasman (Australia), and other geosynclines (downward flexures of the Earth's crust), Lower Paleozoic sediments accumulated to great thicknesses. The seas extensively flooded the Canadian, Angara, and other cratonic areas (stable areas of the Earth's crust that generally form the central mass of a continent). Rocks of the ancient Precambrian basement are now extensively exposed in the Canadian and Baltic shields, but the thin Lower Paleozoic veneer largely mantles the basement of the central Siberian Platform, leaving only the southern border of the Angara Shield and the Anabar area in the north exposed. These large cratons and geosynclines already were differentiated before the deposition of the Sinian, Beltian, and other late Precambrian formations. From Korea and China to the Tarim and Fergana basins of Central Asia, cratonic and quasi-cratonic areas are sutured by various folded mountains. There the distribution of Lower Paleozoic formations and their thicknesses and interrelationships are highly complicated. The stratigraphy also is complex in central Europe and the Iberian Peninsula; Ordovician fossils are the oldest in the Alps. Little is known of the basement of the festoon islands of the western Pacific Ocean, but Middle Cambrian and Ordovician fossils have been discovered in New Zealand, and Silurian fossils occur in Japan and in West Irian, New Guinea.

Lower Paleozoic formations are tremendously thick in the geosynclines, attaining several thousands of metres or more in northeast Kazakhstan (Soviet Union) and elsewhere. The rocks are chiefly composed of terrigenous (land-derived) materials and various volcanic and pyroclastic sequences. Graywackes (sandstones containing fragments of other rocks and a clay-rich matrix) and greenstones (altered volcanics) are common members. Radiolarian chert, graptolitic shales (graptolites are extinct colonial organisms), and some limestone also occur. The characteristics of these deposits change horizontally as well as vertically, indicating the instability of the trough area. Conglomerates and arkose sandstones also

record the complicated geology of their hinterland source areas.

Compared with the geosynclinal sediment accumulation, the cratonic veneer is usually thin. It is composed of shallow marine sediments rich in various benthonic (bottom-dwelling) organisms. Fossiliferous carbonate rocks are common in Scandinavia, for example, and a reef facies is represented in the Lower Cambrian by the archaeocyathid (primitive marine invertebrates that resemble sponges but that may not be related to them) limestone. The coralline reefs become important sediment types in the Upper Ordovician and were well developed in the Niagaran of the Great Lakes and other Silurian formations. *Cryptozoon* (algal?) limestones are present in several Cambrian and Ordovician sections in North America. Limestone conglomerates are common in the Upper Cambrian Chaumitien Limestone of eastern Asia.

Sedimentary environments and ore deposits. In Lower Paleozoic sequences, conglomerate generally is rare, and shale is more common than sandstone. The St. Peter Sandstone, which is composed of water- and wind-worn quartz sand, is astonishingly persistent in the Mississippi Valley, showing how slowly and widely the Middle Ordovician sea flooded the almost level land. Some of the sand is pure enough to use in glass manufacture.

Oil shales, uranium, iron, and salt

The Middle and Upper Cambrian Alum Shale Group of southern Sweden includes oil shales containing lenses of uranium-bearing, bituminous nodules. Uranium deposits also are known from the Lower Cambrian of the Estonian S.S.R. and from the Silurian of Fergana, Central Asia.

Black carbonaceous shales of Ordovician age, containing extraordinary numbers of graptolites, are thought to represent a pelagic sediment—*i.e.*, one containing fossils of free-swimming or floating organisms—deposited in poorly ventilated basins in which no benthos could survive. The Middle Cambrian Burgess Shale of the Canadian Rockies, which yields a unique planktonic (floating) fauna, may have formed in a similar environment.

Oolitic (consisting of cemented small round grains) iron ores are common in the Ordovician formations in Bohemia (Czechoslovakia), Thuringia (East Germany), Bretagne (France), and the northern part of the Iberian Peninsula. The Wabana Series iron ore is found in the Lower Ordovician in Newfoundland; and Clinton Iron Ores, in the Middle Silurian of the Appalachian Mountains from New York to Alabama. Of interest to geochemists endeavouring to determine the history of seawater was the discovery of salt-crystal casts in the lowermost Cambrian (?) Spiral Creek Formation, eastern Greenland, below the *Olenellus*-bearing Bastin Formation. Evaporites, including salt and gypsum, were deposited throughout the Lower Paleozoic—in the Cambrian of Morocco and Iran; the Cambrian, Ordovician, and Silurian of central Siberia; the Ordovician of the Shansi Basin, northern China; and the Upper Silurian Salina Group of central New York.

LOWER PALEOZOIC LIFE

Three important steps in the history of the development of the biosphere occurred in the Lower Paleozoic. The Cambrian animals were exclusively marine, and all major invertebrate phyla were represented. The Ordovician animals also were aquatic; most classes of fossil invertebrates and many important orders in the Paleozoic fauna were included. Furthermore, bony armour plates of primitive jawless fishes have been discovered in Middle Ordovician rocks at three localities of western North America. The oldest fish tooth found is that of *Palaeodus* from the Lower Ordovician marine glauconite sand of the Estonian S.S.R. A third evolutionary step occurred during Silurian time. Corals and other reef builders developed rapidly in the Silurian, but the scorpions *Palaeophonus* and *Proscorpius* from the Upper Silurian of Sweden, if not themselves air breathers, are clearly ancestors of air-breathing scorpions; this reflects the dawn of animal life on land.

Overview of invertebrates and evolutionary trends. As far as the fossil record is concerned, life seemingly developed very suddenly during the transition between the Cryptozoic and Phanerozoic aeons (*i.e.*, at the Cambrian–Precambrian boundary, 570,000,000 years ago), although this interval may have been longer than a transition between any two eras. All of post-Precambrian time is presumably less than one-fifth the time span of the Cryptozoic. Precambrian fossils consist principally of ambiguous problematica or inorganic pseudo-fossils. The Ediacara coelenterate and annelid (worm) fauna is an exception because of its excellent preservation. It is so distinct from later faunas that its age is considered older than Cambrian.

The next oldest fossils are Lower Cambrian archaeocyathids and the small, straight-coned *Volborthella, Hyolithes,* and other primitive invertebrate shells. Trilobites (extinct arthropods whose external skeleton is divided into three longitudinal lobes) are not always associated but are well represented in the next younger faunas. Three Lower Cambrian faunal provinces (regions characterized by the occurrence of specific fossils or fossil assemblages) have been distinguished, each of which is based on the occurrences of trilobites: the Olenellidian fauna of the Arcto-American–Atlantic province; the Redlichian fauna of eastern and southern Asia and Australia; and the intermediate fauna of Siberia and the Mediterranean region, in which redlichid and olenellid types of trilobites are coexistent and protolenid types are common.

Trilobites and faunal realms

Trilobites were the dominant group in the Cambrian fauna; in witness to this fact, two-thirds of the named Cambrian genera are trilobites. Next in abundance were the brachiopods, shelled marine invertebrates in which the two valves, or shells, are of unequal size. About 10 to 15 percent of the fauna would include the archaeocyathids, the arthropods exclusive of trilobites, and other minor groups. Chitinous (chitin is a nitrogenous substance similar to mammalian fingernails and claws) and horny phosphatic shells are especially common among Cambrian animals. The Middle Cambrian Burgess fauna of the Canadian Rockies, containing coelenterate—*i.e.*, without hard parts, such as jellyfish—medusae (*Peytoia*), annelids (*Canadia, Ottoia*), the oncopod *Peripatus*-like *Aysheaia,* the arrowworm *Sagitta*-like *Amiskwia,* and other rarities, is undeniable evidence for the contention that the fossil fauna is only a part of past life and that pelagic animals are very poorly represented in the fossil record.

Graptolites, cephalopods, brachiopods, and trilobites were four important groups that subsequently profused, in the Ordovician fauna. The dendroid (branching) graptolites probably appeared in Middle Cambrian time, and they then flourished to become the most valuable index fossils for interprovincial correlations of the Ordovician record. The most primitive cephalopod (excluding *Salterella*) is the late Upper Cambrian *Plectronoceras* in China. The successive development of nautiloid groups can be seen in the Ordovician Tsinan Limestone of northern China and southern Manchuria. Post-Cambrian trilobite branches appeared suddenly in the Cambro-Ordovician transitional times when many Cambrian branches died out. The shell-bearing brachiopods flourished greatly during and after the Ordovician. Four rock series of the Silurian of the Oslo region, Norway, are based on brachiopod ranges.

Rise of bryozoans, echinoderms, and reef builders

Bryozoans (marine colonial animals) that appeared in the Upper Cambrian became important members of the Ordovician fauna. Among the mollusks, the earliest gastropods appeared in the Lower Cambrian, and pelecypods became common in the Ordovician. The ostracods (small bivalved crustaceans) and eurypterids (very large, extinct arthropods) also appeared at this time, and the echinoderms, marine invertebrates with five-fold symmetry, developed greatly in the Ordovician. Stromatopoods and tabulate corals became important reef builders during the latter part of the Ordovician and in the Silurian.

Among the Lower Paleozoic animals to die out and become extinct was the phylum Archaeocyatha; some brachiopods, cephalopods, and arthropods also disappeared. The ammonites (a large group of extinct mollusks), insects, amphibians, and higher vertebrates had yet to appear.

Table 1: Correlation of the Cambrian System Based on Trilobite Zones

Trilobite stage	Europe and North Africa	eastern Asia		Siberia Kazakhstan	North America	
Olenidian	*Acerocare* Zone *Peltura* Zone *Leptoplastus* Zone *Parabolina spinulosus* Zone *Olenus* Zone *Agnostus pistiformis* Zone	Chaumitian	Fenshanian	Shidertinian	Croixian	Trempealeauan
			Daizanian			Franconian
			Paishanian	Tuorian		Dresbachian
Paradoxidian	*Paradoxides forchhammeri* Zone *Paradoxides paradoxissimus* Zone *Paradoxides oelandicus* Zone	Fuchouan	Kushanian	Maya	Albertan	
			Taitzuan			
			Tangshihan	Agma		
Olenellidian	Issafenien*	Mantoan	Lungwangmian	Lena	Waucoban	
			Tsanglangpu			
	Soussien*		Chiungchussu	Aldan		
			Yulutsun			

*In North Africa.

The plant record. Within the plant kingdom, *Girvanella* occurs in the Lower Cambrian of China and elsewhere, and various kinds of algal remains have been described from the Cambrian of Siberia and Canada. Red and green algae were greatly developed in the Ordovician; brown algae became common in the Silurian.

Spores that presumably were derived from mosses or ferns have been reported from the Cambrian of the Baltic region and from the Urals. *Aldanophyton* has been discovered in the Cambrian of Siberia; and *Biophyton*, in the Ordovician of Bohemia. *Baragwanathia* and *Saxonia* from the Upper Silurian of Victoria, Australia, and Saxony, Germany, respectively, are considered to be lycopods. *Zosterophyllum*, a member of the psilopsids, appeared in the Silurian. The Lower Paleozoic plant record, however, is meagre in comparison with the extraordinarily developed Devonian land flora.

STRATIGRAPHIC CORRELATION

Trilobite zones and guide fossils

The Cambrian System. The Cambrian System is divided into zones chiefly on the basis of the trilobites that are contained in the strata (see Table 1). In Scandinavia, the Olenid Series is divided into eight zones in addition to 31 lesser units called subzones, and the *Paradoxides* Series is divided into three stages and nine zones. The Olenellid Series is best divided in the Anti-Atlas Mountains, Morocco, where eight zones and 18 horizons (stratigraphic layers of little thickness) of trilobites were recognized, in addition to an archaeocyathid limestone below them. This indicates the great accuracy of Cambrian time correlation, which is provided with about 60 key horizons in the European Mediterranean region. Assuming that the duration of the Cambrian Period was 75,000,000 years (it is actually nearer to 70,000,000 years in duration), its zone time is a little longer than 1,000,000 years (zone time is the duration of a period divided by the maximum number of zones or key horizons that exist in the corresponding rock system). This average provides some concept of the time scale relative to other periods (see STRATIGRAPHIC BOUNDARIES).

Some trilobite species or genera are distributed widely, but most are provincial. *Centropleura* is a late Paradoxidian genus known in northwestern Europe, Siberia and adjacent islands, Australia, and in Nevada and Vermont in North America. Thus its distribution is almost worldwide, although its occurrence is uncommon at these places: *Glyptagnostus reticulatus* is a better guide fossil for an early Upper Cambrian horizon because it is distributed widely in Europe, northern and eastern Asia, Australia, and western, southern, and eastern North America. This is the oldest global zone that is defined by the life range of a species.

In North America, the Lower Cambrian is divisible into four zones, the Middle Cambrian into nine zones, including four subzones, and the Upper Cambrian into 13 zones, including six subzones. Nearly all of the index fossils are trilobites, but the lowest zone is the *Obolella* brachiopod Zone in the Appalachian-type section. The Upper Cambrian, or Croixian, which typically is exposed in the upper Mississippi Valley, is divided into the Dresbachian, Franconian, and Trempealeauan stages.

In eastern Asia, the best display of the Mantoan Series is seen in Yunnan, South China, where ten zones, including the *Hyolithes* Zone at the base and the archaeocyathid zone at the top, are distinguished. In North and Northeast China, the Fuchouan and Chaumitian series are each divided into three stages, and 17 and 11 zones are distinguished in them, respectively. The Cambrian zone time is about 2,000,000 years for eastern Asia and 2,500,000 years for North America.

Much work remains to be done before adequate interprovincial correlation can be achieved, but the Chaumitian base may be a little higher and the Croixian base lower than that of the Olenidian Series. It still is questionable whether or not the top of the Trempealeauan is older than the base of the Tremadocian (Ordovician).

The Ordovician and Silurian systems. In the Ordovician and Silurian systems, graptolite and shelly facies are distinguished. In Great Britain, 15 graptolite zones are recognized in the Ordovician system. In Victoria, Australia, 18 zones are found, ranging from Tremadocian to Llanvirnian (see Table 2), whereas only eight zones are known in the equivalent formation in Great Britain. The Silurian system contains 21 zones in Great Britain, 22 zones in the Soviet Union, and 42 zones in Bohemia (see Table 3). The Ordovician zone time is about 2,500,000 years. The shortest zone time of the Lower Paleozoic Era is 850,000 years for the Silurian zone time of Bohemia, assuming the Silurian Period to be 35,000,000 years (the Silurian Period began 430,000,000 years ago and ended 395,000,000 years ago).

Graptolite zones and zone time

Dictyonema flabelliforme, *Glyptograptus teretiusculus*, *Nemagraptus gracilis*, and *Pristiograptus nilssoni* are, for instance, cosmopolitan species, but many others are regional or local. The *Didymograptus bifudus* Zone is known from Europe and North America, but the zone in the Marathon region of west Texas may be older than that of Great Britain.

The intercalation (alternation of layers) of graptolite shales in the sequence of shelly facies, or the reverse, as seen in New York state and central China, is of great value for the biostratigraphic correlation because it provides evidence of the time equivalence of fossils. The Cambrian and Tremadocian faunas of the Acado-Baltic province are sharply separated from the North American ones. The Tremadocian is related to the Arcto-North Atlantic faunas on one side and the Andine faunas on the other in Cambro-Ordovician time. The Durness Limestone of North Scotland and the Hecla Hoek Formation of Spitsbergen and Bjørnøya (Bear) Island, Norway,

Table 2: Correlation of the Ordovician System Based on Graptolite and Shelly Facies

Great Britain	zones	Balto-Scandinavia	Bohemia		Siberia Kazakhstan	South China		Southeast Australia	North America		
Ashgillian	*Dicellograptus anceps* *Dicellograptus complanatus*	Harjuan	*Tretaspis*	Kosov Králův Dvůr (Zidice)	Chokpar	Chientang-kiangian	Yuchin	Bolindian	Cincinnatian		Gamachian Richmondian Maysvillian Edenian
Caradocian	*Pleurograptus linearis* *Dicranograptus clingani* *Climacograptus wilsoni* *Climacograptus peltifer* *Nemagraptus gracilis*	Viruan	*Chasmops*	Bohdalee Loděnice Chrustenice (Nučice)	Dolborsk Mangazeisk		Huangnikang	Eastonian Gisbornian	Champlainian	Mohawkian	Trentonian Black Riveran
Llandeilian	*Glyptograptus teretiusculus*		*Ogygiocaris*	Drabov	Krivolusk	Neichia-shanian	Yenwashan Hulo				
Llanvirnian	*Didymograptus murchisoni* *Didymograptus bifudus*			Šárka				Darriwilian Yapeenian			Chazyan
Arenigian	*Didymograptus hirundo* *Didymograptus extensus* *Dichograptus*	Oelandian	*Asaphus*	Klabava	Chunisk	Ichangian	Ningkuo	Castlemainian Chewtonian Bendingonian	Canadian		Cassinian Jeffersonian Demingian Gasconadian
Tremadocian	*Bryograptus kjerulfi* *Dictyonema sociale*		*Ceratopyge*	Milina Třenice (Krušna)	Ustuisk		Yinchufu	Lancefieldian			

yield various fossils of North American affinities. Trilobites and other fossils of Alaska and the Canadian Rockies are important links in correlating the Cordilleran with the Siberian and Korean–Chinese faunas.

The Tien Shan and Himalayan troughs were two important routes of migration. The Ordovician trilobites of eastern Asia show faunal connection with those of northern and southern Europe through the former and latter route, respectively. The eastern Asian fauna also was connected with the Andine fauna through both the Mediterranean and the Australian and New Zealand faunas.

LOWER PALEOZOIC ENVIRONMENTS

Distribution of land and sea. Epeirogenic movements (uplifts and downwarps of the Earth's crust over wide regions) took place at different times among the three northern megacratons (the shield areas of North America, Europe, and Asia), as did orogenic disturbances (more intensive vertical movements of the crust, associated with mountain building) in the various geosynclines. The Lower Cambrian sea generally was transgressive, and marine inundation was common during Middle Cambrian time. Laurentia, the entire northern landmass, was generally emergent until the Upper Cambrian transgression, at which time, however, the sea covered its southern half. In the Siberian Platform, on the other hand, the Cambrian sequence is complete except for the uppermost part, thus indicating that the area was probably above sea level. In the Baltic region, the Cambrian strata thin toward the east, and the basal Ordovician sandstone directly overlies the Lower Cambrian *Fucoid* and *Eophyton* sandstones in the Estonian S.S.R. The *Eophyton* sandstone is disconformably underlain by the blue mud (marine sediment) of Leningrad, which, on the contrary, becomes thinner toward the west from the Russian Platform.

Transgressive–regressive sequences in the Ordovician and Silurian

The Ordovician sequence is unbroken in the Baltic region, and, in the interior lowland of North America, the Canadian Series is widely underlain by the Croixian. The Middle Ordovician limestones lie above the basal St. Peter Sandstone. The Richmondian sea flooded the continent, and the Cordilleran sea became confluent with the Arctic sea. In the Siberian Platform, the Tremadocian Ustuisk Stage is transgressive, and the Arenigian Chunisk Stage is regressive. The Krivolusk Stage is transgressive again; a local unconformity exists at the base of the Mangazeisk in the southwest. The regressive Dolborsk Stage contains gypsum deposits. The sea retreated toward the Arctic at the end of the Caradocian, and the Silurian overlies the Ordovician transgressively.

The Lower Silurian in the interior lowland of North America overlies the Ordovician, and the Middle Silurian overlaps the Lower Silurian extensively. A large shallow sea with many coral reefs existed there. The salt-bearing Upper Silurian red and gray formations in central New York reveal a regressive facies that may represent an arid period of time. Like the Ordovician, the Silurian sediments of the Baltic region constitute a continuous sequence that terminates with the Downtonian sandstone facies at the top. In similar fashion, the Silurian marine sediments are well represented in the Siberian Platform and are thin and free from volcanic material. The Upper Ludlow, however, is lagoonal and saliferous.

Mountain building and volcanism. Generally speaking, the Cambrian was a quiet period, but, in the Tasman Geosyncline in eastern Australia and Tasmania, strong volcanism occasionally occurred during the Cambrian and later periods. The Tyennan–Jukesian orogeny took place in Tasmania in the Middle and Upper Cambrian; the Benambran movements in the Upper Ordovician and the Bowning movements near the end of the Silurian Period occurred in Victoria and New South Wales.

Crustal deformation in Asia

The Lower Paleozoic formations in the Sayan–Kazakhstan region are tremendously thick and are composed chiefly of terrigenous and volcanic materials. The Upper Cambrian and late Middle Cambrian formations are found in limited areas in the Sayan–Altai region. There, crustal movements were repeated from late Middle Cambrian to Ordovician time, the so-called Salair movements in the Kuznetsk Basin. Influenced by these deformations, the Ordovician sequence and facies vary greatly among the Kuznetsk Alatau, Chorie, and Salair mountains.

Geosynclinal sediments are traceable to the east into the Tien Shan, Nan Shan, and the Tsinling Shan. Near the end of the Silurian Period, there was a strong crustal deformation called the Kwangsi movement in South China and also in Nan Shan. Silurian volcanism is recorded in the Chichibu Geosyncline of Japan.

In Korea and North China, the Cambro-Ordovician Chosen Korean Group is underlain by the Sinian and overlain by the Permo-Carboniferous. The Silurian is absent (if Silurian-derived fossils in North Korea are excluded). The three Lower Paleozoic systems are, however, well represented in central and South China.

Events in Europe and North America

In Europe, the Paradoxidian of the Pribam Syncline in central Bohemia is underlain by a thick formation of sandstone and conglomerate; it is overlain by a volcanic and pyroclastic formation, and this, in turn, is overlain by the Tremadocian. The Lower Paleozoic systems are well developed in the Montagne Noire, southern France, where the Tremadocian is concordant with the Arenigian. In Sardinia, on the other hand, the Arenigian transgresses the folded Cambrian formation. The Arenigian lies on the older rocks in the Armorican Massif, northwestern France, and the Anglesey Island, Wales.

Table 3: Correlation of the Silurian System Based on Graptolite and Shelly Facies

Great Britain	zones	Norway	Sweden		Bohemia		Soviet Union	China	North America	
Ludlowian	Downtonian	Ringerike Series	Oved-Ramsåsa Series		Budňany	Přídoli	*Monograptus* Beds	Shamo Series	Cayugan	Keyser Limestone
	Saetograptus leitwardinensis *Pristiograptus tumescens* *Monograptus scanicus* *Pristiograptus nilssoni* *Pristiograptus vulgaris*	Upper Spiriferid Series	*Colonus* Series			Kopanina	*Clonograptus* Beds			Tonoloway Limestone Salina Group
Wenlockian	*Cryptograptus lundgreni* *Cryptograptus ellesae* *Cryptograptus linnarssoni* *Cryptograptus rigidus* *Monograptus riccartonensis* *Cryptograptus murchisoni*	Lower Spiriferid Series	*Cryptograptus* Series	*Flemingi* Beds	Liten	Motol	*Monograptus flemingi* Beds	Lojaping Series	Niagaran	Guelph Dolomite Lockport Dolomite
				Retiolites Beds			*Retiolites* Beds			Clinton Ores
Llandoverian	*Monoclimacis crenulata* *Monoclimacis griestonensis* *Streptograptus crispus* *Spirograptus turriculatus* *Monograptus sedgiwcki* *Demirastrites convolutus* *Pristiograptus gregarius* *Pristiograptus cyphus* *Orthograptus vesiculosus* *Akidograptus acuminatus* *Glyptograptus persculptus*	*Pentamerus* Series	*Rastrites* Series			Zelkovice	*Rastrites* Beds	Lungmachi Series	Albion	Tuscarora Sandstone
		Stricklandia Series					*Akidograptus* Beds			

In the British Isles, the intrageosynclinal volcanism was violent during the Ordovician Period. The sea expanded in the Caradocian toward the south, and in the Ashgillian it expanded toward the north. The period closed with an extensive emergence of the lands. Then the Silurian transgression was renewed, but volcanism almost had ceased. Crustal movement was intensified toward the end of the period. In Norway, also, volcanism occurred repeatedly in the Caledonian Geosyncline in the Ordovician Period. There, five disturbances are distinguished: namely, pre-Arenigian Trondheim; post-Caradocian Ekne; post-Ashgillian Horg; Wenlockian; and Ardennian. As a result, the arcuate Caledonian range extending from Ireland to Spitsbergen, and probably farther to the north, was produced.

In the latter part of the Ordovician Period, the rising of Appalachia supplied sediments to form the Queenston Delta in New York and Pennsylvania. The Taconic disturbance culminated toward the end of the period in dislocations in the region between Quebec and Albany, New York. In the Lower Silurian Period, the northern Appalachian Mountains supplied terrigenous material to the western lowland, where the Shawangunk Conglomerate and Tuscarora Sandstone accumulated. In North America, the Late Paleozoic Era closed quietly, whereas the Caledonian disturbance was strong in northwestern Europe and some other geosynclinal zones.

Climate. Evaporites, such as the salt and gypsum in Lower Paleozoic rocks, indicate an age of relative aridity. The worldwide distribution of Lower Paleozoic limestones and dolomites, and particularly the Middle Silurian reef limestones, show that warm seas were widespread. Long before the development of hermatypic corals (warm-water types), archaeocyathid limestones were deposited in all continents but South America, including Antarctica, in the Lower Cambrian and probably during early Middle Cambrian. The optimum temperature for this extinct phylum is assumed to have been no less than 25° C (77° F). There is no record of a frigid climate in the Lower Paleozoic Era. The glaciations previously reported to be of Cambrian age actually were late Precambrian in age.

BIBLIOGRAPHY. K.A. VON ZITTEL, *History of Geology and Paleontology to the End of the Nineteenth Century*, trans. by M.M. OGILVIE-GORDON (1901; orig. pub. in German, 1899), the standard reference to the history of this subject; C.O. DUNBAR, *Historical Geology*, 2nd ed. (1960), one of the best textbooks for North America; M. GIGNOUX, *Géologie stratigraphique*, 4th ed. (1950; Eng. trans., *Stratigraphic Geology*, 1955); R. BRINCKMANN, *Geologic Evolution of Europe* (1960; pub. orig. in German, 1956), two of the best textbooks for Europe; N.A. BELIAEVSKY *et al.* (eds.), *Structure géologique de l'U.R.S.S.* (1959; orig. pub. in Russian, 1958); D.A. BROWN *et al.*, *The Geological Evolution of Australia and New Zealand* (1968); and F. TAKAI *et al.* (eds.), *Geology of Japan* (1963), good references to the U.S.S.R., Australia, New Zealand, and Japan; COMMISSION DE STRATIGRAPHIE, CONGRES GEOLOGIQUE INTERNATIONAL, *Lexique stratigraphie international* (1957–), detailed accounts of stratigraphic terms in different countries.

Advanced references to the Lower Paleozoic include: J. RODGERS (ed.), *El sistema cámbrico, su paleogeografía y el problema de su base*, 20th International Geological Congress, Mexico, 3 vol. (1956–61), the most comprehensive publication on the Cambrian system of the world and its base; *Report of the Twenty-First Session, Norden*, International Geological Congress, pt. 7, "Ordovician and Silurian Stratigraphy and Correlation," and pt. 8, "Late Pre-Cambrian and Cambrian Stratigraphy" (1960), many important papers on the late Precambrian and Cambrian stratigraphy and the Ordovician and Silurian stratigraphy; *Report of the Twenty-Third Session*, Czechoslovakia, International Geological Congress, proceedings of sect. 9, "Stratigraphy of Central Europe in Lower Paleozoic" (1968), valuable papers on the Older Paleozoic formation of Europe and the intercontinental correlations; and H.K. ERBEN (ed.), *Internationale Arbeitstagung über die Silur/Devon-Grenze und die Stratigraphie von Silur und Devon* (1962), many papers discussing the Silurian and Devonian systems and their boundary problem.

Articles on the correlations of the Cambrian, Ordovician, and Silurian formations of North America may be found in the *Bull. Geol. Soc. Am.*, 55:993–1004 (1944), 65:247–298 (1954), and 53:533–538 (1942).

(T.K.)

Paleozoic Era, Upper

The Upper Paleozoic Era refers to the segment of Earth history lasting from 395,000,000 to 225,000,000 years ago. The Paleozoic Era was first proposed by Adam Sedgwick in 1838 for the rocks and life of the Cambrian through Silurian systems (395,000,000–570,000,000 years ago), the term meaning time of ancient life. Later, the Paleozoic Era was extended to include the Devonian, Carboniferous, and Permian systems as well, and these systems are informally called the Upper Paleozoic, as distinct from the Lower Paleozoic. The era ends before the Triassic Period of the Mesozoic Era. The Upper Paleozoic Era is distinguished from the lower division chiefly by its more advanced life forms; woody plants and vertebrates were prominent for the first time, trilobites and graptolites declined, and there was an upsurge of brachiopods and ammonites among marine organisms.

Upper Paleozoic rocks are a most important source of economic minerals. Much of the mining conducted by the Phoenicians and Romans in such places as Spain and Cornwall, England, derived copper, gold, and other mineral ores from Devonian volcanics. A similar source has provided copper, lead, and zinc since AD 900 in the Rhine region, and in nearby Thuringia the science of geology was considerably advanced with respect to mining the copper and salt deposits of Permian age. Both the Devonian and Permian were periods of widespread mineralization in most continents. Carboniferous coal has been important in the development of Western technology. Petroleum also is abundant in Upper Paleozoic rocks, especially in the U.S., the U.S.S.R., and Canada. The first well drilled deliberately for oil, in 1859 at Titusville, Pennsylvania, was in Devonian sands.

Knowledge of the Upper Paleozoic Era has increased enormously in the past decade or so, partly because of renewed interest in the theory of continental drift (*q.v.*). Evidently the Earth has been remarkably mobile throughout geological time. Such movement in the Upper Paleozoic caused periodic crises that affected climate and life forms and the distribution of lands and seas.

This article treats the Upper Paleozoic Era as a whole, tracing the development of the Earth and its diverse life forms and climatic regimes throughout this time interval. For additional detail on each of the time periods encompassed see DEVONIAN PERIOD; CARBONIFEROUS PERIOD, LOWER; CARBONIFEROUS PERIOD, UPPER; and PERMIAN PERIOD. See also FOSSIL RECORD; STRATIGRAPHIC BOUNDARIES.

UPPER PALEOZOIC ROCKS

Over the Precambrian shields, Upper Paleozoic rocks formed thin glacial and terrestrial deposits in peninsular India, Africa, and Antarctica, or thin shelf deposits on the Siberian Platform and parts of the Canadian Shield (see CONTINENTS, DEVELOPMENT OF). The thickest deposits, largely marine, occur in geosynclines. One of these major belts, the Cordilleran, passes along the west coast and Rocky Mountains of North America; the second, much larger belt, passes east–west, across southern Europe, through south Asia into the South Pacific islands of Indonesia, New Guinea, New Caledonia, and New Zealand (Figure 1). This second area is called the Tethyan Geosynclinal Complex, named after the Tethys or ancient Mediterranean Sea. Three major geosynclines also were present in South America from Venezuela to Patagonia; these were linked with North America and possibly with the Tethys as well. Other areas of thick marine deposition were the Uralian Geosyncline of the Soviet Union, the east Australian geosyncline, and some large geosynclines in Siberia and the American Arctic.

Shelf deposits, as a rule thin and laid down in very shallow waters, also were widespread over the central United States and Soviet Union. They lie above Lower Paleozoic deposits that rest in turn on Precambrian rocks. Thousands of feet of sediment also were deposited in basins located near the edge of continental shields, as in western Australia and western Canada.

Miogeosynclines and eugeosynclines

The greatest volume of Upper Paleozoic rock occurs as thick piles of volcanic lavas and sediment largely derived from the lavas. The piles accumulated under the sea in rapidly sinking troughs, termed eugeosynclines.

The lavas, rich in iron, magnesium, and calcium, and moderately poor in quartz, poured out from fissures under the sea or from narrow chains of volcanoes, which emerged from the sea as island arcs (*q.v.*). The accompanying sediments were rapidly deposited graywackes (*q.v.*) and mudstones eroded from the volcanoes and the narrow, locally upheaved parts of the geosyncline or its borders. Fossils are rare. Beds generally were deposited in deep water from great clouds of sediment carried in density currents (*q.v.*) originating on the continental slopes.

Sediments formed in miogeosynclines closer to the ancient shields also were thick but lacked large quantities of volcanic material. They are composed largely of detritus, carried from the shields, or of carbonates, derived from broken marine shells and algae. Such miogeosynclinal

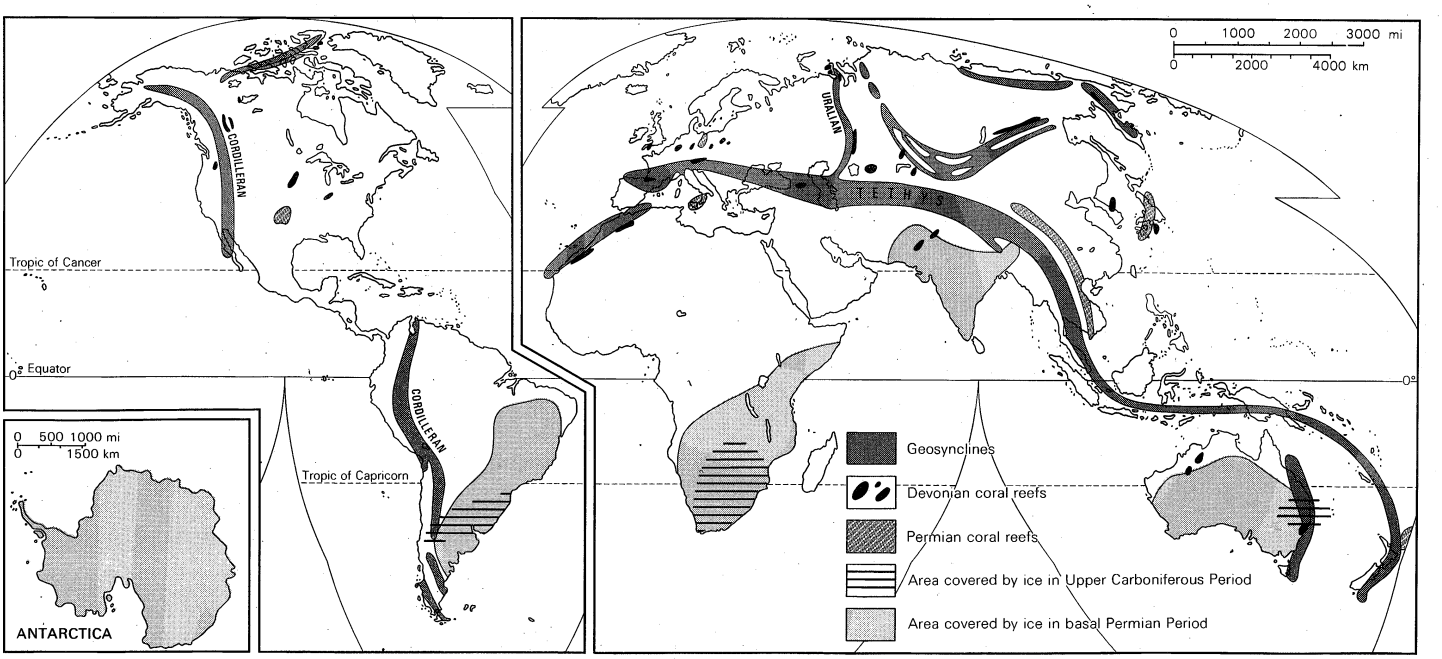


Figure 1: General paleogeography during the Upper Paleozoic Era, showing locations of geosynclines, coral reefs, and glacial deposits.

deposits are found along the eastern Rocky Mountains and through Europe north and south of the eugeosyncline.

Sediments accumulating on shelf regions, in basins, and on shields were mainly carbonates or lagoonal deposits. They are thin and extensive, often crammed with fossil shells. Whereas geosynclinal deposits for a single Upper Paleozoic period may total 3,000 to 18,200 metres (10,000 to 59,700 feet), shelf deposits that represent similar time intervals are only 300 to 1,800 metres (1,000 to 5,900 feet) thick. Frequently the shelves were emergent as low coastal plains or were submerged by only a few tens of feet as sea level changed or continents rose and fell.

The Upper Paleozoic is characterized by its diversity of sedimentary rock types, which fall into five categories (one formed on land, the other four, generally just above or below sea level): terrestrial desert deposits; marine salt deposits and coral reefs; coal measures; and glacial deposits. These special kinds of deposition are discussed below.

Deposition in arid regions

Pioneer geologists in Europe and eastern North America first distinguished the Upper Paleozoic by its red sandstones, which they divided into "Old Red Sandstone" and "New Red Sandstone." The sandstones commonly are coloured red by the oxidation of iron compounds. The sand grains are well rounded like those of present-day sand dunes. Associated beds consist of silt, with fish and plant remains. Cobbles and pebbles formed thick beds locally. Such rock groups may be matched today in mountainous deserts where depressions are occupied by lakes, and valley floors are covered by sand and crossed by ephemeral streams, loaded during floods with rubble from steep mountain sides. Five mountain basins were formed in Great Britain at the start of the Devonian by the Caledonian mountain-building episode, and somewhat similar elongated Devonian basins and uplands were formed in the eastern United States and Canada. The desert lands interfingered with normal marine sediments and river and tidal flat sediments in Ireland, Greenland, Norway, and Quebec.

Renewed uplift raised much of the sea floor above sea level, beginning in the Carboniferous of Europe and eastern North America and continuing into the Permian, in Antarctica, then India, then South America and Africa, and, finally, elsewhere in the world until only a few parts of the Tethys and Siberia remained submerged. Deserts covered most of this land area, apart from western Canada and northern Siberia.

Salt deposits and coral reefs

Before the final emergence of the sea floor by uplift, basins of seawater nearly were cut off from the sea. During the Devonian this occurred in the Michigan Basin and New Brunswick, and during the Permian the Zechstein (salt-stone) Sea of northern Europe formed, and the Delaware and Midland marine basins of Texas became severely constricted. Elsewhere, seas became so shallow that circulation became restricted, as in the Permian of the Russian Platform. Under warm conditions the water in these semienclosed areas became highly saline because of high rates of evaporation. Vast thicknesses of salt, such as sylvite, halite, and carnallite, were precipitated, yielding valuable resources of salt, sodium, and potassium. In Saskatchewan, Canada, there are 6,500,000,000 tons of potassium-rich salt of Devonian age. The complex histories of various basins and the geochemical relationships involved are detailed in the article EVAPORITES.

Coral reefs are deposits that are characteristic of shelf and miogeosynclinal deposits. They are formed by now extinct coral groups and marine plants. The coral islands, coral reefs, and atolls of Indonesia and the South Pacific are modern counterparts. Notable Late Paleozoic developments include the famous Devonian reefs of Alberta, Canada, which are rich in oil, and the Permian reefs of Texas and East Germany.

Some reefs grew to be tens to hundreds of miles long and hundreds of metres (thousands of feet) thick, showing that the growth of the reef kept pace with sea level, either because sea level rose, or the sea floor was slowly sinking. Many reefs, such as the Devonian examples of Germany and the Ural Mountains, grew on volcanic islands between the miogeosyncline and eugeosyncline. Others grew at the edge of the miogeosyncline or close to the Precambrian Shield, as in the Devonian of Alberta. The Permian reefs of Germany grew on folded Devonian rock around the edge of the Zechstein Sea.

Reefs, like desert deposits, have an interesting distribution in time and space (Figure 1). In the Devonian they occur as far north as Ellesmere Island and yet are absent from Africa south of the Sahara and from South America. They are especially abundant in the Middle Devonian, and very rare in the topmost Devonian Fammenian Stage. In the Lower Carboniferous they are restricted to England and a few other places where development is relatively feeble. They flourished again in the Lower Permian, especially Artinskian and lower Tatarian, before vanishing entirely.

Coal measures and glacial deposits

Special shelf deposits on areas close to sea level are coal measures, which form wherever vast amounts of vegetal matter have accumulated and have been buried and transformed by heat and pressure into coal (see COALS). Coal measures generally are rhythmic deposits called cyclothems (*q.v.*) and are made up of limestone and sandstone or shale as well as coal seams. Marine limestones intercalate the coal seams of Pennsylvania and Illinois showing that swampy forest regions, like those of Florida today, suddenly were invaded by the sea and were buried in shelly limestone. A withdrawal of the sea or infilling by sand and mud allowed the forest swamps to become re-established, and the rhythm was repeated many times. In Europe, the Carboniferous swamps often were buried by river-born sediment. Thus a very delicate oscillation of land and sea level was required to allow the forests to grow and then be buried.

Coals of the Upper Carboniferous are restricted to eastern North America, Europe, the U.S.S.R., and parts of China. In the Permian, vast coal deposits were formed in Siberia, India, Australia, Africa, and South America.

The most singular rocks of the Upper Paleozoic are glacial deposits (Figure 1). These rocks provide the best evidence for climatic change during the Upper Paleozoic. The earliest Upper Paleozoic glacial deposits are known from Brazil, in the western Maranhao Basin, of basal Devonian (Gedinnian, or Siegenian) age, with further deposits of possible Upper Devonian age. The deposits of tillite, consisting of boulders of diverse rock types scattered in a clay ground by ice (rock flour), are identical with tills formed by glaciers today. Finely layered shales, called varved deposits (*q.v.*), also are present; these are akin to varves formed in glacial lakes today. In addition, glacially striated pavements are found, where ice has ground over ancient rock and polished it smooth. In some outcrops are glacial bombs, which are boulders rafted by ice and suddenly released onto the underlying sediment on the sea floor or lake floor.

In the Upper Carboniferous such beds—in South Africa, east Australia, and South America—are more extensive than in the Devonian. The area covered by ice increased still further in the Lower Permian, when the rock record indicates that ice covered Africa up to the present Equator, the whole of Australia, east Antarctica, India–Pakistan into the Himalaya Chain (including Sikkim and Nepal), and South America east of the Andes. Younger glacial deposits, in the Middle and Upper Permian, are limited to eastern Australia, including Tasmania.

UPPER PALEOZOIC LIFE

The formation of extensive land masses with mountains and lakes at the outset of the Upper Paleozoic greatly increased the different kinds of dwelling places for life (Table 1). In response, plants and animals changed, in order to live out of water. A second only slightly less significant change was the development of cold regions, capped by ice, resulting in further diversification of life. Some forms adapted to cold conditions and others to warm; this had an important effect on the direction of evolution and the fossil record.

Plants. The simplest plants with woody tissue are of Upper Silurian age. They had erect shoots growing out rhizomes, with simple naked branching axes bearing terminal sporangia and no specialized leaves (*Cooksonia,*

Table 1: Summary of the Upper Paleozoic Fossil Record

Silurian	Devonian	Carboniferous	Permian	Triassic	Period
		So. Allogromiina			Foraminiferida
		So. Textulariina			
		Parathoramminacea			
		Sf. Endothyracea			
1	6	9 Sf. Fusulinacea	8F	1	
				Sf. Nodosariacea	
		Sf. Spirillinacea			
		Cl. Hexactinellida			Porifera (sponges)
		Cl. Demospongia			
		Cl. Heteractinida	8	8	
6	8	7	O. Favositida		Anthozoa (corals): Subcl. Tabulata
2	4	O. Syringoporida 3	3 5F	0	
		O. Auloporida		0	
5	O. Heliolitida 6	6	4	?1-3??	
8	4 O. Halysitida				
2	1	Sf. Cyathaxoniacea			Rugosa
4	6	10 Sf. Zaphrentiacea	8	0	
7	9 3	O. Columnariida 6	4		
4	4 O. Cystiphyllida	2	3 3	0	
3	5	Subcl. Heterocoralia			
		O. Ctenostomota			Bryozoa
3	3 3	3 O. Cystoporata 3	3F	1	
4	2	1 O. Trepostomata 1	1 1	0	
6	4 5	1 O. Cryptostomata 1	1 1	0	
9	9	13 O. Cyclostomata	5 5F	1	
3	3 3	3 Cl. Inarticulata 2	2 2	2 3F	Brachiopods
7 (top)	4	4 Sf. Eichwaldiacea	3 1	0	
1	1 1	1 Sf. Davidsoniacea			
			Sf. Lyttoniacea		
			Sf. Richthofeniacea		
		Sf. Strophalosiacea			
		Sf. Productacea			
		Sf. Pophomenacea			
		Sf. Chonetacea			
	Sf. Plectambonitacea				
		Sf. Enteletacea			
	Sf. Orthacea				
	Sf. Porambonitacea				
	Sf. Pentameracea				
		Sf. Stenoscismatacea			
		Sf. Rhynchoporacea			
		Sf. Rhynchonellacea			
6	9 Sf. Atrypacea 10	7	4F 4F	1	
3	3 Daytacea				
		Retziacea			
2	2	1 Sf. Athyridacea		1	
3	3	2 Sf. Spiriferacea	1	1 Koninckiacea	
4	10	Sf. Spiriferinacea 8F	5F	0	
		Sf. Reticulariacea			
2	3	3	3	?2	
		Sf. Stringocephalacea			
0	3	1 Sf. Dielasmatacea	1 1	0	
	2	5 Sf. Cryptonellacea	4	1	
		O. Palaeoloricata			Amphineura
		O. Neoloricata			
		O. Tryblidioidea			

Table 1: Summary of the Upper Paleozoic Fossil Record (continued)

Period	Silurian	Devonian	Carboniferous	Permian	Triassic
Cl. Monoplacophora	Archinacelloidea		O. Bellerophontoidea		
Gastropoda			O. Archaeogastropoda		
			O. Mesogastropoda		
			O. Opisthobranchia		
				Sf. Subufitacea	
Nautiloidea		O. Discosorida			
		O. Actinoceratida			
	O. Ascoceratida				
	O. Endoceratida				
		O. Michelinoceratida			
		O. Barrandeoceratida			
		O. Oncoceratida			
				F. Parabactrididae	
			F. Bactrididae		
Ammonoidea		Sf. Mimosphinctacea			
		Sf. Mimoceratacea			
		Sf. Pharciceratacea			
		Sf. Anarcestacea			
		Sf. Prolobitacea			
		Clymeniina			
		Proeglyphioceratacea			
			Sf. Dimorphoceratacea		
			Sf. Goniatitacea		
			Sf. Shumarditacea		
			Sf. Cyclolobitacea		
			Sf. Schistoceratacea		
			Sf. Agathiceratacea		
			Sf. Cheiloceratacea		
			Sf. Prolecanitacea		
			Sf. Medlicottacea		
					Sf. Arcestacea
					Sf. Noritacea
					Sf. Clydonitacea
				Sf. Otoceratacea	
Bivalvia			O. Conocardioida		
			O. Nuculoida		
			O. Solemyoida		
			O. Mytiloida		
			O. Pterioida		
			O. Trigonioida		
			O. Unionoida		
			O. Actinodontoida		
			O. Arcoida		
			O. Veneroida		
			O. Hippuritoida		
			O. Lucinoida		
			O. Praeclardoida		
			O. Pholadomyoida		
			O. Myoida		
			O. Poromyoida		
			Subcl. Scaphopoda		
Trilobita		Sf. Illaenacea			
		Sf. Proetacea			
		So. Harpina			

Table 1: Summary of the Upper Paleozoic Fossil Record (continued)

Period	Silurian	Devonian	Carboniferous	Permian	Triassic
Trilobita		So. Cheirurina			
		So. Calymenia			
		So. Phacopina			
		So. Lichida			
		So. Odontopleurida			
	(Brachiopoda) 1				
Crustacea		O. Leperditicopida			
		O. Palaecopida			
Subcl. Ostracoda	2	21	O. Myodocopida 16	11	0
	5		O. Podocopida 6		
Echinodermata	9	12	10	10	7 Subcl. Euthycarcinoidea
Cl. Carpoidea			O. Thoracica		
			O. Acrothoracica		
			O. Cyclina		
		O. Soluta			
		O. Mitrata			
		Class Rhombifera			
		Class Diploporita			
Cl. Crinoidea			O. Disparida		
			O. Poteriocrinitina		
		O. Dendrocrinina			
			O. Cyathocrinitina		
		O. Taxocrinida			
			O. Sagenocrinitida		
		O. Eudiplobathrina			
			O. Tanaocrinina		
					Isocrinida
					Millericrinina
Cl. Blastoidea	1	(4)	O. Fissulato 4	4	
	1(F)	4 Cl. Machaeridia	O. Spiraculata 6	3	
		Cl. Edrioasteroidea			
		Cl. Ophiocystioidea			
			Cl. Echinoidea		
	2F	3	Cl. Holothuroidea 4	3	1
		2	Cl. Stelleroidea 5	6	5
	6	6	Cl. Ophiuroidea 7	6	6
	5	6	4	3	3
Cl. Graptolithina			O. Dendroidea		
		Monograptidae			
Amphibia			O. Conodontophorida		
			Subcl. Labyrinthodontia		
		1	Subcl. Lepospondyli 10 4	10 4	5
			O. Cothylosauria		Lissamphibia
			1 O. Areaoscelidia	6	1
Reptilia				O. Protorosauria	
				O. Mesosauria	
				O. Millerosauria	
				O. Eosuchia	
					O. Prolacertilla
					O. Rhynchocephalia
					O. Squamata
					O. Thecodonta
				O. Pelycosauria	
				O. Therapsida	
				27	17 10

Taeniocradia of the Rhyniophytina; see PSILOPSIDA). These plants persisted into the Devonian, and, with so much land and freshwater available, flourished and multiplied, especially by Siegenian times. One Early Devonian (Gedinnian) group had spikelike clusters of reniform sporangia (Zoosteophyllophytina), which evolved into the Lycophytina (*Protolepidodendron*) in the Siegenian (see LYCOPSIDA). The Trimerophytina (*Dawsonites*) arose in the Siegenian from the Rhyniophytina. Emsian times saw further ramification of groups, with the development of large leaves.

Carboniferous forests

Carboniferous plants flourished in extensive forests, and surviving relatives remain to this day, notably among the ferns. Other inhabitants of the forests included *Calamites* (Sphenopsida *q.v.*), related to scouring rushes, with whorls of slender simple leaves. Scale trees (Lycopsids) grew over 30.4 metres (100 feet) high, in a tall, slender trunk branching repeatedly near the top, forming a crown of stubby twigs with slender straplike leaves (*Lepidodendron*). Another kind, *Sigillaria,* had a thick high trunk with large bladelike leaves. *Cordaites* up to 36.5 metres (120 feet) tall, with leaves inches to 1.5–1.8 metres (5–6 feet) long and seeds in racemes, was a forerunner of conifers (*q.v.*).

In the Permian, the conifer group Voltziales (such as *Walchia*) gradually replaced *Cordaites,* and seed ferns (pteridosperms) became rare, dying out in the Jurassic. Scale trees perished by the end of the Permian Period. Primitive cycadeoids, which were related to the sago palm, became widespread, possibly arising in the Pennsylvanian. The onset of Permian glaciation encouraged the development of a specialized southern flora, typified by *Glossopteris* and *Gangamopteris,* with simple tongue-shaped leaves. The *Glossopteris* Antarctic flora penetrated as far as Siberia in the Middle Permian as well as into southern Asia but greatly declined at the end of the period.

Vertebrate life. Fish remains are rare before the Devonian, though known from Middle Ordovician. Sharks, known chiefly from teeth and fin scales, were common in Devonian seas. Some developed into the largest of Devonian animals, the Arthrodires, of which *Dinichthys* was over 6.08 metres (20 feet) long. Massive armour of bony plates over the head was hinged to the armour of the shoulder region of the body. This group died out by the end of the period. A significant development occurred in the Choanichthyes (also called Sarcopterygii, fleshy finned bony fishes), which are able to breathe air through openings in the roof of the mouth and thus survive in deserts when pools dry out. Of one branch, Dipnoi, five genera remain in deserts of the world. These bear weak fins and no true teeth. A second branch, the crossopterygian fishes (*q.v.*), had more highly organized limbs, giving greater mobility. An Upper Devonian example, *Eusthenopteron* from Escuminac, Gaspé Peninsula, Quebec, had sturdy fin bones, armour-plated head, and infolded teeth. This rhipidistid group of crossopterygian fishes gave rise to amphibians. The Late Devonian of Greenland has yielded skeletons 0.9 to 1.2 metres (three to four feet) long of crocodile-like amphibians called labyrinthodonts from the infolded enamel of their teeth. Nearly all Paleozoic amphibians belonged to this widespread extinct group. In the Carboniferous and Permian, the Amphibia developed considerable diversity, ranging from animals a few inches to ten feet long, and up to 226.5 to 271.8 kilograms (500 to 600 pounds) in weight, including such examples as *Eryops,* a stubby crocodile-like creature ten feet long, "with much belly and little leg, like Falstaff in his old age," according to T.H. Huxley.

Devonian amphibians

The renewal of desert conditions in the Permian gave further impetus to vertebrate evolution. The Amphibia were virtually at a dead end, because most of them were chained to the water for reproductive purposes. Even today, most of their young pass through a tadpole or swimming stage in water. They thus clung to river valleys and lake shores.

The key to future exploitation of the land lay with an offshoot that rapidly became more important and successful than the amphibia. This was the reptile, distinguished from the amphibian by its ability to lay eggs in desert conditions, from which the young emerged fully able to fend for themselves and independent of water. This greater degree of independence, achieved in the Late Carboniferous, saw a rapid expansion of the reptiles into various habitats. In the Permian some were agile like lizards (*Varanops*), some semiaquatic (*Limnoscelis*), others had thick legs and stubby tails. Some were carnivorous, others such as turtles were herbivorous, and others ate shellfish. The most bizarre kind, *Dimetrodon,* had a finback or sail of neural spines. One small group, the theriodonts of the order Therapsida, walked on two hind legs and had specialized teeth—incisors, canines, and molars. They were precursors to the mammals and some authorities consider that the mammals, or closely related forebears, also were present in Permian times.

Invertebrate life

Insects. One of the Late Paleozoic habitats awaiting invasion was the air. Vertebrate life had to await the following era before developing winged reptiles and birds. First to conquer the air were the insects, during Carboniferous times. From the coal measures of Upper Carboniferous age over 500 species are known, including giant dragonflies over two feet (.6 metres) across. Numerous other winged insects are found, together with spiders, scorpions, centipedes, and cockroaches, some as long as 7.62 to 30.48 centimetres (3 to 12 inches).

Marine invertebrate animals. Marine life was extremely varied by the close of the Lower Paleozoic and, unlike terrestrial animal and plant life, quite as diverse as today. It underwent moderate but not drastic change at the start of the Upper Paleozoic. The trilobites, most prolific animals in Early Paleozoic seas, continued their slow decline in numbers and perished in the Permian. The dominant Ordovician and Silurian fossils, called graptolites, persisted into the Devonian as a remnant. Ammonites, related to octopus, became prominent, arising in the Devonian and continuing through the Mesozoic Era, and brachiopods flourished, reaching a peak of diversity first in the Devonian, then again in the Permian. Other dwellers of the sea floor included bivalves, gastropods, and crinoids and corals, both compound corals, which built up sizable reefs, and solitary corals. Microscopic life was also abundant. One group of protozoans, the Foraminiferida (Foraminifera), developed the complex chambered, grain-shaped Fusulinacea, of great value to stratigraphers. Another, still mysterious in origin, formed millions of tiny fossils called coccoliths.

Some invertebrate life forms adapted to freshwater. A more significant development was the adjustment to cold marine waters such as those surrounding the southern ice cap of South America. By contrast, other life became adapted to tropical conditions. Animal life in warm tropical waters was greatly varied, in terms of numbers of orders and species and in morphology. Compared with life in cold waters, shells tended to be small, thin, and intricately ornamented and perhaps coloured as well, but colour is seldom preserved in fossils. These observations accord moderately well with present life, which is most diverse, numerous, and intricate at the Equator and least diverse at the poles.

One of the most exciting and ironic aspects of evolution in the Upper Paleozoic is that the high degree of specialization entailed in adaptation for warm or cold conditions did not pay off in the expected way. A sudden twist to Earth history almost ended in disaster for all life.

SEQUENCE OF FAUNAL CHANGES

The southern cold-water fauna developed first in the Devonian of South America, and by Emsian times some distinctive elements had entered Tasmania, New Zealand, and Antarctica. In the north, by contrast, coral reefs had begun to grow. The waters became much warmer in Eifelian times, to judge from the entry of northern species of brachiopods from North America into South America, the disappearance of the characteristic tillites, and the widespread development of coral reefs in Canada, western Australia, Siberia, the Urals, Europe, and North Africa (Figure 1). Near the close of the Devonian, coral reefs disappeared from most of these areas, and with them

Climatic changes

perished many brachiopods, trilobites, and other forms.

Two explanations of this drastic change have been suggested. Some authorities attribute it to a sudden exposure to ultraviolet light or increase in radioactivity due to mountain building or the impact of a huge meteor that crashed into the earth and raised enormous tidal waves that swept the seas. A less dramatic and perhaps more realistic suggestion is that profound refrigeration occurred, sufficient to kill off many of the warm-water-loving animals, especially coral reefs. Support for this is provided by the recognition of uppermost Devonian tillites in South Africa.

Coral reefs were slow to develop and seldom were prominent in the Carboniferous Period, but there is no strong evidence for refrigeration in the Lower Carboniferous, perhaps because the record is poor for this period in Africa and South America. In the Upper Carboniferous, glaciation covered east Australia, South Africa, and South America, leaving typical glacial rocks and strongly affecting marine life in the surrounding seas. The Northern Hemisphere also was affected: at least in the Yukon of Canada, where a sudden cold phase chilled the waters and changed the fossils in the Bashkirian and again in the late Moscovian.

The major surge of ice occurred in the basal Permian. Ice scraped rock bare and left glacial deposits over South America, Africa, Antarctica, and Australia and moved north in the Indian subcontinent. Simultaneously, cold-water life flourished in neighbouring seas and invaded the northerly waters of Siberia, Alaska, and Canada in force. Warm-water life was forced to retreat to equatorial waters and became threatened with complete extinction. Gradually conditions warmed, the ice melted, and warm-water life was saved to flourish again, eventually to enter a second great phase of reef building, second only to the Middle Devonian, especially in Asia, Texas, Canada, and Russia. These reefs were killed by a second Permian glacial episode, a second flourishing of cold-water life, and constriction of warm-water life. But this time the glacial phase was of lesser extent, ice covering only eastern Australia. A second prolonged warm period followed, but reefs were restricted to Asia, Texas, and Germany, because by then Russia and Arctic Canada had become land. A third short-lived cold phase in the Upper Permian formed tillites in Tasmania and New South Wales and encouraged a short-lived boost for cold-life forms. Then the crisis approached. There would be no further glaciation for at least 100,000,000 years. Life adapted to cold waters appeared to be doomed, and a great deal of life was doomed. The end of the Paleozoic Era was the greatest crisis in life history. There were few major groups that were not at least halved in variety, not to mention numbers. Among marine invertebrate life, Fusulinacea, half the Bryozoa, many ostracods, half of the insect life, corals (including reef builders), most of the elaborate brachiopods, most of the ammonites, all trilobites, many crinoid orders, and all blastoids perished completely. Life that changed but survived included Radiolaria; Foraminifera; sponges; half the Bryozoa; inarticulate, loop, and spire-bearing brachiopods; chitons; gastropods; bivalves; nautiloids; conodonts; and the echinoderm groups Ophiuroidea, Stelleroidea, and Holothuroidea, most of these being minor groups. Even within these groups a great number of species and genera perished.

Mass extinctions

Among the fish, many had perished in the Devonian, and some in the Carboniferous, with no outstanding death toll in the Permian, though half the bony fish of class Actinopterygii died. Many new groups appeared in the Triassic.

Land animals suffered more. Of the labyrinthodonts, order Anthracosauria, including the mainly tropical embolomeres, perished entirely, and the Rhachitomi (suborder including *Eryops*) were greatly reduced. Over half of the reptile orders died, according to A.S. Romer. Cotylosauria and subclass Synapsida were greatly reduced, and almost all Archosauria died. Norman D. Newell calculated that 75 percent of the amphibian families and over 80 percent of the reptile families had disappeared by the end of the Permian.

Some of the survivals of the era tell us little. Animals such as fish and reptiles were so mobile that they were able to flee from disaster areas and chance upon a refuge. More significant is what happened to more sedentary animals. The life that died was, on the whole, the tropical life, not the cold-water life. The compound corals, and many of the life forms that lived on them, died. The warmwater Fusulinacea and the elaborately decorated and evolved brachiopods all died. But they did not die because it became too cold. They were not exterminated by competition with cold-water life. There is no evidence for any glaciation except the minor affair in the Late Permian. There probably was a great increase in temperature in the uppermost Permian until conditions became too hot, especially in the tropics, for a great deal of life to exist; this is the simplest explanation that is compatible with the available evidence.

Warming effects

STRATIGRAPHIC CORRELATION OF UPPER PALEOZOIC UNITS

The correlation of Upper Paleozoic deposits, like those elsewhere in the geological record, depends on many variables and can be attempted by several methods.

The most important boundaries involved lie at the base and top of the Upper Paleozoic, and these have undergone change in recent years. The lower boundary has long been controversial because early workers never precisely defined it, and, as a result, their ideas changed from time to time. An international subcommission of geologists decided to fix the boundary at the base of the graptolite *Monograptus uniformis Zone*. This marks an arbitrary boundary, with the species, rather than genera, differing above and below.

The upper boundary also has been controversial. Pioneers again failed to fix a boundary, which lies somewhere in terrestrial beds without marine fossils. Marine successions that cross the boundary into the Triassic are extremely rare and are restricted to Armenia, Iran, and South China. Basal Triassic is well developed in Arctic Canada and Nevada, and uppermost Permian in New Zealand. In spite of the well-marked faunal change between the Permian and Triassic, a number of American, Australian, British, Japanese, and Soviet workers have inadvertently included Triassic rocks in their Permian systems thereby masking the differences. Chinese, New Zealand, and Canadian workers have brought out the faunal difference through worldwide studies, and their work apppears to be becoming more widely accepted by American and Soviet authorities.

Boundaries between the periods are arbitrary (Table 2). In many ways the uppermost Devonian Stage, the Famennian, has strong links with the basal stage, Tournaisian Stage of the Carboniferous, and the exact boundary, still unfixed, will be placed by consensus rather than on any strong faunal change. The same is true of the base of the Permian. The Upper Carboniferous or Pennsylvanian faunas closely resemble Lower Permian faunas, and the boundary between the two is now placed by investigators of ammonites at the base of the Asselian. Some workers on Fusulinacea prefer a higher base and include the Asselian in the Upper Carboniferous.

UPPER PALEOZOIC ENVIRONMENTS

The curious distribution of Upper Paleozoic rocks has been one of the prime sources of evidence in support of the theory of continental drift, which claims that continents have shifted their position relative to each other through time. Geologically there are at least three independent lines of evidence to suggest that the continents and oceans were differently disposed during the Upper Paleozoic Era (see PALEOGEOGRAPHY). Most striking is the occurrence of glacial deposits in the "southern continents" of Antarctica, Africa, Australia, South America, and the subcontinent of India. One would expect to find corresponding glaciation around the North Pole, but there is none.

The distribution of lands and seas

A second line of evidence lies in the similarity of Upper Paleozoic rocks along both shores of the Atlantic Ocean, both having had a similar history. In the past many have proposed that the two were linked by continents, or land

Table 2: World Correlation of Commonly Cited Upper Paleozoic Stratigraphic Units

	division	Europe	Soviet Union	China	United States	South Africa	New South Wales	Argentine
Permian Period	Upper	Zechstein	Tatarian	Changsing	Ochoan	Lower Beaufort Group	Newcastle Coal Measures	Upper Patquia
				Wuchiaping			Tomago Coal Measures	
	Middle		Kazanian	Maokou	Capitan (Guadalupian)		Mulbring Muree Belford	Middle Patquia
					Word (Guadalupian)		Fenestella Shale	
		Kreuznach	Ufimian		Ranger Canyon	Upper Ecca Shales	Eiderslie	
	Lower	Wadern Tholey Lebach Cusel	Baigendzinian	Chihsia	Leonard	Ecca Coal Measures	Greta Coal Measures	Lower Patquia
			Aktastinian			Lower Ecca Shales	Farley	
			Sakmarian		Wolfcamp-Lenox Hills		Rutherford	
			Asselian	Maping	Neal Ranch	Upper Dwyka Tillite	Allandale Lochinvar Drift	
Carboniferous Period	Pennsylvanian	Stephanian	Gzhelian	Chuanshan Limestone	Missourian	? Dwyka	Seaham	Pituil Group
			Kassimovian		Desmoinesian			
		Westphalian	Moscovian	Huanglung Limestone	Atokan		Patterson Volcanics	San Eduardo Group
					Morrowan		Mount Johnstone	
		Namurian	Bashkirian				Gilmore Volcanics	Magaditas Group
	Mississippian	Viséan	Visokovsky Serpukhosky Venevsky Michailovsky Alexinski Tulsky Stalinogorsky	(Hunan) Tzemenchiao Series Tsehshui Series	Chesterian	Tillite	Wallaringa Wiragulla Beds Bingleburra	
					Meramecian			
		Tournaisian	Cherepetsky Agayevsky Upinsky Malevsky Khovansky	Shihtengtze Series Mengkungao Series	Osagean			
					Kinderhookian			
Devonian Period	Upper	Famennian	Dankov Lebejan Elezk Zadonsk	Makunao Limestone Tutzutang Limestone Changlungchieh Shale	Bradford Cassadaga	Basal ? Dwyka Tillite	Tangaratta-Luton Mandowa	absent
		Frasnian	Liven Evlanov Voronezh Buregi Semiluk Sargajev Paschija Kynov	Shetienchiao Beds Lungkouchung Beds	Cohocton Finger Lakes		Baldwin	
	Middle	Givetian	Cheslav Afonin	Chitzechiao Limestone Ichiawan Shale	Taghanic Tioughnioga	Witteberg Group	Yarrimie	
					Cazenovia			
		Eifelian	Bija Kojva Vjasovaja	Tiaomachien Group	Onesquethaw		Silver Gully	
	Lower	Emsian	Russian platform Takata Formation	Yukiang		Bokkeveld Group	Wogardi Argillite Drik Drik Cope's Creek	Lolen
					Deerpark	Table—Mountain Group		Providencia
		Siegenian	(Krasnokamsk Formation)	Lunghuaskan	Helderbergian		Pipe Clay	Naposta
		Gedinnian			Upper Keyser		(Mountain Creek Volcanics)	Bravard
		Skala						

bridges, now foundered beneath the waves. But geophysical evidence shows that no such bridges existed. It is presently believed that the rocks lay close together in the Upper Paleozoic Era and since that time have been torn apart.

A third kind of evidence is based on the quality and diversity of marine and terrestrial Paleozoic life, which strongly suggests that the tropics then lay well north of their present position in Africa, Asia, and North America. Upper Paleozoic coral reefs are now concentrated at 30°–40° north of the Equator in Central Asia, Europe, and North America. It seems likely that the reefs formed near the Equator and then were moved north with the rest of the land. Marine and terrestrial animals also show a peak of diversity in the same regions, suggesting tropical conditions. Thus the distribution of fossil life, like glacial rocks, suggests a northward movement of both North America and Central Asia and Europe of some 35° latitude since the Permian.

Paleomagnetic evidence, which records the former position of the magnetic, and therefore probably the geographic poles, also suggests continental drift (Figure 2). There is particularly good agreement from all lines

From E. Irving, *Paleomagnetism and Its Application to Geological and Geophysical Problems* (1964); John Wiley & Sons, Inc., New York

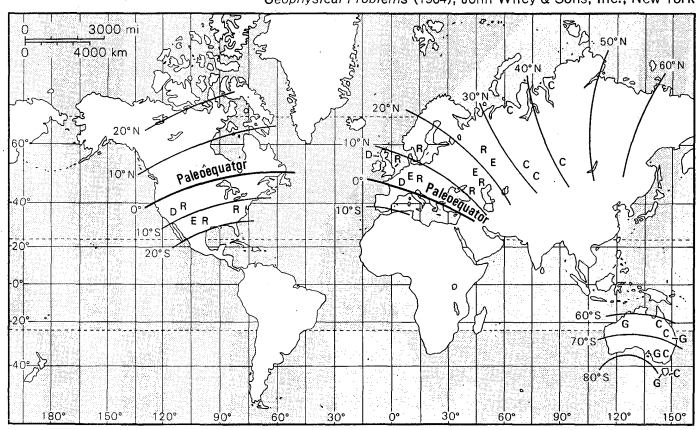


Figure 2: *Ancient latitudes of continents in Permian times.* C=coal deposits; D=desert sands; E=evaporites; G=glacial deposits; and R=red beds.

of evidence for Europe and Australia, two of the most closely studied regions. According to sedimentary, volcanic, paleontologic, and paleomagnetic evidence, Europe and North America lay in contact during Upper Paleozoic times, with the Equator passing through south Europe. According to similar evidence, Australia moved south in Upper Carboniferous times to lie close to the South Pole, placed close to Tasmania. Considerable paleomagnetic data supports the paleontologic and geological evidence that South Africa, India, and South America lay close by.

Igneous activity

Two kinds of volcanic and igneous activity may be distinguished in the Upper Paleozoic record. One accompanied the formation of geosynclines and is indicative of pulling apart of continents either at the edge of oceans (Tethys, Cordilleras) or within continents (Urals). Typically the volcanoes are basaltic in composition, like the lavas of Hawaii or Iceland, and these have poured out in great volume. They occur along the western and southern side of Carboniferous and Permian eugeosynclines, extending from New Zealand through the Himalayas into Europe, and along the geosynclines of British Columbia and the western United States. The extrusion of these lavas was accompanied by the intrusion at depth by chemically identical, more coarsely crystalline, rocks, such as gabbro. Other associated rock types are peridotite, dunite, and serpentinite, ultramafic rocks that probably formed as a heavy fraction from the basalt lavas, either during or before their extrusion. Examples include the ultramafics of the Urals (Devonian 360,000,000 to 380,000,000 years old), the Carboniferous or Permian "Great Serpentine Belt" of New South Wales, and the Kazanian Ultramafic Belt of New Zealand.

The second major kind of igneous and volcanic activity arose from the convergence rather than divergence of continents, in which one continent overrode another or so converged that piles of sediments in geosynclines were squeezed together and forced up into mountains and down deep into the earth. Under pressure and heat, such sediment and volcanics melted and then forced their way up to the surface as molten rock. During the mid-Devonian Acadian disturbance of eastern North America, intermediate to acid lavas and ash showers (rhyolites, ignimbrites—fire stones) were poured out in southern Quebec, Gaspé, New Brunswick, and Maine. The deep-seated roots of these volcanoes are exposed as granites in New England and Nova Scotia. Similar granites are exposed on the opposite side of the Atlantic Ocean in Cornwall and Brittany.

Orogeny

In Europe the Upper Paleozoic Era was ushered in by a drastic change in geography. Lower Paleozoic seas and lands were changed as geosynclines were raised into mountains called the Caledonides across Great Britain and Norway.

In eastern North America, on the other hand, the mountain building occurred a little later as the Acadian disturbance. A mountain range over 160.9 kilometres (100 miles) wide, from which great deltas grew westward in the Middle and Upper Devonian (Catskill Delta), formed from Newfoundland to Cape Hatteras or perhaps as far south as Florida.

Upper Mississippian disturbance was widespread in North America, forming mountains in the Appalachians and the Ouachita Mountains of Arkansas, Oklahoma, and Colorado. Red sandstones were deposited in the Yukon region of Canada. Grits (coarse sands) formed widely in Britain, and the Armorican and Hercynian mountains developed across northwest Europe with culmination of the upheaval at the end of the Carboniferous. Uplift also occurred in New Mexico, Texas (Marathon Uplift), the central interior of the United States, western Canada, eastern Australia, and the Himalayas. Sedimentation and volcanism were most continuous in the Arctic areas of Canada and Siberia and in the Urals, China, and New Zealand.

The Permian is particularly noted for its extensive land areas. During the Permian the seas steadily retreated from all continents, commencing with southern Africa and South America. The early Permian saw extensive submergence, apart from mountainous areas in Europe, followed by widespread deltaic conditions and the formation of coal swamps in the south. Emergence occurred in Australia, Antarctica, India, South Africa, South America, and even Siberia and China, and seas in the Urals, Arctic, and western Canada became considerably shallower. A second submergence of Australia occurred, followed later by similar events in the Urals. Finally, there was widespread emergence of North America, Australia, most of the Uralian area, Europe, and most of Siberia, leaving only parts of Texas, Mexico, the Cordilleran Geosyncline, eastern Siberia, and the Tethys submerged. Only parts of Tethys remained underwater in the last phase of the Permian, however, and even this area probably was briefly emergent.

BIBLIOGRAPHY

Area studies: J. DE VILLIERS, "Devonian of South Africa," in the *International Symposium on the Devonian System* (1967); E.P. PLUMSTEAD, *Gondwana Floras, Geochronology, and Glaciation in South Africa*, Rept. 22nd sess. Int. Geol. Cong., Delhi, India, p. 9. pp. 303–319 (1964); J.B. WATERHOUSE, "World Correlations of New Zealand Permian Stages," *N.Z. J. Geol. Geophys.*, vol. 12, no. 4, pp. 713–737 (1969); K.S.W. CAMPBELL and R.G. MCKELLAR, *Eastern Australian Carboniferous Invertebrates: Sequence and Affinities*, pp. 77–119 (1969); E.W. BAMBER and J.B. WATERHOUSE, *Stratigraphy and Biostratigraphy of the Upper Paleozoic Sequences of the Yukon Arctic Canada* (1971); A.S. ROMER, *Vertebrate Paleontology*, 3rd ed. (1966); N.D. NEWELL, "Crises in the History of Life," *Scient. Am.*, 208:77–92 (Feb. 1963); J.B. WATERHOUSE, "The Permian Faunal Succession in New Zealand," *J. Geol. Soc. Aust.*, 10:165–176 (1963); E.T. TOZER, "Xenodiscacean Ammonoids and Their Bearing on the Discrimination of the Permo-Triassic Boundary," *Geol. Mag.*, 106:348–361 (1969); and D.A. BROWN, "Some Problems of

Distribution of Late Palaeozoic and Triassic Terrestrial Vertebrates," *Aust. J. Sci.*, 30:434–445 (1968).

General texts: M. GIGNOUX, *Géologie stratigraphique*, 4th ed. (1950; Eng. trans., *Stratigraphic Geology*, 1955), a good text for European geology; W.B. HARLAND *et al.*, *The Fossil Record*: A Symposium with Documentation (1967), an elaborately detailed summary of the fossil record, with many errors in detailed ranges, and confusion over the Permian-Triassic boundary; P.B. KING, *The Evolution of North America* (1959), an excellent text; B. KUMMEL, *History of the Earth: An Introduction to Historical Geology* (1961), good for its worldwide coverage; D.H. OSWALD (ed.), *International Symposium on the Devonian System*, 2 vol. (1967), invaluable for an understanding of the Devonian Period; S.K. RUNCORN (ed.), *Continental Drift* (1962), good articles on the continental drift; and A.K. WELLS and J.F. KIRKALDY, *Outline of Historical Geology*, 5th rev. ed. (1966), a modest introduction to the geology of Great Britain.

(J.B.W.)

Palermo

From its foundation by Phoenician traders in the 8th century BC, Palermo (ancient Panormus) has stood as one of the principal cities of Sicily. During the Middle Ages its subjection to successive conquerors and dynasties made it one of the most cosmopolitan cities of the Mediterranean world. The modern city, the capital of the autonomous region of Sicily in the Italian Republic, lies on the northwestern coast at the head of the Bay of Palermo, facing east. Behind the town is a fertile plain, the Conca d'Oro, backed by mountains.

The ancient city

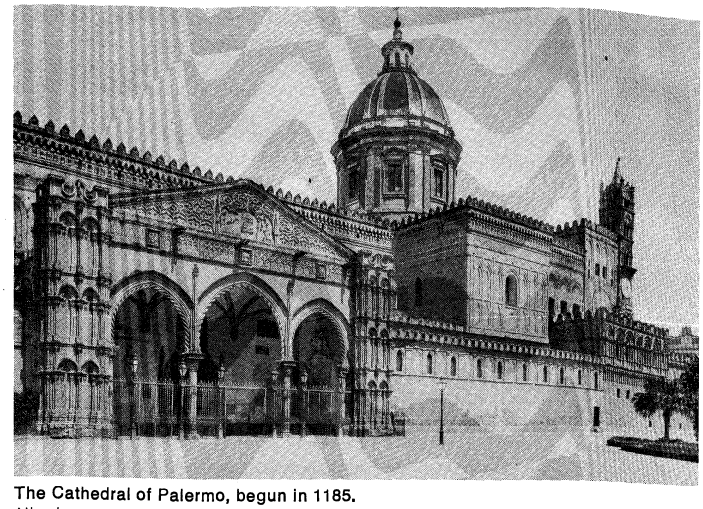
History. The Phoenician city was established along a tongue of land between two former streams to the north and south of the present Via Vittorio Emanuele, and from the beginning Palermo linked its fortunes to its port. The Carthaginians inherited the Phoenician settlements in Sicily in the 5th century BC, and Palermo's thick walls and easy provisioning from the sea secured its independence from the Greeks in Sicily. Romans, however, conquered Palermo in 254 BC after a long siege in the First Punic War. The Roman emperor Augustus (ruled 27 BC–AD 14) established a Roman colony there, but under the empire the city underwent a gradual decay.

A Vandal chieftain, Gaiseric, took Palermo in AD 440 and made it the centre of his repeated raids throughout Sicily. In 476 the city, with the whole of the island, was ceded to Odoacer (Odovacer), the barbarian king of Italy. Palermo prospered mildly under the Ostrogoth kingdom of Theodoric, who had replaced Odoacer on the peninsula. The city was rejoined to the empire by Emperor Justinian's celebrated general Belisarius in 535 and under the Byzantines became the second city of the island, exceeded in political importance only by the capital, Syracuse.

The Arabs conquered Palermo in 831 and made it the capital of the emirate of Sicily. The city prospered under Muslim rule as an emporium of the rich trade with North Africa. Palermo was divided into five new districts, and its population was swelled by merchants from all over the Mediterranean—including many from the maritime cities of Italy. New agricultural techniques were introduced, and luxuriant gardens were planted in the Conca d'Oro.

Thus Palermo was already a flourishing metropolis when it fell to the Norman adventurers Roger I and Robert Guiscard in 1072. But the Norman era (1072–1194) is Palermo's real "golden age," particularly after the foundation of the Norman kingdom of Sicily in 1130 by Roger II. Greeks, Arabs, Jews, and Latins worked together with singular harmony and gave rise to a composite culture of remarkable vitality. The courts of the Norman kings attracted such diverse talents as the Arab poet and geographer Idrīsī, the Byzantine scholar Doxapater, and an Anglo-French man of letters, Peter of Blois. A distinctive Arab–Norman architecture is represented by the Cathedral, founded in 1185 by Gualterio Offamilio (Walter of the Mill) sent to Sicily by the English king Henry II as tutor to William II. From the same period date the Palazzo Reale (Royal Palace), the churches of San Giovanni degli Eremiti (1132) and the Martorana (1143), and several country palaces around the city such as the Cuba and the Zisa.

Palermo's "golden age"



The Cathedral of Palermo, begun in 1185.
Alinari

Norman rule in Sicily was replaced by that of the German Hohenstaufen dynasty with the crowning of the Holy Roman emperor Henry VI as king of Sicily on Christmas Day 1194. Henry's son, Emperor Frederick II (Frederick I of Sicily), shifted the focus of imperial politics to southern Italy and Sicily, and the cultural brilliance of his court at Palermo was renowned throughout western Europe. But Frederick's reign inaugurated a long period of economic decline. The population dwindled, mostly owing to Frederick's mass deportation of Arabs. Building was neglected, and commerce waned despite Frederick's attempts to force immigration from inland towns and to encourage settlement by Italian merchants. Palermo nevertheless supported the Hohenstaufen cause in Sicily until the last Hohenstaufen claimant to the Sicilian throne, Manfred, was crushed by the French Charles of Anjou in the Battle of Benevento in 1266.

Charles punished Palermo's pro-imperial stance by moving the capital of his new kingdom in Naples-Sicily to Naples. Further, he increased taxes and quartered French troops in the city. Popular resentment of Angevin oppression exploded in 1282 in the so-called Sicilian Vespers, in which the French garrison at Palermo was massacred, signalling the beginning of an uprising that within the year achieved the expulsion of the French from all Sicily.

The Sicilian Vespers

A parliament held in Palermo in 1282 invited Peter III of Aragon to the throne. Peter restored the city as the capital of Sicily, but the continuing war with the Angevins in Naples, combined with factional disputes among the Sicilian nobility, undermined royal authority and accelerated Palermo's economic decay. Throughout the 14th century, Palermo, under Aragonese rule, had to struggle hard to keep its predominance over the rival city of Messina.

After 1412 the crown of Sicily was united with that of Aragon (and later Spain), and, with the accession of Alfonso V of Aragon (I of Sicily) in 1416, Palermo took on a new aspect of prosperity. In the 16th and 17th centuries, Spanish viceroys such as Pignatelli, Colonna, Maqueda, and Toledo fostered the reconstruction of the city. Two large intersecting roads redefined the city, and modern squares, such as the Quattro Canti, and churches such as Santa Maria della Catena, Santa Maria di Porto Salvo, La Pietà, Santa Teresa, Santa Caterina, Casa Professa, San Giuseppe dei Teatini, San Domenico, L'Olivella and La Gancia bear witness to the prosperous era of the viceroyalty. But the splendid appearance was belied by repeated risings of the urban poor in the 16th and 17th centuries. The most famous of these was the revolt of 1647, led by the goldbeater Giuseppe d'Alessi and backed by the city's tanners, that threatened Spanish rule throughout Sicily.

In the wake of the War of the Spanish Succession (1701–14), Palermo passed for a short time under the rule of Victor Amadeus II of Savoy and, after 1718, of the Austrian Habsburgs. In the resettlement of Europe

after the Napoleonic Wars, Sicily-Naples was given to the Spanish Bourbons. Revolts against Bourbon rule erupted in Naples in 1820 and 1848; and in 1860 Garibaldi liberated the city. Attached to the newly reformed Kingdom of Italy, Palermo recovered its prosperity and soon expanded beyond its ancient walls.

The modern city. A few public buildings such as the post office were built under the Fascists, and the city suffered severe bombing damage in July 1943, when it was taken by Allied troops. In 1948 regional autonomy was granted to Sicily within the Italian Republic, and Palermo was restored as the leading city on the island. The population leaped from 150,000 to 600,000 and new industries developed in manufacturing, chemicals, and printing; but the problems of endemic poverty in the older parts of the city are still alive. The new port runs both merchant and passenger lines to Tunisia and Naples and handles a lively mercantile trade with the United States, Canada, and Germany. Citrus fruits, cereals, fresh fish, and chemicals are among Palermo's principal exports.

BIBLIOGRAPHY. G.M. COLUMBA, *Storia e topografia antica e medioevale di Palermo* (1910–11), a fundamental work for the ancient and medieval periods; L. GENZARDI, *Il comune di Palermo sotto il dominio di Spagna* (1891), a well documented work for the modern period; I. LA LUMIA, *Palermo* (1875), the best 19th-century guide written by a professional historian; G. PARDI, "Storia demografica della città di Palermo," in *Nuova Rivista Storica* (1919), documents population changes; I. PERI, "Il porto di Palermo dagli Arabi agli Aragonesi," in *Economia e Storia* (1958), an essay on the medieval period; R. ROCHEFORT, "Les bas-fonds de Palerme," in *Annales* (1958), a fundamental study of the city's lower classes; P. VILLA, *Storia della vita urbanistica di Palermo* (1941), a work of some interest that unfortunately has not been brought up-to-date.

(F.Gi.)

Palestrina

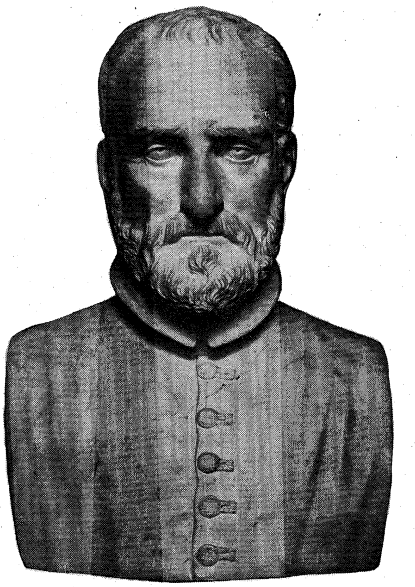
The most celebrated composer of the mid-16th century, Giovanni Pierluigi da Palestrina lived during the period of the Catholic Counter-Reformation and was a primary representative of the 16th-century conservative approach to church music. His music, noted for its transparent serenity, spiritual quality, and careful balancing of consonance and dissonance (the repose and tension of harmonies), is based on the contrapuntal style (*i.e.*, using counterpoint, or interwoven lines of melody) of the Franco-Flemish school of composition, which dominated Renaissance music. Palestrina's music was long taken as a standard of traditional Roman Catholic liturgical music, even by 18th- and 19th-century writers not truly acquainted with his musical traditions.

Life. Palestrina was born about 1525 in the town of Palestrina, near Rome, where his ancestors are thought to have lived for generations. His latinized signature, Joannes Petrus-Aloysius Praenestinus, was taken from Praeneste, the ancient Roman name for Palestrina.

As a child he went to Rome. In 1537 he was one of the choirboys at the basilica of Sta. Maria Maggiore, where he also studied music between 1537 and 1539. In 1544 Palestrina was engaged as organist and singer in the cathedral of his native town. His duties included playing the organ, helping with the choir, and teaching music. His pay was that of a canon and would have been received in money and kind. His prowess at the church there attracted the attention of the bishop, Giovanni Maria Ciocchi del Monte, who later became Pope Julius III.

In 1547 Palestrina married Lucrezia Gori. Three sons were born to them: Rodolfo, Angelo, and Iginio. Only the last outlived his father. In 1551 Palestrina returned to Rome, where he assumed the first of his papal appointments, as musical director of the Julian Chapel choir, and thus was responsible for the music in St. Peter's. Before he was 30 he published his first book of masses (1554), dedicated to Julius III, and the following year he was promoted to singer in the Pontifical Choir. About this time he became composer to the papal chapel. Palestrina repaid the pope's patronage by composing a mass in his honour. Yet he did not neglect the secular side of his art,

First papal appointment



Palestrina, portrait bust by an unknown artist. In the Vatican Museums.
By courtesy of the Vatican Museums

for his first book of madrigals (secular and spiritual partsongs) appeared in 1555, unfortunately at a time when the lenient regime of Julius III had given way to the sterner discipline of Paul IV. A decree of the new pope forbade married men to serve in the papal choir, and Palestrina, together with two of his colleagues, received a small pension by way of compensation for their dismissal.

For the next five years Palestrina directed the choir of St. John Lateran, but his efforts were continually thwarted by singers whose quality was almost as limited as their number, which was restricted because very little money was available for music. Nevertheless, he gained admission for his eldest son, Rodolfo, then about 13, as a chorister. Eventually he broke away from this uncongenial milieu. The chapter archives of St. John Lateran record that in July 1560 he and his son suddenly departed.

A year passed before Palestrina found employment. In March 1561 he accepted a new post at Sta. Maria Maggiore. This post was more congenial to him and he remained at it for about seven years. At the invitation of Cardinal Ippolito d'Este he then took charge of the music at the Villa d'Este in Tivoli, a popular summer resort near Rome. He was in the cardinal's service for four years, at which time he also worked as music master for a newly formed Seminarium Romanum (Roman Seminary), where his sons Rodolfo and Angelo became students.

Palestrina received an offer in 1568 to become musical director at the court of the emperor Maximilian II in Vienna. He refused the position because of the low salary and a disinclination to leave Rome. Palestrina's terms were also too high when he was invited to the court at Mantua in 1583. The composer, however, and the duke of Mantua, Guglielmo Gonzaga, an amateur musician of some pretensions, did become friends, and Palestrina was commissioned to write special compositions for the ducal chapel of Sta. Barbara.

Return to St. Peter's

With the death in 1571 of the composer Giovanni Animuccia, musical director at the Vatican since 1555, there was a chance for Palestrina to return to his old post as musical director of the Julian choir. The chapter, eager to have him back, increased the salary, and he forthwith returned to St. Peter's. When his growing fame as a composer prompted Sta. Maria Maggiore to rehire him, St. Peter's again raised his salary. In acknowledgment of his position as the most celebrated Roman musician, he was given in 1578 the title of master of music at the Vatican Basilica.

The series of epidemics that swept through central Italy in the late 1570s carried off his wife and his two elder sons, both of whom showed great musical promise. He

himself fell seriously ill. Grieving over his wife's death, he announced his intention of becoming a priest, to the delight of the pope, Gregory XIII. After having been made a canon, however, he renounced his vows in order to marry (1581) Virginia Dormoli, widow of a wealthy merchant. Although he spent considerable time administering her fortune, he retained his position at St. Peter's and continued to compose.

Although an attempt in 1585 to make Palestrina musical director of the Pontifical Choir proved abortive, he was considered by all the popes under whom he served as the official composer for the choir, and it is recorded that he marched at the head of the pontifical singers on the occasion of erecting the great Egyptian obelisk in the piazza of St. Peter's.

Pope Gregory XIII had commissioned Palestrina and Annibale Zoilo to restore the plainchant, or plainsong (a traditional liturgical chant sung in unison), then in use to a more authentic form. The task proved too great, and Palestrina's editorial work gave way to a flow of creative music. Much of it was published during the last 12 years of his life, including volumes of motets (choral compositions based on sacred texts), masses, and madrigals. He also helped to found an association of professional musicians called the Vertuosa Compagnia dei Musici.

Two years before Palestrina's death, the new pope, Clement VIII, increased his pension; and the same year, in a singular mark of respect and admiration, fellow composers paid their elderly senior the compliment of writing 16 settings of the Vesper Psalms to his praise. In return, Palestrina sent them a motet on the appropriate text: *Vos amici mei estis* "You are my friends, if you do what I teach, said the Lord." In January 1594 he was taken ill, and he died on February 2, in Rome.

Music. Palestrina's musical output, though vast, maintained a remarkably high standard in both sacred and secular works. His 105 masses embrace many different styles, and the number of voices used ranges from four to eight. The time-honoured technique of using a cantus firmus (pre-existent melody used in one voice part) as the tenor is found in such masses as *Ecce sacerdos magnus*; *L'Homme armé*; *Ut, re, mi, fa, sol, la*; *Ave Maria*; *Tu es Petrus*; and *Veni Creator Spiritus*. These titles refer to the source of the particular cantus firmus. Palestrina's mastery of contrapuntal ingenuity may be appreciated to the fullest extent in some of his canonic masses (in which one or more voice parts are derived from another voice part). His ability to ornament and decorate a solemn plainchant, making it an integral part of the texture and sometimes almost indistinguishable from the other, freely composed parts, is evident from some of his masses based on hymn melodies.

The masses

By far the greatest number of masses employ what has come to be known as the parody technique, by which a composer made use either of his own music or that of others as a starting point for the new composition. Many other masses derive from musical ideas by Palestrina's predecessors or contemporaries. Yet another type of mass is demonstrated by the nine works written for Mantua; in these the Gloria and Credo sections are so arranged that plainsong and polyphony alternate throughout. Finally, there is a small but important group of masses that are in free style, the musical material being entirely original. Perhaps the best-known example is the *Missa brevis* for four voices.

Palestrina's motets, of which more than 250 are extant, display almost as much variety of form and type as do his masses. Most of them are in some clearly defined form, occasionally reflecting the shape of the liturgical text, though comparatively few are based on plainsong. Many of them paraphrase the chant, however, with an artistry that is every bit as successful as that of the masses. On the same level as the canonic masses are such motets as *Cum ortus fuerit* and *Accepit Jesus calicem*, the latter apparently a favourite of the composer's—an assumption justified because he is depicted holding a copy of it in a portrait now in the Vatican.

His 29 motets based on texts from the Song of Solomon afford numerous examples of "madrigalisms": the use of suggestive musical phrases evoking picturesque features, apparent either to the ear or to the eye, sometimes to both. In the offertories, Palestrina completely abandons the old cantus firmus technique and writes music in free style, whereas in the hymns he paraphrases the traditional melody, usually in the highest voice. In the *Lamentations of Jeremiah* he brings effective contrast to bear on the sections with Hebrew and Latin text, the former being melismatic (floridly vocalized) in style and the latter simpler and more solemn. His Magnificats are mainly in four sets of eight, each set comprising a Magnificat on one of the eight "tones": *alternatim* structure is used here as in the Mantua masses.

Although Palestrina's madrigals are generally considered of less interest than his sacred music, they show as keen a sense for pictorial and pastoral elements as one finds in any of his contemporaries. Over and above this, he is to be remembered for his early exploitation of the narrative sonnet in madrigal form, notably in *Vestiva i colli*, which was frequently reprinted and imitated. His settings of Petrarch's poems also are of an exceptionally high order.

Palestrina's reputation

At the end of the 19th century the view that Palestrina represented the loftiest peak of Italian polyphony was in some ways detrimental to his reputation, for it cast his music into rigid preconceptions. Even more unfortunate was the insistence on "counterpoint in the style of Palestrina" in the examination requirements of academies and universities, for such requirements stultified a style that Palestrina had used with great flexibility. Generations of fledgling composers were taught to revere the music of Palestrina as a symbol of all that was pure in ecclesiastical counterpoint. Indeed, the greater part of his musical output, and in particular his masses (where his unerring sense of tonal architecture may be heard at its best), still remains worthy of admiration.

Palestrina, unlike Bach, did not have to be rediscovered in the 19th century, though the dissemination of his achievement was helped by the interest of the Romantic composers. There always was a Palestrinian tradition, mainly because his music supplied the need for a well-regulated formal system to be used for the embryonic composer in presenting himself to the unknown musical world. Strict counterpoint was associated with a technique acquired in this way. It became a standard type of musical discipline that, however, gradually moved farther and farther away from its progenitor. In his day, Palestrina was a senior figure who, utilizing the dominant musical style of his time, created works notable for their spiritual qualities and technical mastery.

MAJOR WORKS

MASSES: 105 masses, including *Ecce sacerdos magnus* (published 1554), *De Beata Virgine* (1567), *Papae Marcelli* (1567), *L'Homme armé* (1570), *Lauda Sion* (1582), *O magnum mysterium* (1582), *Aeterna Christi munera* (1590), *Jam Christus astra ascenderat* (1590), *Dum complerentur* (1599), *Tu es Petrus* (1601), *Veni Creator Spiritus* (1888).

MOTETS: More than 250, for 4, 5, 6, 7, 8, and 12 voices, including a set of 29 motets based on the Song of Solomon (1584); *Assumpta est Maria* (1572); three settings of *Alma Redemptoris Mater*; three settings of *Lauda Sion*.

MADRIGALS: Spiritual madrigals for three, four, and five voices; secular madrigals for various voices.

OTHER VOCAL MUSIC: including 68 Offertories, 13 complete sets, Lamentations, 12 Litanies, 20 Psalms, 35 Magnificats (mainly in four sets of eight, each set comprising a work on one of the eight "tones").

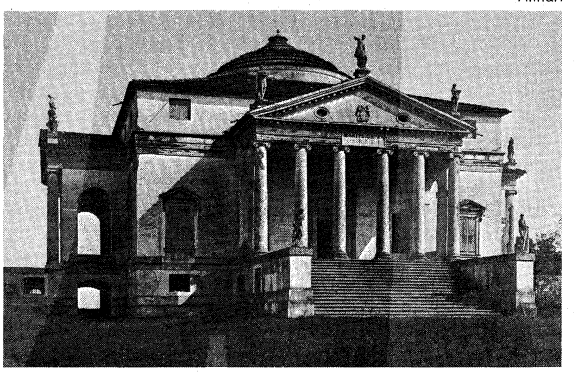
BIBLIOGRAPHY. HENRY COATES, *Palestrina* (1938), is the standard English-language study of the life and works of Palestrina and contains a complete list of works. JEROME ROCHE, *Palestrina* (1971), is a brief, up-to-date biography useful for new details about Palestrina's life. Information on more recent discoveries also appears in the excellent section on Palestrina in GUSTAVE REESE, *Music in the Renaissance*, rev. ed. (1959). KNUD JEPPESEN, *Der Palestrinastil und die Dissonanz* (1925; Eng. trans., *The Style of Palestrina and the Dissonance*, 2nd ed., 1946, reprinted 1970), is a specialized but thoroughgoing study of the technical aspects of the composer's contrapuntal style.

(D.W.S.)

Palladio, Andrea

The greatest architect of 16th-century northern Italy, Andrea Palladio is also one of the most influential figures in the whole development of Western architecture. The qualities that made him influential were numerous and varied. His palaces (*palazzi*) and villas were imitated for 400 years all over the Western world; he was the first architect to systematize the plan of a house and consistently to use the ancient Greco-Roman temple front as a portico, or roofed porch supported by columns (this was probably his most imitated architectural feature), and finally, in his *I quattro libri dell'architettura* (*Four Books on Architecture*), published in 1570, he produced a treatise on architecture that, in popularizing classical decorative details, was possibly the most influential architectural pattern book ever printed.

Alinari



Villa Rotonda, near Vicenza, Italy, by Andrea Palladio, 1550–51.

Palladio was born November 30, 1508, in Padua, in the Italian province of the Veneto, where, as a youth, he was known by his christened name of Andrea di Pietro della Gondola. He was apprenticed to a sculptor in Padua until, at the age of 16, he moved to nearby Vicenza and enrolled in the guild of the bricklayers and stonemasons. He was employed as a mason in workshops specializing in monuments and decorative sculpture in the style of the Mannerist architect Michele Sanmicheli of Verona. There were at the time no classical buildings of note in Vicenza, the Veneto having been badly disturbed by wars until the 1530s.

Early works in the Veneto. Between 1530 and 1538 Count Gian Giorgio Trissino, a Humanist poet and scholar, was rebuilding his villa at Cricoli outside Vicenza in the ancient Roman, or classical, style. Andrea, working there as a mason, was noticed by Trissino, who undertook to expand his practical experience with a Humanist education. The Villa Trissino was rebuilt to a plan reminiscent of designs of Baldassarre Peruzzi, an important High Renaissance architect. Planned to house a learned academy for Trissino's pupils, who lived a semi-monastic life studying mathematics, music, philosophy, and classical authors, the villa represented Trissino's interpretation of the ancient Roman architect and theorist Vitruvius (active 46–30 BC), whom Palladio was later to describe as his master and guide. The name Palladio was given to Andrea, after a Humanist habit, as an allusion to the mythological figure Pallas Athena and to a character in Trissino's poem "Italia liberata dai goti." It indicates the hopes Trissino had for his protégé.

Patronage of Trissino

At the Villa Trissino, Palladio met the young aristocracy of Vicenza, some of whom were to become his patrons. By 1541 he had stylistically assimilated the Mannerist works of Michele Sanmicheli and the High Renaissance buildings of Jacopo Sansovino, whose library of St. Mark's in Venice had been begun in 1536. He had probably been introduced in Padua to Alvise Cornaro, whose designs were the first to import the Roman Renaissance style to northern Italy. Palladio may also have met a prominent Mannerist architect and theoretician, Sebastiano Serlio, who was in Venice at that time and whose third and fourth books on architecture (*L'Architettura;* 1537 and 1540) were to be an inspiration to him.

In about 1540 Palladio designed his first villa, at Lonedo for Girolamo de' Godi, and his first palace, in Vicenza for Giovanni Civena. The Villa Godi has a plan clearly derived from the Villa Trissino but with similarities to traditional Venetian country houses. It contains all the elements of Palladio's future villa designs, including symmetrical flanking wings for stables and barns and a walled courtyard in front of the house. In elevation the Palazzo Civena is close to the High Renaissance palace type developed in the early 16th century in Rome. In plan it resembles Sanmicheli's Palazzo Canossa (*c.* 1535) in Verona. An innovative feature is the use of traditional arcaded pavement of northern Italy behind the main elevation, an idea that Palladio reinterpreted in imitation of an ancient Roman forum.

Visits to Rome and work in Vicenza. In 1541 and again in 1547 Palladio visited Rome with Trissino. These visits greatly affected his palace designs. On them, he saw the work of the greatest architects of the Roman High Renaissance style, Donato Bramante, Peruzzi, and Raphael, generally more remembered for his painting than for his architecture. He also measured ancient Roman antiquities, notably the baths. Palladio's principal ideas on palace design were formed between his first works of 1540 and his visit to Rome in 1554–56.

In 1546 Palladio prepared designs for the reconstruction of the 15th-century town hall in Vicenza, known since then as the Basilica, and in 1548 these plans were accepted, though much earlier designs, drawn in 1534 by the Mannerist architect and painter Giulio Romano and by several other distinguished architects, had been previously rejected. This was his first major public commission, and the work, which was not actually finished until 1617, involved recasing a vast hall with a two-story arcade of white stone to serve as a buttress to the old structure. Suited to both the Gothic style of the original structure and the dimensions of the classical orders, Palladio's arcade was of great proportional subtlety. The architectural motifs used were taken from Serlio and from Sansovino's library of St. Mark's in Venice. Up to 1556 Palladio produced three basic palace types. The first, in 1550, was the Palazzo Chiericati, in which he extended his Palazzo Civena forum idea of a block with its axis parallel to the pavement, which it envelops in a loggia, or roofed open gallery. The tripartite division of the colonnaded elevation, which gives the building a definite central focus, was an innovation. The second, in 1552, was seen in the Palazzo Iseppo Porto, Vicenza, in which he stated in its clearest form his reconstruction of a Roman house. The facade was closely based on the Roman Renaissance palace type, such as Bramante's House of Raphael (*c.* 1514), which Palladio had drawn in Rome. But it was planned in what Palladio believed to be the ancient Roman style. Two tetrastyle halls with four columns each were placed on opposite sides of a court surrounded by a giant colonnade of Corinthian columns. The third, in 1556, was in the Palazzo Antonini in Udine, which has a square plan with a central four-column tetrastyle hall and the service quarters asymmetrically to one side. The facade has six columns, which are attached to the wall rather than free-standing and which are centrally placed on each of the two floors, surmounted by a pediment or a low-pitched gable—a device normally used in his villas.

Three types of palaces

Palladio further developed the basic plan of his Palazzo Iseppo Porto in the Palazzo Thiene (*c.* 1545–50), Vicenza, the largest and most problematical of his palace designs, of which only the side and rear blocks were completed. Four wings, containing a combination of rectangular rooms and small octagons, similar to those of the Roman public baths, are symmetrically placed around a huge court. The elevations are of a grandeur unequalled in Palladio's other work. The design is the first in which Palladio was influenced deeply by the prevailing contemporary style of Mannerism and especially by Giulio Romano, who was in Vicenza when the project was begun.

During his stay in Rome, 1554 to 1556, Palladio in 1554 published *Le antichità di Roma* ("The Antiquities of Rome"), which for 200 years remained the standard guidebook to Rome. In 1556 he collaborated with the

classical scholar Daniele Barbaro in reconstructing Roman buildings for the plates of Vitruvius' influential architectural treatise (written after 26 BC) *De architectura* (*On Architecture*). The new edition was published in Venice in 1556.

Palladio's elevations have always a central emphasis that reflects the axial symmetry of the plan. This is developed in the Palazzo Valmarana, Vicenza, of 1565, along with an increasing use of stucco surface reliefs and giant orders, or columns, extending more than one story. The latter are both Mannerist elements, used particularly by Michelangelo. Giant orders were also used in the massive and unfinished Palazzo Porto-Breganze of *c.* 1570 and finally in the Loggia del Capitanio of 1571. The latter was built in emulation of many similar loggias, such as those of Florence and Venice. The lower floor was to be a raised platform open to the square and the upper a meeting hall. The original decoration was adapted to symbolize the contribution of Vicenza to the Venetian victory over the Turks at Lepanto in 1571, and a triumphal-arch motif was added to the side elevation. But the cost of the victory so impoverished the government that only three bays, or sections, were built of a possible five or seven intended.

Though Palladio absorbed contemporary Mannerist motifs, his plans and elevations always retained a repose and order not associated with Mannerist architecture, particularly that of Michelangelo and Giulio Romano. When the simplicity of his early designs was abandoned, it was largely to incorporate details warranted by the examination of buildings of the late Roman Empire, reflecting archaeological study common to his period.

Villas

Palladio's villas were less affected by his visits to Rome. For practical reasons these buildings were always of stuccoed brickwork with a minimum of carved-stone detail. His aim was to recreate the Roman villa as he had come to understand it from Latin descriptions in the writings of Pliny and Vitruvius. His villas were built for a capitalist gentry who, during the period of Palladio's maturity, gained in prosperity and found new economic outlets in agricultural improvement and land reclamation. He developed the prototype plan of Villa Trissino with many variations at Cricoli. The plan could change in scale and function to serve as a summer residence of an urban aristocrat or the estate headquarters of a gentleman farmer. Included in the former category are the least typical and most widely copied of Palladio's villa designs, the villa for Giulio Capra, called the Villa Rotonda, near Vicenza. This was a hilltop belvedere, or summer house, with a view, of completely symmetrical plan with hexastyle, or porticoes on each of four sides and central circular halls surmounted by domes. The Villa Trissino at Meledo, of the same type, was to have curved wings attached to the main portico. This was a device Palladio usually used when less consideration had to be given to farming and agricultural use of the land. Although the Villa Trissino was not built, it was a most influential design because it was illustrated in the *Quattro libri*.

Palladio adapted the classical temple front to the facades of his villas because it had the dignity suitable for an entrance. He reasoned that, since ancient temples such as the Pantheon in Rome had pedimented porticoes, houses, which preceded temples, would also have had them. Sometimes, as at the Villa Cornaro (*c.* 1560–65) at Piombino Dese and the Villa Pisani (*c.* 1553–55) at Montagnana, the portico is two-storied, with principal rooms on two floors. Normally (as at the Villa Foscari at Mira, called Malcontenta [1560]; the Villa Emo at Fanzolo [late 1550s]; and the Villa Badoer), the porch covers one major story and the attic, the entire structure being raised on a base that contains service areas and storage. In a third type the temple front covers the whole front of the house, as at the Villa Barbaro (*c.* 1555–59) at Maser, which Palladio designed for his friend the scholar Daniele Barbaro. This villa retains the contemporary fresco interiors painted by the Venetian master Paolo Veronese (*c.* 1528–88) and is one of the few interiors to survive from Palladio's day.

At the Villa Thiene (*c.* 1550) at Quinto, he started to build a grandiose house planned on the lines of his reconstruction of a Roman villa shown in the *Quattro libri*, but it was never finished. At the Villa Sarego (*c.* 1568–69) at Santa Sofia a similar inward-facing complex was also planned but not completed. This design differs from the normal villa in its two-story rusticated colonnade forming loggias to rooms arranged around three sides of a court. It is reminiscent of the court to the Pitti Palace in Florence, built in 1550 by the Mannerist architect and sculptor Bartolommeo Ammannati (1511–92). The rustication, or boldly hewn masonry, of Villa Sarego is probably derived from a style used on the Venice mint (begun in 1537) by Sansovino.

Palladio's villas were planned as total complexes but could be built in part to satisfy the owner's immediate requirements. He attached great importance to the courts that flanked or stood in front of the house, since they extended its axial symmetry and proportion.

I quattro libri

At the end of 20 years of intensive building, Palladio in 1570 published *I quattro libri dell'architettura*. This was a summary of his studies of classical architecture. He used many of his own designs to exemplify the principles of Roman design. The first book contains studies of materials, the classical orders, and decorative ornaments; the second, many of Palladio's designs for town and country houses, together with his classical reconstructions. His executed designs are often corrected, particularly in the case of early works like the Villa Godi. They are marked with dimensions according to a system of mathematical ratio. The ratios used are based upon the musical intervals in use in Palladio's day, and it was believed that numerical equivalents would result in a beautiful building, since it would be designed within a universal mathematical order. The third book contains designs for bridges, ancient town planning, and basilicas, or ancient Roman oblong halls for public assembly, later adopted as a prototype for the Christian church. The fourth book concerns the reconstruction of ancient Roman temples.

Venetian period. After 1570 Palladio's life was centred on the building of churches in Venice. In the Veneto, because of a war with the papacy, few churches had been built in the first half of the century, and there are no church designs in his early drawings. Palladio's first design was for the facade of S. Pietro di Castello (1558) in Venice—a design that does not survive. Around the 1560s he was working on monastic commissions in Venice for Sta. Maria della Carità and for the refectory and cloisters of S. Giorgio Maggiore. In the early 1560s he designed the facade for S. Francesco della Vigna, at Venice, which had been built according to Sansovino's designs of 1534 but was never finished. Palladio's facade became a design prototype for classical churches with a high nave, or central aisle, and lower aisles. He resolved this by intersecting classical temple fronts—one joining the side aisles and the other, grander front superimposed upon it and covering the higher elevation of the nave. This ingenious solution was refined and perfected in the facades of S. Giorgio Maggiore (1566, completed in 1610) and Il Redentore (1576, completed in 1592). The liturgical revival of the Counter-Reformation opposed the centrally planned church requiring separate functions for different parts of a Latin-cross church. Palladio's proposals for a circular church for Il Redentore, therefore, were rejected. In both churches the nave is a hall of gray stone columns, lit from windows at high level and covered with a plain stucco barrel vault. The interiors are a chaste white with no decoration. In Il Redentore the apse is lit from the dome above and from the choir, which stands behind a semicircular screen of columns.

At the end of his life, in 1579, Palladio designed a central-plan church as a chapel at Maser. It is a shallow Greek cross covered by a circular dome. Internally, the complex decoration of all surfaces relates it in style more closely to Palladio's late palace designs than to his churches. This was followed by a similar unexecuted project, S. Nicola di Tolentino (1579) in Venice. These demonstrate Palladio's ideal church plan and follow his reconstruction of the Pantheon in the *Quattro libri* and paralleling designs by Giacomo da Vignola (1507–73), the leading architect in Rome after Michelangelo.

With the death of Sansovino in 1570, Palladio became the leading architect of the Veneto region. Until then he had failed to gain official state patronage, and his designs for palaces in Venice, known from the *Quattro libri* and from drawings, had never found patrons. His later civic work in Venice consisted of advice on fortifications, designs for decorations used on state occasions, and interiors for the Doges' Palace. In 1572 his two sons died, and afterward he lived a secluded life, publishing only an illustrated edition of Julius Caesar's *Commentaries* as a memorial.

Palladio's last commission came in 1579–80—to build a theatre in Vicenza for the Accademia Olimpica for the performance of classical dramas. The design of the Teatro Olimpico was in the nature of an academic exercise, being based on the reconstruction of the ancient Roman theatre at Orange, in France. In August 1580 Palladio died, leaving a considerable number of unfinished buildings: the Basilica in Vicenza, the two Venetian churches, the Villa Rotonda, and the Teatro Olimpico. These were continued by his followers, notably Vincenzo Scamozzi (1552–1616) and O. Bertotti-Scamozzi (1719–90), but because of the changing taste of the period they were not strictly in accordance with Palladio's designs.

The influence of Palladio's buildings and publications reached its climax in the architecture of the 18th century, particularly in England, Ireland, the United States, and Italy, creating a style known as Palladianism, which in turn spread to all quarters of the world. His immediate influence on his contemporaries was largely confined to his pupil Vincenzo Scamozzi, who designed a number of villas in his master's style, notably the Villa Molin (*c.* 1597), near Padua.

MAJOR WORKS

VILLAS: Villa Godi, Lonedo (*c.* 1540–42); Villa Marcello, Bertesina (*c.* 1540–44); Villa Poiana, Poiana Maggiore (*c.* 1545–50); Villa Thiene, Quinto (*c.* 1550); Villa Pisani, Bagnolo (1540s–60s); Villa Rotondo, Vicenza (1550–51); Villa Pisani, Montagnana (*c.* 1553–55); Villa Badoer, Fratta Polesine (1554–63); Villa Chiericati, Vancimuglio (1554–57); Villa Emo, Franzolo (late 1550s); Villa Barbaro, Maser (*c.* 1555–59); Villa Cornaro, Piombino Dese (*c.* 1560–65); Villa Valmarana, Lisiera (*c.* 1565–66).

PALACES AND PUBLIC BUILDINGS: Palazzo Civena, Vicenza (1540–46); Palazzo Thiene, Vicenza (*c.* 1545–50); Basilica, Vicenza (design accepted 1548, finished 1617); Palazzo Chiericati, Vicenza (1550); Palazzo Valmarana, Vicenza (1565–66); Teatro Olimpico, Vicenza (1579–80).

CHURCHES: S. Giorgio Maggiore, Venice, refectory (1560–62), church (1566, completed 1610); S. Francesco della Vigna, Venice (*c.* 1565); Il Redentore, Venice (begun 1576); Tempietto, Maser (design 1579, begun 1580).

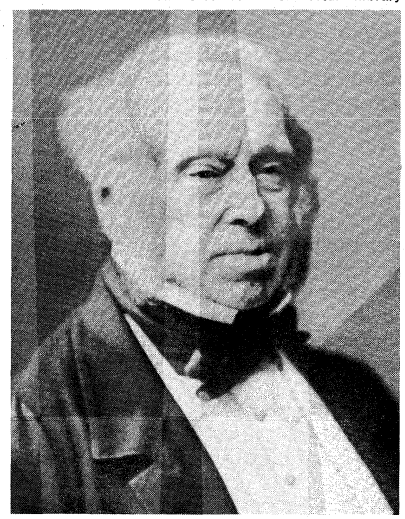
BIBLIOGRAPHY. ROBERTO PANE, *Andrea Palladio*, 2nd ed. (1961), although not translated into English, is useful for its lavish illustrations; the text is subjective. RUDOLPH WITTKOWER, *Architectural Principles in the Age of Humanism*, pt. 3–4, 3rd ed. (1962), provides the most illuminating analysis of Palladio's design ideas, particularly his use of mathematical proportions. The following works by the Italian scholar G.G. ZORZI, although not translated into English, provide an essential documentary archival foundation for the study of Palladio. Each contains a catalogue raisonné of the buildings and publishes all the original material (drawings) and documentary evidence: *I disegni delle antichità di A.P.* (1959); *Le opere pubbliche e i palazzi privati di A.P.* (1965); *Le chiese e i ponti di A.P.* (1967); *Le ville e i teatri di A.P.* (1969). JAMES S. ACKERMAN, *Palladio* (1966), is the most reliable recent historical-critical essay on Palladio's *oeuvre* as a whole. An introduction on the major influences on Palladio's career is followed by chapters on the different building types. The final chapter gives the author's definition of Palladio's principles, with full bibliography. The same author's *Palladio's Villas* (1967) is an expanded version of ch. 2 in *Palladio* above, with a catalog of the villas and documentary evidence.

(M.A.R.)

Palmerston, Lord

That "Palmerstonian" should become an adjective to describe any British foreign-policy chauvinist who is brusque in tone, propagandist and bullying in method, suggests the distinctiveness of the man who held the offices of foreign secretary in 1830–34, 1835–41, and 1846–51 and of prime minister in 1855–58 and 1859–65. In fact, Palmerston's policies were more traditional and orthodox, his methods more responsible and—when it mattered—usually more temperate than this caricature suggests; and in 1851 and 1858 he sacrificed his position by "truckling" to the French when it seemed to him that the public interest—not understood by the public—required it. But he did, uniquely, in the 1850s, represent Britain's confidence in its commercial ascendancy; in the fleet that maintained the "Pax Britannica" and protected British lives and property in many parts of the world; in the superiority and in the suitability for export of its liberal institutions; and in its right (though its military forces were meagre) to play a leading role in international relations. Becoming prime minister and (incidentally) leader of the Liberal Party in 1855, as a result of public pressure on Parliament, he remains the only British premier who twice increased his parliamentary majority by an appeal to the electorate.

Radio Times Hulton Picture Library



Palmerston, *c.* 1860.

Early life. The christening of Henry John Temple (born October 20, 1784, Broadlands, Hampshire) in the "House of Commons church" of St. Margaret, Westminster, was appropriate. His father, a cultured grand seigneur and dilettante politician, failed in his ambition to convert his Irish peerage into a United Kingdom peerage, which would have given him and condemned his son (known as Harry) to a seat in the House of Lords. Instead, with a break of less than a year (in 1835), Harry Temple was to sit in the Commons from 1807 until he died as prime minister on the eve of his 81st birthday (October 18, 1865, Brocket Hall, Hertford). After two years in Italy and Switzerland with his family, young Temple went to Harrow School in May 1795. Its classical curriculum was supplemented by French, Italian, and some German from a tutor brought home from Italy. In November 1800, Temple entered the University of Edinburgh, lodging with the Scottish philosopher Dugald Steward and imbibing the free-trade teaching of Adam Smith.

Succession to title and estates

In April 1802 Temple succeeded to his father's title and estates as 3rd Viscount Palmerston and to a burden of debt that conspired, along with a sense of public duty, to make him seek public office; the fact was, he could never afford to be out of office long. He soon began to extend and embellish the house and gardens of Broadlands in Hampshire and, from the mid-1820s, improved his Irish estates in County Sligo. Having survived a youth of ill health, he later displayed a rare stamina, cultivated by regular exercise. (His chronic liability of eyestrain, however, explains one of the exacting demands that made him unpopular with foreign office clerks and diplomats—a good bold hand. In his last 20 years he suffered severely from gout.) Entering St. John's College, Cambridge, in

October 1803, Palmerston was still an undergraduate when he contested the vacancy in the university parliamentary representation resulting from the death of William Pitt, the family political hero, in January 1806. He lost then and again in the general election of 1807, but he sat for the University of Cambridge from 1811 to 1831.

Political life, 1807–30. Only after being made—through the patronage of a guardian (Lord Malmesbury)—a junior lord at the admiralty in the Tory government of 1807 did Palmerston become a member of Parliament by a transaction with the patron of the pocket borough of Newport, Isle of Wight. Studious at Cambridge, though no recluse, the young junior minister was thought a bit of a prig. As he passed into middle age 20 years later, he was thought of politically as a dull dog; for, after refusing the chancellorship of the exchequer from the prime minister Spencer Perceval in 1809, he took the office of secretary at war, worth £2,480 a year, and found himself condemned to "tread water" for nearly 20 years. The office was humdrum, and its parliamentary duties were light. He rejected the post office in 1821 because it would have taken him to the House of Lords; and he resisted other offers, from the prime ministers Lord Liverpool and George Canning, because they would have taken him to Dublin, the Caribbean, or Calcutta. Palmerston would not forgo the delights of the London society centring around Almack's social club, the three principal hostesses of which—Lady Jersey, the Princess Lieven, and Lady Cowper (whom he married in 1839 after she was widowed)—were all probably his mistresses; he was known widely as "Lord Cupid." Because in 1827 the ultra-Tories, including the Duke of Wellington and Sir Robert Peel, refused to serve under Canning, Palmerston at last reached the Cabinet. But the offer of the exchequer by Canning was withdrawn, and it was as secretary at war that Palmerston sat in the cabinets of Canning, Viscount Goderich, and Wellington (who early in 1828 temporarily reunited the Tories).

Early career in Parliament

Known initially as a Pittite (but assumed at Cambridge in 1806 to be a supporter of the slave trade) and condemned by Radicals as a follower of Viscount Castlereagh, who was hostile to civil liberties, Palmerston acquiesced in the switch from unmoving resolution to "Liberal Toryism" over which Liverpool presided in 1821–23. But Palmerston was not an intimate of Canning and his follower, the financier and statesman William Huskisson, nor of Peel. Only after Canning's death, when Palmerston applauded "the great strides which public opinion has made in the last few years," could he be considered a Canningite. As such, he resigned, reluctantly, when Wellington drove Huskisson out of office in mid-1828; but he never closed the door to overtures from Wellington, though stipulating first that the Canningites as a body, and then that some Whigs as well, should come in. He had owed his return to the Cambridge seat in Parliament in 1826 to the Whigs—he was opposed by two ministerial colleagues hostile (as he was not) to Catholic Emancipation—and of the Whigs who sat with him in the cabinets of Canning and Goderich, he said that he liked them much better than the Tories and agreed with them much more. Palmerston, whose maiden speech in February 1808 had been a defense of a British attack on the Danish fleet to deny it to Napoleon, became in cabinet entranced with foreign affairs and adopted a position on events in Greece and Portugal in advance of the other Canningites. This he made public in the House on June 1, 1829, when he complained that Wellington and his foreign secretary, the Earl of Aberdeen, had made Britain the keystone of the arch of European absolutism. The circulation of his speech as a pamphlet indicates that Palmerston had now decided to play for high political stakes. By attacking and interrogating the Tory government from a Whiggish point of view all through 1830 and by rejoicing at the Paris revolution of that year, he qualified himself to become foreign secretary when Wellington's resistance to all parliamentary reform led to the creation of a Whig–Canningite coalition under Earl Grey.

Opposition to Wellington

Views on liberalism and conservatism. The Reform Bills of 1831 and 1832 were really more considerable than Palmerston liked, and he tried to modify them. Failing, he blamed "the stupid old Tory party" for making them necessary by refusing minor concessions, emphasized the "final" nature of the 1832 Act, and proclaimed his confidence that the landed interest would continue to prevail in politics as he thought it should. From 1849 to 1865 he came to personify the opposition of the landsmen and many of the middle classes to the enfranchisement of trade unionists and to resist fiscal and legislative assaults on landed property, opining (with regard to Ireland) that "tenant right was landlord's wrong." After one term (1832–35) as member for South Hampshire, he was defeated; and for the rest of his life he represented the Devon market town of Tiverton. There, at a time fixed for the convenience of the London press, he would expound opinions, platitudes, and prejudices, which caused it to be said that "he was not a common man, but a common man might have been cut out of him." A racehorse owner, he combined the conservatism of the countryman with a concern for the expansion of the manufacturer's foreign markets. His standard text was that reform in 1832 had prevented social revolution and that enlightened legislation thereafter (the new poor law of 1834, municipal reform in 1835, educational grants, corn-law repeal in 1846, factory acts, and other reforms, to some of which he contributed as home secretary and prime minister) was producing social peace. This made him proud of his country and more than ever inclined to exhort foreign autocrats and bureaucrats to behave like sensible Whigs and Canningites.

Attempts to export the British government system

Palmerston believed that something like the British system of responsible government would be good for all European states and that it would become the norm (as by the first decade of the 20th century it had). What he hinted at in speeches of 1829–30—that in office he would become an active verbal participant in Europe's ideological war—he confirmed in a speech of August 2, 1832, widely disseminated by German liberals. No English ministry was doing its duty, he declared, if inattentive to the interests of constitutional states, which were Britain's natural allies. Unlike Canning, he had persuaded himself that "the selfish interests and political influence of England were best promoted by the extension of liberty and civilisation," and even, against the evidence, that constitutional governments in Spain and Portugal, for instance, would be pro-British. Rebuked for "missionary diplomacy" not intended to lead to action but to inflame international relations, he retorted that ineffective protest was better than tacit acquiescence in wrong and that opinions were mightier than armies. He was charged with "disturbing the peace of Europe by giving encouragement to every revolutionary and anarchical set of men." For the Liberal and Radical following he gained by such allegations, repeated by the Tories at home and accepted as just by the Court, he was grateful. Yet his opinions on foreign and domestic matters were all of a piece. He did not want democratic or republican, least of all "Red," regimes abroad any more than at home. But he regarded the mission of Lord Minto to the Italian courts, on the eve of the revolutions of 1848, as mediatorial, not inflammatory; its object was to show the rulers of Europe that they should have their 1688 (English revolution against James II, led by William of Orange) and their 1832, lest worse befall them.

Views on nationalism. Palmerston was a British nationalist; he said that the country had no permanent allies, only permanent interests. The idea that because he applauded the cause of Liberalism in Europe, he wished to tear up the Treaty of Vienna in the interests of national self-determination, is nonsense. It was true that he was instrumental in securing confirmation of the independence of Greece and Belgium and the removal of Lombardy from the Austrians; but for Polish, Magyar, and Romanian patriotic causes he lifted not a finger, and he surprised the German liberal nationalists of Frankfort, in 1848, by a refusal of patronage, largely because German nationalism conflicted with the Vienna frontiers in Schleswig-Holstein.

Palmerston was a philhellene; but by the time he became

foreign secretary the only question was whether Greece should be a viable size, wholly independent of Turkey and under the surveillance of Britain, France, and Russia. By 1832 he had achieved this objective, only making the mistake of providing Greece with a king—the Bavarian Otho—who was not prepared to administer a constitution of London–Paris design; this mistake was remedied in 1863, when the Dane George I became king of Greece.

The Belgian revolt of 1830 was a *fait accompli*, and it had become a British interest to secure Dutch recognition of it without allowing the French to profit by intervening. In this matter, as chairman of the London Conference, Palmerston first showed his diplomatic proficiency. Twice the French marched—but with British goodwill only because the French were pledged to withdraw when they had dealt with the Dutch; "the French must go out of Belgium, or we have a general war . . . in a given number of days," wrote Palmerston on August 17, 1831. The outcome was an independent constitutional Belgium, with its neutrality guaranteed by the Five Powers in a famous "scrap of paper."

Palmerston's view of Austria

If he wanted Italian federation or unification—in 1848–49 as well as in 1859–61—it was from no addiction to the national principle in the abstract; and, if he wanted the Austrians out of Italy, it was not primarily because they were illiberal. His view was that Austria had been put into northern Italy in 1815 to provide, with the kingdom of Sardinia, a barrier against French aggression. Through mismanagement the Austrians had contrived to raise so much "national hatred" against themselves that their presence in Italy was an incitement to the French and a danger to the general peace, and it was weakening the Habsburg Empire as well. An able speech in Parliament on July 21, 1849, gave the coldest comfort to the Hungarians, against whose bid for independence Austria had to seek Russian aid. Palmerston said, wholly sincerely, that "the political independence and liberties of Europe are bound up . . . with the maintenance and integrity of Austria as a great European Power." Austria was, after all, Britain's natural ally in the Balkans, because of a lack of conflicting interests.

Palmerston's fears of France and Russia. Of the Great Powers, Palmerston felt that only Russia and France might directly threaten British interests, which he interpreted very widely and in which he certainly included all the routes to India and the Far East via the Mediterranean; from concern for India sprang Persian and Afghan wars as well as the Crimean War (see below). It was Palmerston's objective never to find France and Russia arrayed together against Britain and to practice, whenever direct confrontation could be avoided, the technique of "restraint by cooperation." The France of Louis-Philippe (ruled 1830–48) acted for most of the '30s as Britain's ally, and Palmerston's riposte to Metternich's coalition of the three emperors (of Austria, Prussia, and Russia) at Münchengrätz in 1833 was the Quadruple Alliance of Britain and France with the constitutional parties in Spain and Portugal, which were engaged in war with absolutist pretenders backed by Metternich. This alignment of 1834 was intended to "defeat the Holy Alliance in the arena which they themselves had chosen." But the French became irked at the restraining element in British cooperation and did not see why they should not be as predominant in Spain as the British were in Portugal. Relations, therefore, deteriorated even before there was an open breach in 1839–40 on the Eastern Question (regarding the Ottoman Empire). In 1833 Palmerston had been unable to act when Russia, by saving the Ottoman sultan from Muḥammad ʿAli Pasha, the ruler of Egypt, had gained predominance at Constantinople. At the end of the decade, Palmerston's fear that the tsar and Muḥammad would divide the Turkish Empire between them gave way to the fear that Muḥammad would aggrandize himself as the protégé of France. Palmerston's mobilization of the Powers to isolate France and confine Muḥammad to Egypt gave him (1840) a major diplomatic and parliamentary triumph, achieved amid the doubts and fears of Cabinet colleagues and the downright opposition of some.

Triumph over Muḥammad ʿAli Pasha

Relations with France were unnecessarily bad when Palmerston left office in 1841. His condemnation of Lord Aberdeen (Tory foreign secretary, 1841–46) for appeasing France and the United States also contributed to a feeling in the highest Whig circles that he ought not to return to the foreign office; and his refusal to take any other appointment was made the excuse for the prime minister Lord John Russell declining to form a government to repeal the Corn Laws in December 1845. In mid-1846, when Russell did form a government, Palmerston became foreign secretary again. He was worsted by the French in the Spanish Marriages affair (the son of Louis-Philippe of France married the sister of Isabella II of Spain) and also tangled with the French minister François P.G. Guizot and the princess Lieven over Switzerland and Italy, when the French showed signs of alliance with the illiberal Powers. After the revolution in 1848, as in 1830, Palmerston was concerned both with protecting the new French regime and deterring it from going to war (*e.g.*, in Italy). He and the tsar, both standing for the Treaty of Vienna and the balance of power, saluted one another from the twin rocks that stood amid the revolutionary tide.

The popular hero. In 1848–49 Palmerston was more intent upon preserving the general peace than upon patronizing Liberalism, but in 1849–51 he won Radical applause for his denunciations of the cruelty of counter-revolutionaries; for his release of British arms to Sicilian insurgents (for which he had to apologize to King Ferdinand II, 1849) and his later endorsement of William Ewart Gladstone's exposure of Ferdinand's treatment of political prisoners; by his evident approval of the hostile reception given to the Austrian general Julius, Freiherr von Haynau (known as Hyena), the villain of the counterrevolutionary repression in northern Italy and Hungary, when, against Palmerston's advice, he visited Britain in 1850; by his pressure on the Turks, first, not to yield up and, later, to release, Polish and Hungarian refugees; by his acceptance, when the defeated Hungarian patriot Lajos Kossuth visited Britain, of addresses describing the rulers of Austria and Russia in Kossuthian terms (1851).

This propagandist diplomacy infuriated Prince Albert, consort of Queen Victoria, whose moderating influence Palmerston sought to evade, and it embarrassed Cabinet colleagues who, like the Queen, were not kept fully informed. But Palmerston defeated Russell's intention of removing him from the foreign office by a famous dusk-to-dawn speech on July 8, 1850, in which, against the weight in the Commons, he defended the British bombardment of Athens, where the regime had proved less liberal than Palmerston had hoped, and the sabotaging of an agreement reached in London with France and Russia over British subjects' claims against Greece—including debts allegedly owed to an Iberian Jew, Don Pacifico. Exploiting with great skill every vein of British chauvinism, he "fearlessly challenge[d] the . . . House as representing . . . a commercial, a constitutional country," on the issue of whether a British subject in whatever land he might be could, like the Roman of old, "hold himself free from indignity when he could say 'Civis Romanus sum.' " His popularity as "the most English minister who ever governed England" was such, especially outside Parliament, that Russell did not dare dismiss him until December 1851, when Palmerston, to stand well with the ruler of France, approved the coup d'état by which President Bonaparte overthrew the constitution of the Second Republic.

Palmerston's defense of his policies

Palmerston at once brought about the fall of the Russell government and might have joined the minority Tory government if the prime minister Lord Derby had been willing to abandon his protectionist policies. He served as a reforming home secretary in a Peelite–Whig coalition under Aberdeen, which in 1854 took Britain into the Crimean War against Russia and allied with France in defense of Turkey. Often accused of what was later called "brinkmanship," Palmerston benefitted by the public opinion that if he had been in control, Russia would not have taken the risks that led to war with the West; and his resignation in December 1853, avowedly in opposition to

Russell's reform bill, was ascribed (especially by newspapers primed by Palmerston) to discontent with an infirm diplomacy. A switch to a more belligerent posture was regarded as the price of his immediate return.

Premierships. Under Aberdeen, Palmerston was a more loyal and reasonable colleague than was Russell. When Russell resigned as leader of the House of Commons because he would not oppose a motion for inquiry into the misconduct of the Crimean War—which was carried by the staggering majority of 305 to 148 (January 1855)—Palmerston succeeded him. With public opinion behind him, Palmerston became prime minister. He had to grant the inquiry into the war (and so lost the support of most of the Peelite ministers). His attempts to galvanize the war effort and remedy gross defects in many branches of the services were partly nullified by bad appointments at home and in the Crimea and by a characteristic refusal to bow to the anti-aristocratic campaign for administrative reforms. He was pressured by the French to make peace (1856) on terms he thought inadequate but which forced Russia to give up its control of the mouth of the Danube. He submitted to restraints by colleagues in quarrels with the United States, which he disliked as a democracy, as a competing imperialism, and as less than wholehearted in the suppression of the Atlantic slave trade, on which he himself was a passionate crusader. But when the Tory Opposition, with Peelites and Cobdenites (followers of the free-trade activist Richard Cobden), narrowly defeated him on the China War (the second in which the men in China had involved him), Palmerston confidently appealed to the electors against "an insolent barbarian" at Canton violating British persons and property. (Palmerston was not an imperialist in the late-19th-century sense, though he was responsible for adding Aden and Hong Kong [1841] and Lagos [1861] to the British Empire, but he never would admit the right of any part of the world to hold aloof from the impact of Western commerce and civilization.) The considerable majority achieved in the April 1857 election was a personal triumph, but it melted away when he did not make the lion's roar sufficiently loud in response to French attacks on Britain for harbouring refugees conspiring to murder Napoleon III; and Palmerston's government resigned after defeat in the House (February 1858).

First premiership

After the election of 1859 denied the Tories a majority, Palmerston resumed the premiership (June), in harness with Russell and the Peelite Gladstone, all being pro-Italian against Austria. This triumvirate ruled until Palmerston died, though there was often a prospect of Gladstone resigning in opposition to naval defense expenditure (on which Palmerston insisted), which cut across his passion for economy, and there were eventually differences of opinion on parliamentary reform and the Irish Church. Palmerston knew, however, that he would be able to rely on the Tories for support if Gladstone resigned and linked himself with the Radical John Bright. Although his determination that Britain "should count for something in the transactions of the world" was successfully challenged by Bismarck in the Schleswig-Holstein affair, 1863–64, Palmerston retained great prestige at home; and on the eve of his death he greatly enlarged the Liberal majority in an election on the cry "Leave it to Pam." It was rightly said, after his death, that "the exceptional sway of Lord Palmerston could not be reproduced by any other statesman, or any combination" and that "the reign of *moderate* Liberalism" was over. He had been a conservative statesman using radical tools and keeping up a show of liberalism in his foreign policy; after him the defense of the Conservative cause would revert to the Conservative Party.

Second premiership

BIBLIOGRAPHY. The official *Life*, 5 vol. (1870–76), was by LYTTON BULWER (vol. 1 and 2 and a skeleton of vol. 3) and EVELYN ASHLEY. Original sources were used by H.C.F. BELL in *Lord Palmerston*, 2 vol. (1936); and JASPER RIDLEY in *Lord Palmerston* (1970), neither of whom convey either the animation or the inner consistency of the man. Some are published in PHILIP GUEDALLA (ed.), *Gladstone and Palmerston* (1928); and BRIAN CONNELL (ed.), *Regina vs. Palmerston: The Correspondence Between Queen Victoria and Her Foreign and Prime Minister, 1837–1865* (1961). Guedalla's *Palmerston* (1926) is a piece of artistry, apt to mislead, unlike the less pretentious *Lord Palmerston* of W. BARING PEMBERTON (1954). WALTER BAGEHOT'S essay in *Biographical Studies* (1856); ANTHONY TROLLOPE, *Lord Palmerston* (1882); and the MARQUESS OF LORNE, *Palmerston* (1892), are still worth reading. Of works which are not avowed biographies, DONALD SOUTHGATE, *"The Most English Minister" . . . , the Policies and Politics of Palmerston* (1966), has as its subject Palmerston's effect upon British foreign policy, 1829–65; also expounded in BASIL KINGSLEY MARTIN, *The Triumph of Lord Palmerston* (1924). SIR CHARLES WEBSTER, *The Foreign Policy of Palmerston, 1830–1841*, 2 vol. (1951), has found no successor. A.J.P. TAYLOR'S first and best book, *The Italian Problem in European Diplomacy 1847–1849* (1934); and CHARLES SPROXTON, *Palmerston and the Hungarian Revolution* (1919), are perceptive; and GAVIN B. HENDERSON, *Crimean War Diplomacy, and Other Historical Essays* (1947), corrective. DEREK BEALES, *England and Italy, 1859–60* (1961); W.E. MOSSE, *The European Powers and the German Question, 1848–71* (1958); and E. PREVELAKIS, *British Policy Towards the Change of Dynasty in Greece, 1862–1863* (1953), are illuminating in their several fields.

(Do.S.)

Pamir Mountain Area

The Pamir mountain area belongs mainly to the Gorno-Badakhshan *avtonomnaya oblast* (autonomous region) of the Tadzhik Soviet Socialist Republic of the U.S.S.R. It is surrounded by the Central Asian mountain systems of the Hindu Kush, Kunlun, Gissar-Alai and Tien Shan.

The territory of the Pamirs is bounded on the north by the Trans Alai Range (Zaalaysky Khrebet); on the east by the Sarykol Range, which forms the border between China and the U.S.S.R.; on the south by Lake Zorkul, the Pamir River, and the source of the Pyandzh River bordering Afghanistan; and on the west by the north–south segment of the Pyandzh Valley. The eastern parts of the Peter I and Darvaz ranges on the northwest are included within the Pamirs, since their glaciers comprise a single system with the glaciers of the northwestern Pamirs. The derivation of the word Pamir has not been definitely established.

Relief. The Pamirs are a combination of east–west and north–south ranges, with the former predominating. The east–west Trans Alai Range, which forms the northern frame of the Pamirs, falls steeply to the intermontane Alai Valley. The high central part of the range, between the Tersagar Pass on the west and Kyzylart on the east, averages between 19,000 and 20,000 feet, reaching its highest point at Lenin Peak, 23,400 feet (7,134 metres). South from the Trans Alai extend three north–south ranges. Of these, the western, the Academy of Sciences Range, and the central, Zulumart, are relatively short; and the eastern, Sarykol Range, forms the eastern border of the Pamirs. The area east of the Sarykol Range is sometimes called the Chinese Pamirs.

Principal mountain ranges

The north–south Academy of Sciences Range enters into the northwestern Pamir system, where it rises into a huge barrier, reaching 24,584 feet (7,495 metres) in Communism Peak (the highest point in the U.S.S.R.). The eastern slope of the range is covered on the south face by the Fedchenko Glacier. The western slope intersects other ranges that lie still farther to the west: the Peter I Range, with Moscow Peak (22,300 feet); the Darvaz Range, with Arnavad Peak (20,000 feet); and the Vanchsky and Yazgulemsk ranges, with Revolution Peak (22,900 feet). The ranges are separated by deep ravines. To the east of the Yazgulemsk Range, in the central portion of the Pamirs, is the east–west Muzkol Range, reaching 20,400 feet in the Soviet Officers Peak. South of it stretches one of the largest ranges of the Pamirs, called Rushan on the west and Bazar-Dara or Northern Alichur on the east. Still farther south are the Southern Alichur Range and, to the west of the latter, the Shugnan Range. The extreme southwestern Pamirs are occupied by the Shakhdarin Range, composed of north–south (Ishkashim Range) and east–west elements, with the Mayakovsky Peak (20,000 feet) and Karl Marx Peak (22,100 feet). In the extreme southeast, to the south of Lake Zorkul, lies the east–west Vakhan Range.

It is customary to divide the Pamirs into a western area

Eastern and western Pamirs

and an eastern area, distinguished by their forms of relief. In the eastern Pamirs a medium-mountain relief predominates on a high raised foundation. While the heights above sea level average 20,000 feet or more, the relative heights of the peaks above their foundation do not in most cases exceed 3,300–5,900 feet. The ranges and massifs have mainly rounded contours, and the wide and flat-bottomed valleys and troughs between them, situated at heights of 12,100–13,800 feet, are occupied either by quietly running, meandering rivers or by dry channels. The valleys and slopes of the ranges are covered by layers of loose material.

In the western Pamirs the relief is high mountain and sharply disjointed, alternating between low ranges and alpine ridges capped by snows and glaciers; and there are deep, narrow ravines with high, rapid rivers. The valleys and depressions are filled with debris, so that almost the only suitable places for human settlement are the alluvial fans in the valleys of tributaries of the Pyandzh River. The transition from the eastern-Pamirs type of relief to the western-Pamirs type occurs gradually. The conventional boundary is a line joining the ridge of the Zulumart Range with Kara-Bulak Pass in the Muzkol Range; from Pshart Pass, it follows the ridge of the Northern Alichur Range to Lake Zortashkol, where it turns south to the valley of the Pamir River.

Geology and climate. Geologists divide the Pamirs into three zones according to the characteristics of their rock formations: the northern, central, and southern Pamirs. The southern zone consists of metamorphic rocks (gneiss, quartzite, marble, and others) to which a majority of researchers attribute a Precambrian age (more than 570,000,000 years ago). The zone on the whole represents a huge anticlinorium, or series of stratified arches. The central zone of the Pamirs contains limestone, sandstone, and shale rocks of the Jurassic, Triassic, and Permian periods (136,000,000 to 280,000,000 years ago) and also red-coloured terrestrial rocks of the Lower Cretaceous Period (136,000,000 years ago). There are some marine rocks of the Lower and Middle Paleozoic Era (345,000,000 to 570,000,000 years ago) and lava and tufaceous rocks of the Paleocene (54,000,000 to 65,000,000 years ago). The structure of the central Pamirs is that of a huge synclinorium (an inverted arch caused by fracturing); it is separated from the northern Pamirs by a deep fracture.

In the structure of the northern Pamirs, two subzones can be discerned: a Paleozoic zone, which stretches to the ridge of the Trans Alai Range, and a zone beyond, which is composed of more recent deposits. The Paleozoic subzone of the northern Pamirs is a huge anticlinorium with a complex internal structure. It is separated from the Trans Alai subzone by the Karakul fracture. The Trans Alai subzone is very complex. Its western part is a fan-shaped anticlinorium, in the centre of which emerge Jurassic deposits; radiating outward are more recent, dislocated rocks of the Lower Cretaceous Period. The eastern part has Cretaceous and Paleocene deposits in a system of conflicting folds. Because of the numerous overthrusts, or horizontal faults, in some places the layers overlap each other. On the north the Trans Alai subzone is bounded by the deep Gissaro-Trans Alai fracture, separating the Pamirs from the Gissaro-Alai system.

Severity of climate

The climate of the Pamirs is arid and continental. These features are more pronounced in the eastern part, where there are broad, closed basins in which cold air is retained and a barrier of high ranges intercepting moist air currents. There is a mixed circulation of air—western (cyclonic) and southern (monsoonal). In the valleys of the western Pamirs the amount of annual precipitation is four to ten inches, and in the eastern Pamirs two to five inches. In the high altitudes and on slopes of mountains, the amount of precipitation increases, reaching 32 to 40 inches on the Fedchenko Glacier. The thickness of the snow blanket in the western Pamirs reaches 20 to 28 inches, and in the eastern it reaches 1.5 to four inches. The average January temperature in the eastern Pamirs at heights around 11,500 feet is 0° F (−17.8° C). Winter here is long (October through April) and severe, and temperatures have been recorded at below −58° F (−50° C)—in Bulunkul in 1953, −64.9° F (−53.3° C). In the short summer, temperatures do not rise above 68° F (20° C). The climate of the valleys of the western Pamirs is milder. The average temperature in January at heights of about 6,900 feet is 18.7° F (−7.4° C), and in July it is 72.5° F (22.5° C), but temperatures vary greatly. The period of vegetation (with temperatures of 41° F [5° C]) is 223 days in Khorog and 140 days in Murgab.

The Pamirs have 1,085 glaciers, covering an area of 3,105 square miles. The largest centres of glaciation occur in the Academy of Sciences, Trans Alai, Rushan Northern Alichur, Yazgulems, Darvaz, Peter I, and Zulumart ranges. The largest valley glacier, Fedchenko (length 44.3 miles), starts in the Academy of Sciences Range; other glaciers in this range include the Grumm-Grzhemaylo (23 miles), Garmo (17 miles), Sagran (15 miles), and Geographic Society (13 miles). The main glaciers of the Trans Alai Range are the Sauk-Dara (16 miles), Korzhenevsky (around 13 miles), October (10.9 miles), and Lenin (5.9 miles). In the eastern Pamirs shallow bog glaciers and snow fields predominate.

The rivers belong mainly to the basin of the Amu Darya (the ancient Oxus), which forms the frontier between the U.S.S.R. and Afghanistan and flows into the Aral Sea. Its upstream extensions are the Pyandzh and the Pamir. The area also contributes to the basin of the Tarim, which flows eastward into China. Lakes include Karakul (a salt lake), Rangkul, Shorkul, Zorkul, Yashilkul, and Sarez.

Mountain flora

Plant and animal life. Vegetation in the Pamirs, especially in the eastern Pamirs, is poorly developed. Bare cliffs or a cover of surface rubble predominate. At high altitudes the vegetation changes with increasing altitude from mountain desert flora to mountain drought-resisting plants, then to mountain steppe vegetation, and finally to cold-resistant plants. Native forms predominate, but some forms from Central and Southwest Asia are found. The eastern Pamirs are a cold, high-mountain desert, with woody vegetation completely absent and with low-growing plants that are adapted to the severe conditions. On the dry mountain slopes there are low shrubs of winter fat, the only form of vegetable fuel in the area, and plant cover, such as the Pamir tansy, oxytropis, astragalus, local species of wormwood, the bulbous iris, and meadow grass. On the bottoms of the moist valleys, sedge and cobresia meadows are abundant.

The vegetation of the western Pamirs is richer, although on the mountain slopes and bottoms of valleys there is a prevalence of wormwood and haloxylon. At heights over 8,500 feet, spiny pillow-form plants (*e.g.*, acanthus, spiny astragalus) are widespread. At heights over 10,500 feet, yugan, kamol, fescue, and feather grass are found. At 12,500 to 14,100 feet are alpine cobresia meadows; above 14,500 feet vegetation is sparse. Along the rivers of the western Pamirs are dense growths of willow, sea buckthorn, birch, poplar, and hawthorn. A thinner wood–shrub vegetation reaches heights of 12,500 feet, including Juneberry, almond, birch, and juniper. On irrigated lands there are cultivated plantings of grape, apricot, apple, pear, walnut, and mulberry trees.

The fauna of the Pamirs are not numerous. In the eastern Pamirs are found the *arkhar* (a mountain sheep), the long-tailed marmot, and the large-eared Tibetan wolf. Birds include the Tibetan mountain turkey, the Tibetan raven, the horned lark, and the snow vulture. The western Pamirs have the mountain goat (*kiik*), brown bear, wolf, fox, snow panther, lynx, weasel, marten, *zayats-tolay* (a hare), dormouse, and flying mouse. Birds include the Indian oriole, bluebird, dark-breasted pheasant, stone partridge, and paradise flycatcher. The Pamirs have few fishes; only the carp and the Tibetan loach are known.

Population and industry. Some 98,000 people live in the Pamir area, and the average population density is four persons per square mile. More than 90 percent of the population are Tadzhiks living in the western Pamirs. Their languages belong to the Iranian group, and those who are religious are Shīʿite Muslims. The inhabitants of the eastern Pamirs are Kirgiz, who speak a Turkic language and are Sunnite Muslims.

Agriculture

Nearly all are peasants whose chief occupation is farming and livestock breeding (in 1969 there were 52 collective farms and one state farm). The small amount of arable land is planted with grain, beans, gourds, and potatoes, and there are orchards of apples, pears, apricots, and mulberry trees. Sheep and goats are the predominant livestock.

Industry in the Pamirs includes several small hydroelectric power stations and a few mines. In the eastern Pamirs, brown coal and common salt are mined. There is evidence of industrial deposits of gold, gems, jasper, lazulite, mica, asbestos, and talc. Thermal and mineral springs are common. There are auto routes across the Pamirs from Dushanbe in the west to Khorog in the south and from there to Osh in the east. Another route runs from Khorog south through the once impassable ravine of the Pyandzh–Pamir rivers. Buses go from Khorog to several regional centres. Many other routes are accessible only to pedestrians and pack animals.

Exploration

Modern exploration began with the Russian A.P. Fedchenko, who in 1871 succeeded in reaching the northern foot of the Pamirs from the Alai Valley. In 1877 the Russian geologist I.V. Mushketov visited the valley of the Muksu River and the vicinity of Lake Karakul. In 1877–78 the Russian zoogeographer N.A. Severtsov, penetrating into the depth of the mountain country, made a chart of its structure. In 1878 an expedition under the Russian naturalist V.F. Oshanin discovered the large valley glacier subsequently named A.P. Fedchenko. From 1884 through 1887, the Pamirs were explored by the zoologist G.Y. Grum-Grzhimaylo, who provided valuable data on glaciers of the northwestern and northeastern Pamirs.

Under the Soviets, explorations in the Pamirs have become systematic. In 1928 an expedition of the Academy of Sciences of the U.S.S.R. explored the region of the Fedchenko Glacier, making possible the first accurate topographical maps of the western Pamirs. In 1933 the first high-mountain glaciological observatory in the world was constructed on the Fedchenko glacier, at a height of about 14,800 feet. The Tadzhik–Pamir Expedition of 1932 resulted in important monographs on the geology, geomorphology, and hydrogeology of the Pamirs. Soviet alpinists have contributed much to the investigation of the Pamir region—a part of the world not easily accessible. (T.K.Z.)

BIBLIOGRAPHY. OLE OLUFSEN, *Through the Unknown Pamirs: The Second Danish Pamir Expedition, 1898–99* (1904); MALCOLM SLESSER, *Red Peak: A Personal Account of the British-Soviet Pamir Expedition 1962* (1964).

Panama

The Republic of Panama is a country in Central America, situated on the S-shaped Isthmus of Panama, which joins North and South America. Its area is 29,783 square miles (77,138 square kilometres), of which 29,208 square miles is on the mainland and 575 square miles is divided among the islands. The area given does not include the 558 square miles of the Canal Zone, a ten-mile-wide strip of land stretching across the Isthmus; through the zone runs the Panama Canal, connecting the Atlantic and Pacific oceans. Panama has granted to the United States jurisdictional rights over the zone, limited to the specific purposes of maintenance, operation, sanitation, and protection of the interoceanic waterway.

Location and boundaries

Panama extends for 384 miles from the Costa Rica frontier in the west to the Colombian border in the east. The shortest distance across the Isthmus is 31 miles, stretching from the mouth of the Río Nergalá (Necategua), which flows into the Golfo de San Blas on the northern (Caribbean) shore, to the mouth of the Río Bayano on the Pacific coast. Panama is bounded on the north by the Caribbean Sea, on the south by the Pacific Ocean, on the east by Colombia, and on the west by Costa Rica. Panama City, the capital, is located on the Pacific coast to the east of the Panama Canal. The state was ruled in the early 1970s by a provisional government junta.

Economically, Panama is a developing country: most of the population is engaged in agriculture and cattle raising. Although there is considerable international trade because of the country's strategic location, industry in the early 1970s was just beginning to develop. (For an associated feature, see PANAMA CANAL; for historical aspects, see CENTRAL AMERICAN STATES, HISTORY OF THE.)

THE LANDSCAPE

The natural environment. *Physiography.* The Panamanian landscape consists of three distinct areas: the lowlands, or hot lands, lying at altitudes below 2,300 feet, which are 87 percent of Panamanian territory; the lands of the temperate zone, situated at altitudes of between 2,300 and 4,900 feet, which comprise about 10 percent of the land; and the highlands, or cold lands, lying more than 4,900 feet above sea level and covering 3 percent of the whole.

The land surface may also be divided into two geological categories—the highlands and mountains of volcanic origin and the lowlands, plains, and hills of sedimentary origin.

The highlands include the Barú, or Chiriquí, volcano, 11,408 feet (3,478 metres) high; the central mountain range, a continuation of the Cordillera de Talamanca (Costa Rica), which stretches to the central part of the Isthmus at elevations of between about 3,300 and 10,000 feet; the northeastern mountain arc, including the Cordillera de San Blas; the southeastern mountain arc, including the summits of Chimán, Río Congo, Serranía del Sapo, and Pirre; and the southern volcanic chains, including the mountain systems of the Península de Azuero and the Macizo (massif) de Canajagua.

The lowlands include the plains of the provinces of Panamá and Chiriquí, the plains and hills of the central Isthmus of Panama; the eastern section, occupied by the Bayano, or Chepo, and the Chucunaque river basins; and the northern plains of the Caribbean region.

Panama's Pacific seaboard is 1,015 miles in length, and its Atlantic seaboard extends for 774 miles. On its continental shelf, wide in the Pacific and narrow in the Atlantic, are more than 1,600 islands. Its coasts have many bays, gulfs, peninsulas, headlands, and capes.

The principal archipelagoes off the Caribbean coast are those of Bocas del Toro and Mulatas (San Blas). Off the Pacific coast the principal islands are the Archipiélago de las Perlas (Pearl Islands), Taboga, Taboguilla, and—largest of all Panamanian islands—Coiba.

Rivers

Drainage and soils. Panama has approximately 500 rivers: 350 flow to the Pacific and 150 to the Caribbean sea. Most of the rivers have their sources in the highlands. They tend to erode their banks upstream, forming meanders in the middle parts of their courses and deltas and inlets at their mouths. Their volume of flow increases during the rainy season (May to December) and decreases during the dry season (from January to April). None of the rivers is navigable, except the Tuira (which is navigable for about 100 miles), and the Bayano (navigable by small boats for 35 miles). The most important rivers flowing to the Caribbean are the Sixaola, the Changuinola, the Río Indio, the La Miel, and the Chagres. (The Chagres Basin is the best known because of its use for the maintenance of the Panama Canal.) The most important rivers flowing to the Pacific are the Chiriquí Viejo, San Félix, San Pablo, Santa María, the Bayano, the Chucunaque, and the Tuira.

About half of the isthmian territory is formed of volcanic rock. There are also many clay and slate shales in the eastern and central provinces and some sedimentary rocks in the western provinces, in the central part of the Isthmus, the Canal Zone, and adjacent areas.

Soil types include soft, red clayey soils (situated in Chiriquí province in the southwest), which are easily eroded, containing some humus (decomposed plant and animal matter); hard clay soils, which are dry and unsuitable for cultivation; and savanna (grassy parkland) soils, which are good for pasture, even if they are dried out during the dry season. There are also alluvial soils (clay, silt, sand, and gravel deposited by rivers), limited to small areas

along the rivers, which are suitable for the cultivation of sugarcane, plantain (a starchy, banana-like fruit), and bananas; volcanic ash soils, which are fertile and contain great quantities of organic matter; and swampy soils.

Temperatures

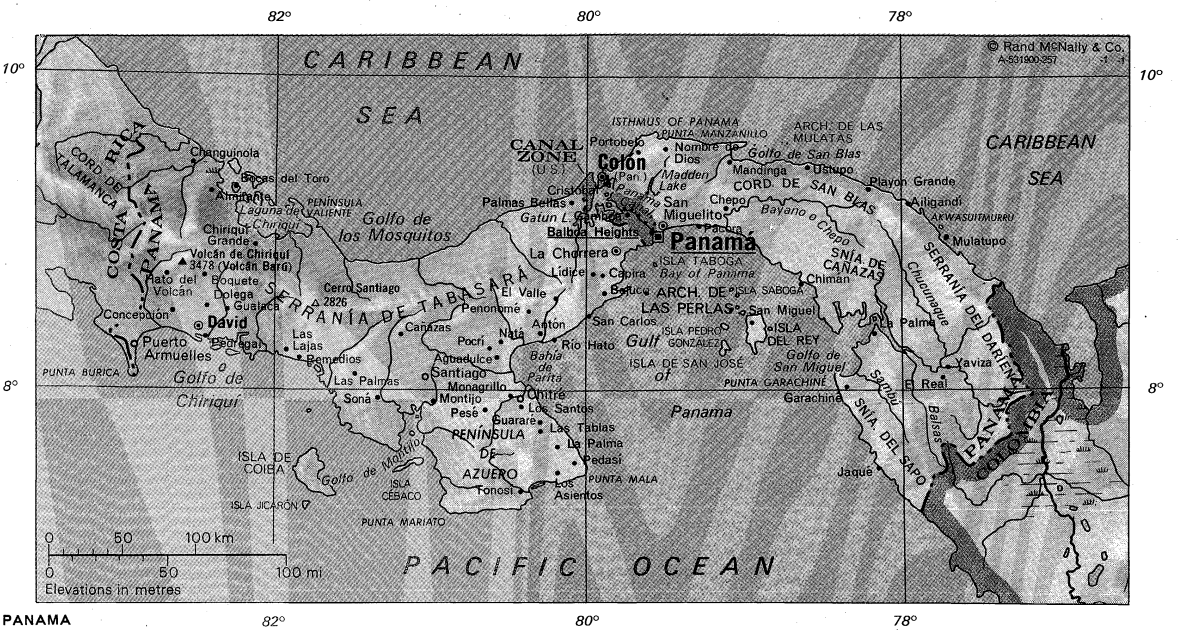
Climate. Panama has a tropical, rainy climate; except for the Caribbean coast and the high mountains, where rain falls at all seasons, there are two well-defined seasons —dry and rainy. The mean temperatures in the Caribbean region are 81° F (27° C) in the dry season and 84° F (29° C) in the rainy season; in the Pacific region the mean temperature is 81° F (27° C) in both seasons. The Caribbean region has the most rainfall, receiving between 59 and 138 inches (150 and 350 centimetres) a year. The more populated Pacific region receives between 59 and 79 inches a year.

Four climatic regions can be distinguished—the very wet tropical zone, the wet tropical zone, the arid tropical region, and the temperate wet zone. Wet tropical forests are found in the Caribbean lowlands, on the Pacific coast, and in the Bayano and Chucunaque depressions; the Canal Zone also experiences the same kind of climate. The arid tropical region is situated on the Pacific coast; it covers the savanna area and has a dry and a rainy season. The temperate wet zone occurs in the highlands of the central mountain range in the west and in the Alto (High) Darién region in the east.

Vegetation and animal life. The tropical rain forest of the lowlands occupies half the country. This vegetation, which includes a variety of species, is suitable for commercial forestry exploitation. The swamp woods, including plantations of mangrove trees, cativo (a large tree of the legume family), and orey (a tall hardwood tree), are abundant.

Tropical savannas are found mainly on the Pacific seaboard. Subtropical woods associated with heavier rainfall occur in the western provinces, in the Cerro Azul and Cerro Jefe regions, in the mountain ranges of San Blas, and in the sierras (mountains) of Alto Darién to the east.

Panamanian animal and plant life is abundant and var-



PANAMA

MAP INDEX

Cities and towns

Physical features and points of interest

ied. The animal life includes poisonous reptiles, crocodiles, armadillos, ocelots (spotted leopard-like cats), small tigers, monkeys, and birds with colourful plumage. Fish of many kinds are plentiful.

Regions of settlement. The jungle regions (lowlands with a rainy, tropical climate) comprise the Chagres Basin, the eastern part of the province of Panamá on the south central part of the Isthmus; the province of Darién in the east; and the Caribbean lowlands. The inhabitants are Negroes and Amerindians. Subsistence agriculture is common.

The savanna region, in the Pacific lowlands to the west of the Canal Zone, has a wet, hot climate. Natural vegetation is rich and varied. Occupied by rural mestizos (people of Spanish and Amerindian descent), the land is intensively cultivated. Rice, beans, sugarcane, plantains, ñame (a type of yam), and yuca (cassava) are grown, and cattle are raised.

The highlands comprise the central mountain range, the region around the Barú or Chiriqui volcano, and Alto Darién, a region of jungles and abundant rainfall. The Chagres, or Route, region covers the central Isthmus of Panama and is crossed by the Río Chagres. Through this zone, where the country is at its narrowest and lowest, runs the Panama Canal (*q.v.*).

The Chagres region also contains the Canal Zone; of its total area of 558 square miles, 163 square miles is occupied by the waters of the Gatun Lake. The zone covers five miles on each side of the canal and does not include the cities of Panama and Colón. The population of the Canal Zone varies periodically; in the early 1970s it was about 45,000, including military personnel and civilians.

The Canal Zone

The Canal Zone is administered for the United States Congress by the Canal Zone Government. The operation of the canal and related activities are in the hands of the Panama Canal Company. Both operate under the direction of the governor, who is directly responsible to the United States secretary of the army, and to the president of the United States; besides his administrative duties in the Canal Zone Government, the governor is the president of the board of directors of the Panama Canal Company. The seat of the government is in the town of Balboa Heights, near the Pacific, close to Panama City. Other towns in the zone are Ancon, on the slopes of Ancon Hill, Diablo Heights, Corozal, Pedro Miguel, Gamboa, Coco Solo, and Majagual–Gold Hill. There are army installations at Ft. Clayton and Ft. William D. Davis. The Canal Zone's ports are Cristobal, close to the city of Colón, and Balboa, on the Pacific. The United States Air Force Canal Zone base, Albrook Field, is also located in the zone on the northeastern outskirts of Balboa Heights.

The population of Panama is predominantly rural, but in the early 1970s the urban population was growing. At the time of the 1970 census, when Panama had a population of about 1,400,000, there were about 9,300 inhabited places, of which 14 were over 5,000. The Panamanian population is widely dispersed, with many people living in hamlets and isolated dwellings. These scattered Panamanians are engaged in subsistence agriculture and cattle raising.

Urban centres

The most important urban centres are Panama City and the city of Colón. Fifty-five percent of the total urban population is in Panama City, which had a total population of almost 350,000 inhabitants, while Colón has over 70,000 inhabitants. Panama City is situated on the Pacific coast overlooking Panama Bay. In addition to being the site of the government and the capital of the republic, it is the centre of industrial, commercial, political, and cultural activities.

THE PEOPLE

Ethnic and linguistic groups. In the 16th century, when the Spaniards first came to the Isthmus, it was occupied by Cuna, Guaymí, Chocó, and other American Indian groups. Out of a mingling of the Spaniards with the Indians, there resulted the mestizo. With the introduction of Africans, other mixed types appeared. During the 19th century, with the construction of the Panama–Colón railroad, new groups arrived—North Americans, French, and Chinese. During the construction of the canal, more North Americans came, as well as Negroes from the Caribbean islands, Spaniards, Italians, and Greeks.

The most numerous of the American Indian groups are the Guaymí, who live in the western provinces of Chiriquí, Bocas del Toro, and Veraguas. Their everyday language is called the *movere*, or the "language of the plains." Next in numbers are the Cuna, who live in the Archipiélago de las Mulatas and on the coast nearby as well as in other parts of the isthmus and along the shores of the Río Tuira and its tributaries; they speak the Cuna language. The Chocó (Chocoe) groups live mainly in the province of Darién. They engage in subsistence agriculture, fishing, and hunting. Their everyday language is the *emberá* dialect.

The Hispanic–Indian group (essentially mestizo) represents the largest population stock in Chiriquí province in the southwest and in the central provinces of Panamá and Colón. This group is mixed with elements of African ancestry in Panama City and Colón. The urban element in these two cities also experiences a strong North American influence from those living in the Canal Zone.

The descendants of the Negro slaves who arrived at the Isthmus during the period of Spanish colonization are distributed throughout Panama. The Antillean-form Negroes are a small minority whose ancestors originally came to Panama to work on the construction of the railroad and the canal.

North Americans have influenced both the economy and culture of Panama. They live in the Canal Zone and in Panama City. Other significant ethnic minorities are the Chinese, Hindus, and Jews, all of whom play an important role in commerce and industry and participate in the country's political and professional life.

Spanish is the official language. Roman Catholicism is the religion of 92 percent of the population. There is freedom of worship.

Demography. A striking attribute of Panamanian demography is the high rate of population increase. Numbers have increased from 336,000 in 1911 to 467,000 in 1930, 805,000 in 1950, and 1,400,000 in 1970.

Population growth

Panama, Area and Population

	area		population	
	sq mi	sq km	1960 census	1970 census
*Republic of Panama**				
Territory (*comarca*)				
San Blas†	1,238	3,206	20,000	25,000
Provinces (*provincias*)				
Bocas del Toro	3,443	8,917	33,000	44,000
Chiriquí	3,381	8,758	188,000	236,000
Coclé	1,944	5,035	93,000	118,000
Colón†	1,644	4,259	85,000	110,000
Darién	6,488	16,803	20,000	23,000
Herrera	937	2,427	62,000	73,000
Los Santos	1,493	3,867	71,000	72,000
Panamá	4,360	11,292	372,000	577,000
Veraguas	4,280	11,086	132,000	152,000
Total Republic of Panama	29,783‡	77,138‡	1,076,000	1,428,000§
Canal Zone‖				
Total Canal Zone	372 558	964 1,445	42,000	44,000

*Excluding the Canal Zone. †The Territory of San Blas is administratively part of the Province of Colón. ‡Total includes 575 sq mi (1,488 sq km) of offshore islands not accounted for in the areas of the provinces. §Figures do not add to total given because of rounding. ‖Of the two total area figures, the first is land area, the second is total, including non-tidal inland water.
Sources: Official government figures; Panama Canal Company.

The Panamanian population is very young (the average age is 19) with a high fertility rate. The rate of population growth is about 3.3 percent a year, largely the result of the high birth rate and a low mortality rate.

A second cause of growth has been immigration, which was substantial during World War II and then decreased. In addition, there is migratory movement within the country of rural people in search of land and of young people moving to the city in search of better conditions.

Eighty percent of the population is in the Pacific area, mainly in the southwest of the Isthmus. The greatest flow of population is from the villages to the cities of Panama and Colón.

Excluding the inhabitants of the Canal Zone, the population density is about 50 persons per square mile.

THE NATIONAL ECONOMY

In the early 1970s the production of goods and services was increasing at the rate of about 8 percent annually.

The majority of the economically active population is engaged in agriculture, forestry, cattle raising, fishing, commerce, and general services. Some mineral resources are exploited.

Resources and economic activity. *Mining.* Cement and salt are the leading mineral products, although gold and silver are also produced on a small scale. There are also commercially significant deposits of manganese, copper, iron, bauxite, and limestone. Plans had been made, in the early 1970s, to exploit copper in the Petaquilla area of Colón province. Oil has been found in several places. There are some outcrops of lead, zinc, rutile, and sulfur, but all are less important. Stones used in construction, such as slate, shale, basalt, andesites, and clays, are common, as are such semiprecious stones as quartz, jasper, and agate.

Forestry. The use of woods for industrial purposes is not completely developed. The Ministry of Agriculture and Cattle Raising, through a specialized organization, supervises the exploitation of the national forests. In the early 1970s, mahogany, oak, cedar, and other species were used to supply some 40 sawmills.

Energy. Electricity is provided by the Institute of Hydraulic Resources and Electrification (IRHE). The IRHE has built a hydroelectric plant at La Yeguada in the province of Veraguas, which in the early 1970s produced about 95 percent of Panama's hydroelectric output of 34,000 kilowatts; at that time several other irrigation or hydroelectric projects were also under study.

Agriculture. A seminomadic and subsistence type of agriculture coexists alongside large commercial enterprises using modern agricultural methods. Land tenure is characterized by the scarcity of rented land, only 6 percent of the total.

Principal crops

The most common agricultural products are rice, corn (maize), beans, sugarcane, plantain, and oranges. The growing conditions of the isthmus favour the cultivation of tropical fruit, although commercial production is limited. The industrial cultivation of coffee, cocoa, sugarcane, tobacco, and sesame was increasing in the early 1970s. Other products, such as potatoes, yuca, tomatoes, and other vegetables, are grown especially in the temperate zone, which includes Volcán, Cerro Punta, Boquete, and the Antón Valley.

Livestock raising (cattle, pigs, and poultry) is an important and long-established economic activity. The large cattle-raising farms are in the southwest of the Isthmus. The quality of meat produced has been greatly improved by better stock and better pasture. There are more than 1,300,000 head of cattle, mainly in Chiriquí, Veraguas, and Los Santos provinces. The production of pigs has increased very little in spite of a growing urban market and relatively high prices. The largest producers are the provinces of Chiriquí, Los Santos, and Veraguas. The largest centres for the production of chickens and eggs are Panamá, Chiriquí, and Veraguas.

Fisheries. Fishing has developed rapidly as a commercial venture. Shrimps are one of Panama's most important exports, with up to 7,600 tons of shrimp caught yearly. Lobster fishing has potential for development. The fishing centres are Bocas del Toro, Chiriquí, and the northeast of the Gulf of Panama on the Pacific coast.

Industries. The leading industries are engaged in cement manufacture (in the limestone areas near the Trans-Isthmian Highway), sardine processing, sugar refining, processing of general food products, banana and cocoa production, oil refining, and the production of gas and electricity. Other important industries include clothing and shoe manufacturing, breweries, distilleries, and the production of furniture, paper, hides, chemicals, and cigarettes.

There were almost 600 manufacturing concerns in Panama City in the early 1970s, with considerably fewer in Colón and David. The salt deposits in Aguadulce and Guararé are the basis of a significant industry. Tourism plays a major role in the country's economy, providing funds for industrial development.

Financial services. Financial services are provided by representative branches of some major Latin American, North American, and European banks. Some of these banks have branches in the provinces and provide loans for industrial, agricultural, and cattle-raising ventures.

Imports and exports. The main products purchased from abroad are manufactured goods, machinery, transportation equipment, food products, chemicals, and textiles. Most of the imports come from the United States, Venezuela, and Japan; some also come from South American countries. Exports are limited to food products, such as bananas, shrimps, coffee, cocoa, and petroleum products. The United States takes a large part of these exports; other notable buyers of Panamanian products are West Germany, The Netherlands, Canada, and the Panama Canal Zone itself.

Buying more from abroad than it exports, Panama achieves equilibrium in its commercial balance from "invisible revenues," resulting from the employment of Panamanians in the Panama Canal Zone and from the sale of goods and services to people from the Canal Zone.

The Zona Libre de Colón (Colón Free Zone), established at the northern end of the Canal Zone, has become increasingly important as a re-export centre. Its commerce is composed of chemical products, manufactured articles, machinery, and transportation equipment. Most of these products come from the United States, Japan, England, and West Germany and are sent to the countries of Central and South America, as well as to Panama itself.

Trade unions. The most important labour unions and associations are the Confederation of Workers of the Republic of Panama, the Isthmian Federation of Christian Workers, the Transportation Union, the Printers' Union, the Union of Hotel Workers, and the Commercial Employees Union. In the early 1970s the labour movement was tending toward the creation of a single workers' union. Employers are represented by the National Commission of Private Enterprises (CONEP), which also encourages the commercial development of the country's resources.

Management of the economy. The government is promoting Panama's growth. Among its programs are reforestation, development of commercial fishing, exploration and exploitation of mining deposits, improvement of agricultural techniques, improvement of transportation and communication systems, and electrification.

Developmental schemes

Transportation and communications. *Roads.* The main transportation routes extend from the capital to the central provinces (Coclé, Herrera, Los Santos, and Veraguas) and to the western section of the country (Veraguas and Chiriquí). The principal roads are the Trans Isthmian, or Boyd–Roosevelt, Highway, joining Panama with the city of Colón to the north; the road east to Chepo, which is to be extended to the Colombian border; and the National Highway, a paved road that runs from the capital to the Costa Rican boundary through the most inhabited and prosperous regions of the west. A branch goes south to Chitré, Las Tablas, and Pedasí and another one, north from David to Boquete.

Railroads. A transcontinental railroad, the Panama–Colón line, was inaugurated in 1855 and is 49 miles long. It is administered by the Panama Railroad, a North American company. The Chiriquí National Railroad runs between La Concepción and Puerto Armuelles on the Pacific, transporting agricultural products. In Bocas del Toro there is a railroad serving the banana-growing area.

Ports. The Panamanian coastline has many natural harbours with excellent conditions for sheltering vessels

but inadequate cargo facilities. The best ports are those at either end of the Panama Canal—Cristobal, near the city of Colón, and Balboa, near Panama City. Coastal shipping is important for Panamanian agriculturists and is the only means of transportation in certain regions.

Airports. There were about 140 domestic airports in the early 1970s and two international airlines. Tocumen, Panama's international airport, 16 miles from the capital, is served by several international airlines and is used by more than 400,000 passengers a year; by agreement it also serves as the commercial airport for the Canal Zone. There is also the Enrique Malek Airport in David, Chiriquí. The national airlines fly regularly from Paitilla (Marcos Gelabert) Airport in Panama City to San Blas, Darién, La Palma, Chitré, Bocas del Toro, Santiago, Puerto Armuelles, and Chiriquí.

Communications. The national government, with a few exceptions, operates and owns the communication system, maintaining a system of telephone, telegraph, and radio communication in the most populated regions. Two private companies—Tropical Radio and Intercomsa—operate an international communication system.

ADMINISTRATION AND SOCIAL CONDITIONS

The structure of government. *The constitutional framework.* After the coup d'etat by the National Guard in 1968, the legislative branch of government was suspended, and Panama was administered by a provisional government. Before that time, the 1946 constitution, the third in the country's history, was operative. It provided for an executive composed of the president and two vice-presidents to be elected for a four-year period by popular and direct vote. The Cabinet worked with the president and advised him. Legislative power under the 1946 constitution was in the hands of a National Assembly of 41 assemblymen, one for every 15,000 habitants, elected for a four-year period. A return to constitutional government was begun in August 1972 with the election of a 505-member assembly.

Justice. Judicial power is in the hands of a Supreme Court of nine members. They are appointed by the president with the approval of his Cabinet and of the Assembly. There are also two superior courts, municipal judges, a public attorney's office (under the control of the attorney general), an auxiliary district attorney, district attorneys, and other officials.

Local government. The country is divided into nine provinces and a territory (comarca), San Blas. The provinces are divided into municipal districts, which are subdivided into counties. The head of each province is the governor, appointed by the president.

The armed forces. There is no army or navy, but there are two coastguard patrol vessels. The National Guard, with a strength of about 5,000 men, is reponsible for national defense and security.

Services and social conditions. *Educational services.* Education is separated into four categories—pre-elementary, elementary, secondary, and university education. All children have the right to receive education from the state without discrimination as to race, sex, and economic or social standing. No private school system can teach in a foreign language without the consent of the Ministry of Education. Elementary education is compulsory and free from seven to 15 years. Illiteracy had greatly decreased by the early 1970s.

Universities

The institutions of higher education are the state-run University of Panama and the privately operated University of Santa Maria la Antigua in Panama City. The University of Panama, the official university of the republic, is an autonomous institution with enrollment of about 10,000 students.

Health and welfare. The state has built hospital and health centres and granted funds for their improvement. There are about two dozen hospitals and half a dozen hospital-clinics, 30 health centres, and a few mobile medical units. There are about 850 doctors, 150 dentists, and 1,000 nurses.

Social security provides a wide range of benefits; all employees must be registered. The social security administration maintains a hospital in Panama City for its beneficiaries and also supports medical centres and clinics in several places in the interior. The Saint Thomas Hospital, in Panama City, provides treatment for those unable to pay hospital expenses.

Housing. In the early 1970s there were about 290,000 dwelling units in the republic: one-family houses are common in rural areas, but rented multifamily dwellings are usual in urban areas. Condominium ownership was becoming popular in the 1970s.

Police services. Police services are operated by the National Guard, which works jointly with the National Department of Investigation (DENI) in dealing with thefts, homicides, assaults, and personal injury cases.

Wages and cost of living. About 90 percent of the working population is protected by a minimum wage law; the Price Control Office regulates prices. Salaries and wages are subject to great fluctuations. In the Canal Zone non-North Americans, numbering about 16,000, receive average monthly wages that are less than half those received by the 6,000 North American workers employed there. Most workers and professionals are union members. Collective labour contracts are usual. The Labour Code requires a monthly vacation for every 11 months of work.

Cultural life. The various regions are visited by musical or theatrical groups, poets, sculptors, and other artists.

Cultural institutions include the Panamanian Art Institute (Panarte), the Concert Association of Panama, the National Conservatory, the School of Plastic Art, the National School of Dance, and the University of Panama, which has its own ballet-concert musical and theatrical groups.

The Panamanian Tourist Bureau (IPAT) encourages the preservation of traditional holidays, folk music, and folk dances. Panamanian carnivals are famous for their beautiful costumes and joyous music. The Institute of Culture and Sports (INCUDE) promotes junior olympiads and other interprovincial events.

The republic has about 100 radio stations—about 60 in Panama, 13 in Chiriquí, and nine in Colón. There are two television stations; in addition the Canal Zone has its own television and radio station.

The main newspapers are *El Panamá América,* the *Panama-American, The Matutino, Crítica, La Hora, La Estrella de Panamá,* and *The Star and Herald.* In Colón the *Atlántico* and in David the *Ecos del Valle* are the principal papers.

Prospects for the future. Relations with the United States, which operates the Panama Canal, are governed by the provisions of the treaties concluded in 1903 and 1936. In the early 1970s, the outcome of talks between the Panamanian and United States governments on the future of the canal was awaited with interest. Like many other Latin American nations, Panama in the 1970s was passing through a period of economic, social, cultural, and political change. Efforts continued to be made to further expand tourism, which represented a vital element in the country's economy. Greater economic cooperation between Panama and Nicaragua was also anticipated.

BIBLIOGRAPHY. LAWRENCE O. EALY, *The Republic of Panama in World Affairs, 1903–1950* (1951), on the geographic position of Panama in world strategy and international politics, and *Yanqui Politics and the Isthmian Canal* (1971); PANAMA CANAL COMPANY, *Climatological Data: Canal Zone and Panama* (1940–65); W.P. WOODRING, "Geology and Paleontology of Canal Zone and Adjoining Parts of Panama," *Prof. Pap. U.S. Geol. Surv.* 306-A (1957), the geology of the Canal Zone; CANAL ZONE GOVT. and PANAMA CANAL COMPANY, *A Portrait of the Funnel for World Commerce: The Panama Canal* (1955), operation, organization, and services of the Panama Canal; CHARLES REGINALD ENOCK, *The Panama Canal: Its Past, Present, and Future* (1914); JOSEPH BUCKLIN BISHOP, *The Panama Gateway* (1913); SHELDON B. LISS, *The Canal: Aspects of United States Panamanian Relations* (1967); DONALD BARR CHIDSEY, *The Panama Canal: An Informal History* (1970); MARION J. SIMON, *The Panama Affair* (1971).

(N.M.C.)

Panama Canal

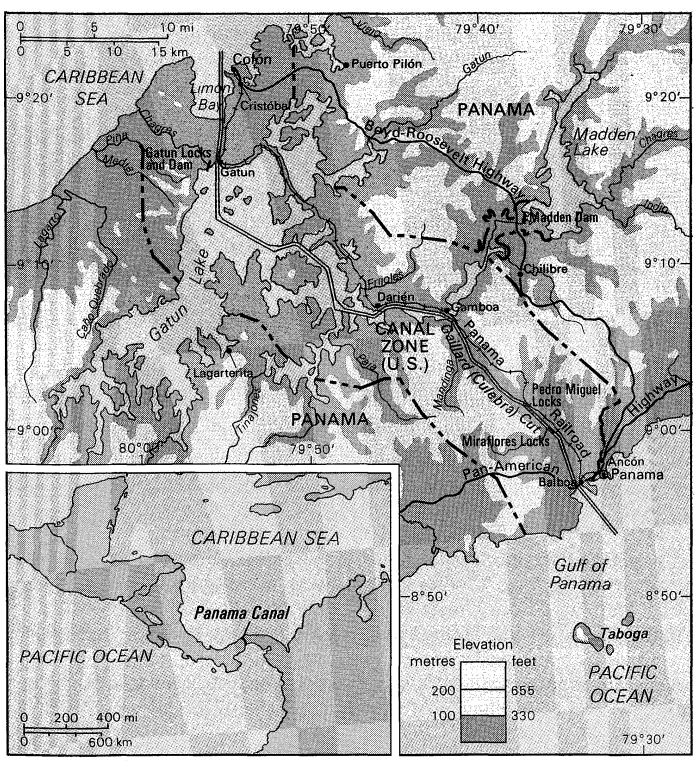
The Panama Canal is an interoceanic waterway connecting the Atlantic and Pacific oceans through the Isthmus of Panama. It is about 51 miles (82 kilometres) in length from deep water to deep water and is one of the two most strategic artificial waterways in the world, the other being the Suez Canal (*q.v.*). It is owned and operated by the United States. For a detailed discussion of canals, see the article CANALS AND INLAND WATERWAYS; for a discussion of the Panama Canal Zone, see the article PANAMA; for a consideration of historical aspects, see CENTRAL AMERICAN STATES, HISTORY OF.

The canal consists of short sea-level sections at each end, three pairs of locks that lift ships to 85 feet (26 metres) above sea level, a 32-mile elevated section that includes Gatun Lake, and a narrow eight-mile-long excavated channel, known as Gaillard Cut, running through the continental divide.

The dimensions of its lock chambers—1,000 feet in length, 110 feet in width, and 41 feet in depth—permit most commercial ships, as well as principal Navy ships, to pass through the canal. Some passenger liners, the largest sizes of cargo carriers and tankers, and the largest aircraft carriers cannot, however, be accommodated. Proposals for enlarging the locks or, alternatively, for building a sea-level waterway have been before the government since 1939.

Ships sailing between the east and west coasts of the United States, which would otherwise be obliged to round Cape Horn, shorten their voyage by about 8,000 nautical miles by using the canal. Savings of up to 3,500 miles are also made on voyages between one coast of North America and ports on the other side of South America. Ships sailing between Europe and East Asia, or Australia, save as much as 2,000 miles by using the canal. In terms of United States interests, the canal, which enables naval units to pass readily from one ocean to the other, is a link in the defense of North America.

The waterway. The Panama Canal lies nine degrees north of the Equator. It extends from north to south, running from the Atlantic terminus at Cristobal to Balboa on the Pacific Ocean. The waterway is set in the middle of a canal zone, which extends five miles on either side of the canal. The zone was granted to the United States in perpetuity by the Republic of Panama by the Hay-Bunau-Varilla Treaty of November 18, 1903, which also provided for the construction and operation of the canal. Waters from the Río Chagres and Gatun and Miraflores lakes are used to refill all the locks. The heavy rainfall of the tropics makes operation feasible despite the use of 52,000,000 gallons (196,560,000 litres) of water for the transit of each ship.



The Panama Canal.

The eight principal features of the canal are (1) a seven-mile dredged channel from the Atlantic terminus in Limon Bay to the Gatun Locks; (2) the Gatun Locks, which raise or lower ships 85 feet to or from Gatun Lake; (3) a 24-mile channel through Gatun Lake to Gamboa; (4) Gaillard Cut, an eight-mile, 500-foot-wide channel winding through the continental divide: (5) the Pedro Miguel Locks, constituting a 31-foot step down from Gaillard Cut; (6) a 750-foot-wide channel running for about a mile across Miraflores Lake; (7) the Miraflores Locks, constituting a 54-foot two-step drop down from the lake level; and (8) a sea-level dredged channel leading over eight miles to the Pacific terminus in Bay of Panama.

The canal locks operate by gravity flow of water from Gatun and Madden lakes. The locks themselves are of uniform length, width, and depth. Each set of locks is built in tandem, to permit simultaneous transit of vessels in either direction. To conserve water, two or more vessels are passed through together, in the same direction, when size permits. Each lock gate has two leaves, 65 feet wide by seven feet thick, set on hinges. The gates range in height from 47 to 82 feet; their movement is powered by motors recessed in the lock walls. They are operated from a control tower, located on the wall that separates each pair of locks, and from which the flooding or emptying of the lock chambers is also controlled. Due to the delicate nature of the lock mechanisms, only small craft are permitted to pass through the locks under their own power. Larger craft are towed through by locomotives. Ships are taken through the canal by a pilot, who boards each ship before it leaves the terminus. In time of war or emergency, armed guards also accompany each ship through the canal.

The canal locks

Navigation. Including waiting time, ships require about 15 hours to negotiate the canal. The average transit time through the canal, once a vessel has been authorized to proceed, however, is seven to eight hours from deep water to deep water. When Gaillard Cut is not being dredged, canal traffic is generally allowed to proceed in both directions.

Marine traffic is handled by a complex manual system that, with the increase in traffic in recent years, has become overburdened. Each ship is boarded by measurers to determine its carrying capacity for toll-collection purposes. Manifests, ship's papers, and other documents are inspected and recorded. Transits are scheduled and are monitored at points along the route. In 1970 an average of 41 oceangoing ships, with an average size of 7,800 tons each, passed through the canal each day. It has been calculated that by 1990 traffic will increase to 70 ships a day.

Tolls. The charter of the canal provides for the Panama Canal Company (an agency of the United States) to establish tolls for the use of the waterway at rates calculated to cover costs of maintenance and operation. Tolls are based upon the measured earning capacity of a ship. The rates have remained virtually unchanged since the canal was opened. The average charge for a ship in transit in 1970 was $6,850. Vessels owned, operated, or chartered by the government of Panama are exempted by treaty from payment of tolls. Changes in the rules for both the measurement and the collection of tolls may be made only after six months' notice, the holding of public hearings, and approval by the president of the United States.

Toll rates

Maintenance. Continual maintenance work on the canal and its associated facilities is needed to keep it in operation in a tropical climate. This includes the dredging of channels, the overhauling of locks, and the repairing of machinery. Especially in the early years of the canal's operation, landslides occurred in the hills adjoining Gail-

lard Cut. A close watch must be kept on soil conditions, since heavy rainfall and the removal of land still make earth slides probable. Preventive measures have frequently had to be taken to keep the channel open—most recently in 1968, when large cracks opened in the hills to the west of the cut.

History. As early as the 16th century the Spanish conceived the idea of constructing a canal across the isthmus. In 1846 the United States concluded a treaty with New Granada (now Colombia), which then exercised sovereignty over Panama, aimed primarily toward securing a canal route; the treaty provided that the United States should guarantee the "perfect neutrality" of the isthmus and its free and uninterrupted transit. In 1855 United States financial interests joined to build the Panama Railroad across the isthmus. In 1880 the French Panama Canal Company, under the leadership of Ferdinand de Lesseps, builder of the Suez Canal, undertook the first attempt to construct a canal. Ten years later, due to lack of planning, graft, and the ravages of yellow fever, malaria, and cholera, which decimated the work force, the company became bankrupt. It was reorganized in 1894 as the New Panama Canal Company but made little progress, due to lack of money. The U.S. Congress, by the Spooner Act of 1902, authorized and directed the President to acquire rights and to build an interoceanic canal. Following Panama's declaration of independence from Colombia in 1903, and the conclusion of the Hay-Bunau-Varilla Treaty between the United States and Panama in that year, canal-building rights passed into U.S. hands. Work began on the canal in 1904 but not until 1906 was it decided that, instead of being at sea level, the waterway would be of the high-level type, using locks. The canal was first opened to traffic on August 15, 1914.

Capital improvements. The first major capital improvement was the construction of Madden Dam and Power Project, which was completed in 1935. This not only stemmed and controlled the flow of water moving into Gatun Lake at the rate of 200,000,000,000 cubic feet (6,000,000,000 cubic metres) a year but also created a large reservoir in Madden Lake, as well as increasing the amount of electric power available.

In 1939 the United States Congress authorized the addition of a third set of locks removed from the existing facilities. Although excavation was begun the following year, work was suspended during World War II and has not been resumed. A highway built across the isthmus—much of it running outside the Canal Zone—has added a supplementary facility to the waterway and railroad. Following an agreement between the United States and Panama in 1955, the Thatcher Ferry Bridge was built, connecting Panama City and Balboa with the north side of the canal, and adding a fresh link to the Pan-American Highway system that connects the capitals and principal cities of South and Central America with the highways of the United States and Canada.

A major improvement to the canal itself, completed in 1970, was the widening of Gaillard Cut from 300 to 500 feet throughout, thus enabling ships to move through at faster speeds.

Canal traffic. Traffic through the Panama Canal is a barometer of world trade, rising in years of prosperity and declining in times of recession. From a low of 807 transits in 1916, traffic rose to a high point of 15,523 transits of all types in 1970. The cargo carried through the canal that year amounted to over 132,500,000 tons. The tolls collected for the fiscal year ending in 1970 totalled over $100,000,000.

Principal trade routes

The principal trade routes served by the Panama Canal, apart from the United States intercoastal trade route, run between the following points: Europe, and the west coast of North America; the United States east coast, and Hawaii and East Asian ports; the United States east coast, and the west coast of South America; the east and west coasts of South America; the West Indies and Asia; Europe and Asia; Europe, and the west coast of South America; Europe, and Australia.

Cargo movements in recent years have reversed the earlier pattern, when Pacific to Atlantic traffic was heavier. From 1968 to 1969, for example, traffic originating in Atlantic ports has been nearly double that from the Pacific. This is due in part to the military buildup in Southeast Asia; to heavy shipments of oil from the Caribbean and Gulf of Mexico ports to the Pacific; to shipments of grain to Asian ports; to the increased transport of ores and metals to Japan; to nitrate and phosphate shipments to the west coast of the United States; and to coal shipments to Japan. Coal and coke are the principal items shipped from the Atlantic to the Pacific, followed by petroleum and petroleum products. Grains, ores, metals, and nitrates rank next in tonnage of cargo.

Panama Canal, Commercial Ocean Traffic, 1971

country of registry	vessel transits (number)	total transits (percent)	cargo (long tons)	total cargo (percent)
Liberia	1,587	11.3	25,201,391	21.2
United Kingdom	1,558	11.1	14,288,579	12.0
Japan	1,462	10.4	13,541,685	11.4
United States	1,368	9.8	8,246,308	7.0
Norway	1,202	8.6	16,011,868	13.5
West Germany	1,069	7.6	4,918,407	4.1
Panama	948	6.8	3,699,065	3.1
Greece	629	4.5	7,735,546	6.5
The Netherlands	494	3.5	2,648,769	2.2
Sweden	479	3.4	3,366,568	2.8

Ships flying the United States flag previously held first place in number of transits among shipping from various countries using the canal route. They currently occupy third place, comprising 11.1 percent of the total. Ships flying the Liberian flag now hold first place, with 11.7 percent. British shipping is second, accounting for 11.6 percent. Norwegian shipping holds fourth place, followed by Japan and others. The large percentage of Liberian shipping results from the practice of ship owners of registering ships under "flags of convenience," to save costs, a practice which is particularly employed for registering large carriers of oil, ores, and grain. Seventy percent of the canal traffic is, however, destined for, from, or between parts of the United States, underscoring the importance of the canal to the United States economy.

Recent traffic statistics also reflect the rapid growth of the Japanese economy; over 45 percent of the cargo passing through the canal came from, or was destined for, Japan in the fiscal year ending in 1970. The canal cannot accommodate the ultralarge tankers and bulk carriers of 100,000 to 300,000 tons deadweight now being built until either larger locks or a sea-level waterway is built; with the greater numbers of large carriers operating on world trade routes and cargo passing through the canal increasing at a rate of 7 percent a year, larger facilities will be needed by the year 2000 if the canal is to remain economically competitive.

The Panama Canal Company. The Panama Canal Company, established by the U.S. Congress, is the corporate agency created to maintain and operate the canal, as well as to conduct business operations incident thereto, as well as those incident to the civil government of the Canal Zone. The company has a board of directors appointed by the U.S. secretary of the army, who is the sole stockholder, under authority delegated to him by the president of the United States.

The basic law requires that the company be self-sustaining. Its obligations include its own operating expenses; the net cost of civil government, interest, and depreciation on United States investment in the enterprise; and $430,000 of the $1,930,000 annuity paid to the Republic of Panama. (The remainder is provided by the U.S. Department of State and is thereby excluded in fixing tolls.)

The company also maintains ship facilities both at Cristobal at the Atlantic terminal, and at Balboa at the Pacific terminal of the canal. These accommodations include docks, warehouses, repair shops, bunkering facilities, and rail and road connections with Panama City and the city of Colón in the Republic of Panama.

The Canal Zone Government is responsible for civil

government in the Canal Zone and is administered by a governor who is directly responsible to the U.S. secretary of the army and to the U.S. president. He is appointed for a four-year term and is ex officio the president of the Panama Canal Company.

Over 15,000 persons were employed by the company and the government in 1970, of whom over 4,000 were U.S. citizens. In recent years an increasing number of Panamanian employees have been paid at U.S. wage rates.

The Panama Railroad

The Panama Railroad runs for 47 miles between Colón and Panama City. The original line, built to transport prospectors bound for the California goldfields across the isthmus, was completed in 1855. Its existence gave Panama a great advantage over Nicaragua when the choice of a canal site was made. The railway was used in the construction of the canal but had to be moved to higher ground on the east side of the canal before the waters began to rise. The railway carries goods and passengers across the isthmus, running close to the canal and crossing Gatun Lake by a man-made causeway.

The net value of the property, plant, and equipment at the Canal Zone was placed at $5,000,000,000 in 1969. The net investment of the United States in the canal enterprise is over $5,000,000,000. Capital expenditures average close to $17,000,000 a year. Direct operating expenses amount to $48,000,000; general and administrative expenses, $17,000,000. Tolls and toll credits produced an income of $100,900,000 in 1970.

International status of canal. Under the terms of the Hay-Bunau-Varilla Treaty of November 18, 1903, the United States received from the Republic of Panama the right to the use, occupation, and control of the Canal Zone "in perpetuity," for the "construction, maintenance, operation, sanitation, and protection" of the interoceanic canal. The treaty further granted to the U.S. "all the rights, power and authority" within the zone, which the U.S. would have "if it were the sovereign of the territory." Although Panama did not yield its ultimate sovereignty over the zone, it gave the U.S. the right to exercise sovereignty there "to the entire exclusion of the exercise by the Republic of Panama." Debates have been conducted how far this extends. It is apparently sufficient to give the U.S. a secure title to the canal, the power to defend it, and the right to decide how the canal and the Canal Zone shall be operated. For this grant the U.S. paid Panama $10,000,000 in cash in 1904, plus an annuity that was originally fixed at $250,000. Since 1955, the annuity has been $1,930,000.

A second component in the international status of the canal is the Hay-Pauncefote Treaty of November 18, 1901, by the terms of which the United Kingdom gave up its interest in an isthmian canal in return for agreement by the U.S. that a canal built by the U.S. would be "free and open" to the vessels of commerce and war of all nations, which observed the rules established for the Suez Canal by the Convention of Constantinople, 1888, by which the Suez Canal was made a corridor for all ships of all nations in peace and in war. This was subsequently interpreted as in no way forbidding the U.S. to take measures for ensuring the safety or defense of the Panama waterway. The Hay-Pauncefote Treaty also provided that there shall be "entire equality" in the treatment of ships of all nations with respect to "conditions and charges of traffic."

The fundamental act authorizing the U.S. government to acquire the rights for a canal route and to build a canal of "sufficient capacity and depth as shall afford convenient passage for vessels of the largest tonnage and greatest draft" then in use, "or such as may be reasonably anticipated," was the act of the U.S. Congress of June 28, 1902, known as the Spooner Act.

Need for a larger waterway

The future. The increasing size of bulk carriers, and of container ships, and the inability of the present canal to accommodate the largest aircraft carriers of the U.S. Navy have led to the consideration of the construction of a larger sized waterway. It remains to be decided whether the larger waterway should be constructed on the site of the existing canal or elsewhere, and whether it should be a lock or sea-level type of canal.

Among 29 conceivable routes across Central America, six are considered the most feasible. The first of these is a Nicaraguan route from San Juan del Norte (Greytown) on the Atlantic up the Río Indio to Lake Nicaragua (Spanish Lago de Nicaragua) and down to the Bay of Salinas (Spanish Bahía de Salinas). Although this is a shorter route, from the point of view of U.S. intercoastal trade, it would require cuts through high mountains and would pose difficult engineering problems. A second possibility is a sea-level route ten miles northwest of the present Canal Zone starting from the mouth of the Río Indio, crossing the continental divide near the town of La Chorrera, and exiting into the Bay of Panama near Taboga Island. Such a route would be about as long as the present canal and could be built up without disrupting traffic. A new United States–Panamanian treaty would be required for this. A third possibility is a sea-level canal within the present Canal Zone, replacing the existing canal either by using the existing route or else by employing a new cut through the divide; construction would seriously disrupt traffic. A fourth possibility is to make a deep-draft lock canal replacing the existing structures. The fifth possibility is a Sasardí-Mortí route in Panama, 100 miles east of the present canal, and the sixth is the so-called Atrato route in northern Colombia. Nuclear engineering is thought to be the only practicable means of construction for both these routes, and this is deemed to be politically unacceptable.

A United States Atlantic–Pacific Interoceanic Canal Study Commission in 1970 recommended building a sea-level canal through Panama just north of the present Canal Zone, retaining the existing canal as a supplementary facility. It believed this route to be the most feasible and to have the advantage of proximity to existing facilities, engineering skills, and a large labour force.

Ecologists have questioned whether a sea-level canal might not disturb the existing balance of nature by bringing organisms from the tropical Pacific into the Atlantic, but the commission did not think this a serious problem. Negotiations have been entered into with the Panamanian government to secure rights for building and operating a new canal. The alternative remains of constructing a sea-level canal in the present zone or of converting the existing canal into a deep-draft lock canal.

BIBLIOGRAPHY

Reports: U.S. CONGRESS, *Report on a Long-Range Program for Isthmian Canal Transits,* 86th Congress, 2nd sess., House Report No. 1960 (1960); "Report on Proposals for Elimination of Pedro Miguel Locks, Panama Canal, Jan. 17, 1944," *Congressional Record,* vol. 102, pt. 8, pp. 10757–10766 (June 21, 1956); report of the ATLANTIC-PACIFIC INTEROCEANIC CANAL STUDY COMMISSION (1970); *Annual Reports* of the PANAMA CANAL COMPANY–CANAL ZONE GOVERNMENT.

Official statements: HON. DANIEL J. FLOOD, *Isthmian Canal Policy Questions,* 89th Congress, 2nd sess., House Document No. 474 (1966); HON. CLARK W. THOMPSON, "Isthmian Canal Policy of the United States-Documentation, 1955–64," *Congressional Record,* vol. 110, pt. 16, pp. 21467–21475.

Books by the canal builders: G.W. GOETHALS *et al., The Panama Canal: An Engineering Treatise,* 2 vol. (1916); W.C. GORGAS, *Sanitation in Panama* (1915); W.L. SIBERT and J.F. STEVENS, *The Construction of the Panama Canal* (1915).

General information: R.R. BAXTER, *The Law of International Waterways with Particular Regards to Interoceanic Canals* (1964), a legal treatise; M.P. DUVAL., JR., *Cadiz to Cathay: The Story of the Long Diplomatic Struggle for the Panama Canal,* 3rd ed. (1968), written by a former marine superintendent of the canal; I.J. KLETTE, *From Atlantic to Pacific: A New Interoceanic Canal* (1967), a brief consideration of issues relating to construction of a new waterway; N.J. PADELFORD, *The Panama Canal in Peace and War* (1942).

(N.J.P.)

Pan Ku

A scholar-official of the Later, or Eastern, Han dynasty (AD 23–220), Pan Ku (in Pin-yin romanization, Ban Gu) was the second of China's three most famous historians, with Ssu-ma Ch'ien and Ssu-ma Kuang. Of the three, Pan Ku is the one whose organization has been most often followed by later historians.

His father's work

He was probably born in AD 32 at Ch'ang-an (modern Sian), the Former, or Western, Han capital, shortly before the interregnum that divided the dynasty into two different periods. His father, Pan Piao (AD 3–54), an intellectual and antiquarian, was given a court appointment by the Emperor during the early years of the restoration of the Han dynasty. Disliking court life, Pan Piao pleaded poor health and retired, thereafter devoting himself to the independent study of history. He collected material for the continuation of Ssu-ma Ch'ien's great history of China, the *Shih-chi*, which had begun with the earliest dynasties and stopped midway through the Former Han dynasty.

After his father's death, Pan Ku continued this historiographical undertaking. In the course of it, however, he was imprisoned for tampering with dynastic records. His twin brother, Pan Ch'ao, an outstanding general who extended China's western frontier to the Pamirs, interceded so successfully that Pan Ku was not only acquitted but was also appointed by the Emperor to the office of official historian.

"Han shu." With all obstacles now removed, he spent the next 16 years compiling and editing the vast *Han shu* (known in English as *The History of the Former Han Dynasty*), which became the prototype for the official histories of successive ruling houses in China, recording the administrations of their predecessors. Although modelled on the *Shih-chi*, the *Han shu* was not merely a supplement to that long-range work but was a new and comprehensive record of the Han Empire, from its beginning down to the regime of the reformer Wang Mang, who had proclaimed his own short-lived dynasty in AD 9. Pan Ku went back to the beginning of the Han, duplicating almost verbatim most of the documents Ssu-ma Ch'ien had used for that part of the Han period he had treated, excising redundancies or simplifying the prose that seemed to him awkward or obscure. And, since Pan Ku's own age enjoyed widening education, bureaucratic proliferation, improved writing materials and techniques, and standardized orthography, he had an even larger body of recent records from which to select. Dealing with a period of roughly 200 years, the *Han shu* is much longer than the *Shih-chi*, which purports to cover 3,000 years.

Plan of the *Han shu*

Both Ssu-ma Ch'ien and Pan Ku were court officials, and they inevitably used the official record of the lives of the emperors and their close relatives (and of the often more decisive activities of their civil and military administrators) to form their main chronological narrative. This constitutes Pan's part 1, the basic annals. He adopted Ssu-ma Ch'ien's methods for other parts: part 2, charts and diagrams of events, genealogies, persons, etc.; part 3, treatises on a wide range of topics, such as court ceremony, music, money and taxes, and navigation; and part 4, single or grouped biographies of memorable persons other than emperors. To these subjects he added new ones on natural phenomena, on geography, and on bibliography, a descriptive account of the books preserved in the Imperial library—invaluable to later scholars trying to judge textual authenticity and family lineages after many works had vanished. Pan Ku eliminated Ssu-ma Chi'en's fifth category of "hereditary houses," since China was no longer a collection of competing states.

Role in contemporary affairs. When the historian felt his task to be essentially completed, he apparently decided to participate more actively in the politics of his day. He had been at least on the periphery of intellectual controversy regarding the interpretation of the Confucian Classics—by no means a mere antiquarian pursuit but one fraught with political implications, as he remarked in one of his few personal observations in the *History*. Editorship of the *Pai-hu t'ung*, or "Symposium in the White Tiger Hall," which deals with this subject, is ascribed to him. In his middle 40s, however, he chose to undertake something more adventurous. Leaving the finishing touches on the *Han shu* to his sister Pan Chao, also an extraordinary scholar (not to be confused with his brother Pan Ch'ao), he joined the staff of the general Tou Hsien and accompanied him in successful campaigns against the northern Hsiung-nu tribes. The following victory inscription composed by Pan Ku was carved in stone 1,000 miles beyond the frontier:

> Our trained soldiery came hither on a campaign against barbarian hordes. We chastised Turkic insolence and restored our supremacy in this distant land. Across these vast plains they went back to their northern home, while our splendid troops set up this trophy that the achievements of our glorious Emperor should be heard of ten thousand generations hence.

The Emperor, however, who was 14 years old and Tou Hsien's nephew, became alarmed at the general's self-importance and, suspecting him of excessive ambitions, exiled him to his own lands. Pan Ku's fate was one common throughout Chinese history; his superior's fall implicated him, and he was incarcerated for interrogation. Falling ill in prison, he died there at the age of 60 (AD 92). His sister duly rounded out the vast *Han shu* manuscript and was officially sanctioned to instruct other scholars on its contents.

Death

Assessment. For centuries the Chinese have debated the relative merits of the history of the single, self-contained dynasty such as Pan Ku's, and the comparatively rare histories that span the rise and fall of successive hegemonies and systems, which are claimed to reflect more effectively the lessons of history. Obviously, the general historian must build on the work of those dealing with shorter periods, and the two kinds of enterprise cannot be compared qualitatively on the basis of scope. As a historian, Pan Ku must be evaluated on other grounds in relation to his predecessor and to the next great long-range narrator, Ssu-ma Kuang, who wrote more than 1,000 years later. Since both were more inclined to offer interpretations and personal comments, their commentaries appear to be more colourful and sometimes more interesting. Pan Ku, on the other hand, is admired for his thoroughness and his virtually complete objectivity.

Indeed, one might call Pan Ku a historiographer rather than a historian. He undertook simply to represent the Han dynasty and empire as factually as possible through an organized compendium of existing documents; hence the title *Han shu*—literally, "Han Documents."

Prose style

Pan Ku's prose style, to which he more or less adjusted the documents he incorporated, was simple, lucid, unecentric, and not especially vivid. It was terse but not lapidary; somewhat more carefully regulated than Ssu-ma Ch'ien's, it was still, probably, not altogether remote from the spoken Chinese of his own day and class. It was a model of what came to be known as the Han style, revived many centuries later in reaction to excessively elaborate prose.

When practicing the dominant literary form of his time, the *fu*, or rhymed prose, however, Pan Ku could be as extravagant, bizarre, and exhibitionistic as others displaying their talents in this fashionable genre. His two rhymed prose compositions on the merits of the successive Han capitals (the new one, of his present masters, of course, winning out) spawned many imitations, especially for their exhibition of unusual words. In a simpler vein he wrote some rather inconsequential verses modelled on the popular folk songs of his day. His name has been attached, spuriously, to a collection of anecdotes and hearsay about the reign of the Han emperor Wu Ti.

BIBLIOGRAPHY. Further information may be found in H.H. DUBS, *The History of the Former Han Dynasty*, 3 vol. (1938–55). Both CH'EN SHOU-YI, *Chinese Literature: A Historical Introduction* (1961); and BURTON WATSON, *Early Chinese Literature* (1962), discuss Pan Ku as historian and writer.

(G.W.Ba.)

Pantheism and Panentheism

Both "pantheism" and "panentheism" are terms of recent origin, coined to describe certain views of the relationship between God and the world that are different from that of traditional Theism (*q.v.*). As reflected in the prefix "pan-" (Greek *pas*, "all"), both of the terms stress the all-embracing inclusiveness of God, as compared with his separateness as emphasized in many versions of Theism. On

the other hand, pantheism and panentheism, since they stress the theme of immanence—*i.e.*, of the indwelling presence of God—are themselves versions of Theism conceived in its broadest meaning. Pantheism stresses the identity between God and the world; panentheism (Greek *en*, "in") holds that the world is included in God but that God is more than the world.

The adjective "pantheist" was introduced by the Irish Deist, John Toland, in a book, *Socinianism Truly Stated* (1705). The noun "pantheism" was first used in 1709 by one of Toland's opponents. The term "panentheism" appeared much later, in 1828, when it was used to characterize the view that the world is a finite creation within the infinite being of God.

Although the terms are recent, they have been applied retrospectively to alternative views of the divine being as found in the entire philosophical traditions of both East and West.

THE NATURE AND SIGNIFICANCE OF PANTHEISM AND PANENTHEISM

Pantheism and panentheism can be explored by means of a three-way comparison with traditional or Classical Theism viewed from eight different standpoints—*i.e.*, from those of immanence or transcendence; of monism, dualism, or pluralism; of time or eternity; of the world as sentient or insentient; of God as absolute or relative; of the world as real or illusory; of freedom or determinism; and of sacramentalism or secularism.

Immanence or transcendence. The poetic sense of the divine within and around mankind, which is widely expressed in religious life, is frequently treated in literature. It is present in the Platonic Romanticism of Wordsworth and Coleridge, as well as in Tennyson, Emerson, and Goethe. Expressions of the divine as intimate rather than as alien, as indwelling and near dwelling rather than remote, characterize pantheism and panentheism as contrasted with Classical Theism. Such immanence encourages man's sense of individual participation in the divine life without the necessity of mediation by any institution. On the other hand, it may also encourage a formless "enthusiasm," without the moderating influence of institutional forms. In addition, some theorists have seen an unseemliness about a point of view that allows the divine to be easily confronted and appropriated. Classical Theism has, in consequence, held to the transcendence of God, his existence over and beyond the universe. Recognizing, however, that if the separation between God and the world becomes too extreme, man risks the loss of communication with the divine, panentheism—unlike pantheism, which holds to the divine immanence—maintains that the divine can be both transcendent and immanent at the same time.

Distinctions viewed from the metaphysical standpoint

Monism, dualism, or pluralism. Philosophies are monistic if they show a strong sense of the unity of the world, dualistic if they stress its twoness, and pluralistic if they stress its manyness. Pantheism is typically monistic, finding in the world's unity a sense of the divine, sometimes related to the mystical intuition of personal union with God; Classical Theism is dualistic in conceiving God as separated from the world and mind from body; and panentheism is typically monistic in holding to the unity of God and the world, dualistic in urging the separateness of God's essence from the world, and pluralistic in taking seriously the multiplicity of the kinds of beings and events making up the world. One form of pantheism, present in the early stages of Greek philosophy, held that the divine is one of the elements in the world whose function is to animate the other elements that constitute the world. This point of view, called Hylozoistic (Greek *hylē*, "matter," and *zōē*, "life") pantheism, is not monistic, as are most other forms of pantheism, but pluralistic.

Time or eternity. Most, but not all, forms of pantheism understand the eternal God to be in intimate juxtaposition with the world, thus minimizing time or making it illusory. Classical Theism holds that eternity is in God and time is in the world but believes that, since God's eternity includes all of time, the temporal process now going on in the world has already been completed in God. Panentheism, on the other hand, espouses a temporal–eternal God who stands in juxtaposition with a temporal world; thus, in panentheism, the temporality of the world is not cancelled out, and time retains its reality.

The world as sentient or insentient. Every philosophy must take a stand somewhere on a spectrum running from a concept of things as unfeeling matter to one of things as psychic or sentient. Materialism holds to the former extreme, and Panpsychism to the latter. Panpsychism offers a vision of reality in which to exist is to be in some measure sentient and to sustain social relations with other entities. Dualism, holding that reality consists of two fundamentally different kinds of entity, stands again between two extremes. A few of the simpler forms of pantheism support Materialism. Panentheism and most forms of pantheism, on the other hand, tend toward Panpsychism. But there are differences of degree, and though Classical Theism tends toward dualism, even there the insentient often has a tinge of Panpsychism.

God as absolute or relative. God is absolute insofar as he is eternal, cause, activity, creator; he is relative insofar as he is temporal, effect, passive (having potentiality in his nature), and affected by the world. For pantheism and Classical Theism, God is absolute; and for many forms of pantheism, the world, since it is identical with God, is likewise absolute. For Classical Theism, since it envisages a separation between God and the world, God is absolute and the world relative. For pantheism, however, God is absolute and relative, cause and effect, actual and potential, active and passive. The panentheist holds that, inasmuch as they refer to different levels of the divine nature, both sets of claims can be attributed to God without inconsistency, that just as a man can have an absolute, unchanging purpose, which gains now one embodiment and now another, so God's absoluteness can be an abstract unchanging feature of a changing totality.

The world as real or illusory. Panentheism, Classical Theism, and many forms of pantheism hold the world to be part of the ultimate reality. But for Classical Theism the world has a lesser degree of reality than God; and for some forms of pantheism, for which Hegel coined the term Acosmism, the world is unreal, an illusion, and God alone is real.

Distinctions viewed from the human standpoint

Freedom or determinism. In those forms of pantheism that envisage the eternal God literally encompassing the world, man is an utterly fated part of a world that is necessarily just as it is, and freedom is thus an illusion. To be sure, Classical Theism holds to the freedom of man but insists that this freedom is compatible with a divine omniscience that includes his knowledge of the total future. Thus the question arises whether or not such freedom is illusory. Panentheism, by insisting that future reality is indeterminate or open and that man and God, together, are in the process of determining what the future shall be, probably supports the doctrine of man's freedom more completely than does any alternative point of view.

Sacramentalism or secularism. Insofar as God is the indwelling principle of the world and of man, as in pantheism, so far do these take on a sacramental character; and insofar as God is separated from the world as in 18th-century Deism (*q.v.*), so far does it become secular, neutral, or even fallen. In contrast, Classical Theism, though basically sacramental, places this quality in an enclave, the church.

DIVERSE VIEWS OF THE RELATION OF GOD TO THE WORLD

On the basis of the preceding characteristics, seven forms of pantheism can be distinguished in addition to Classical Theism and panentheism:

Hylozoistic pantheism. The divine is immanent in, and is typically regarded as the basic element of, the world, providing the motivating force for movement and change. The world remains a plurality of separate elements.

Immanentistic pantheism. God is a part of the world and immanent in it. Though only a part, however, his power extends throughout its totality.

Absolutistic monistic pantheism. God is absolute and identical with the world. The world, although real, is therefore changeless.

Relativistic monistic pantheism. The world is real and changing and is within God (*e.g.*, as the body of God). But God remains nonetheless absolute and is not affected by the world.

Acosmic pantheism. The absolute God makes up the total reality. The world is an appearance and ultimately unreal.

Identity of opposites pantheism. The opposites of ordinary discourse are identified in the supreme instance. God and his relation to the world are described in terms that are formally contradictory; thus reality is not subject to rational description. Whether being is stressed or the void, whether immanence is or transcendence, the result is the same: one must go beyond rational description to an intuitive grasp of the ultimate.

God as related to time and change

Classical Theism. God is absolute, eternal, first cause, pure actuality, an omniscient, omnipotent, and perfect being. Though related to the world as its cause, he is not affected by the world. He is essentially transcendent over the world; and the world exists relative to him as a temporal effect of his action—containing potentiality as well as actuality and characterized by change and finitude. Since all of time is part of God's eternal "Now," and since God's knowledge now includes the total future as though laid out before him like a landscape, it is not clear that, in this system, man can have freedom in any significant sense; for although foreknowledge does not of itself determine anything, it vouches for the existence of such determination. Nonetheless, human freedom is in fact asserted by Classical Theists.

Neoplatonic or emanationistic pantheism. God is absolute in all respects, remote from the world and transcendent over it. This view is like Classical Theism except that, rather than saying that God is the cause of the world, it holds that the world is an emanation of God, occurring by means of intermediaries. God's absoluteness is thus preserved while a bridge to the world is provided as well. In Plotinus (3rd century AD), the foremost Neoplatonist, the Nous (Greek, "mind"), a realm of ideas or Platonic forms, serves as the intermediary between God and the world, and the theme of immanence is sustained by positing a World-Soul that contains and animates the world.

Panentheism. In this alternative, both sets of categories, those of absoluteness and of relativity, of transcendence and of immanence, are held to apply equally to God, who is thus dipolar. He is the cause of the world and its effect; his essence is eternal, but he is involved in time. God's knowledge includes all that there is to know; since the future is genuinely open, however, and is not in any sense real as yet, he knows it only as a set of possibilities or probabilities. In this alternative man is held to have significant freedom, participating as a co-creator with God in the continuing creation of the world.

With only slight attention being accorded to Classical Theism (which is covered in another article), the incidence of the preceding eight forms of pantheism and panentheism in cultural history remains to be explored.

PANTHEISM AND PANENTHEISM IN NON-WESTERN CULTURES

The Vedas and *Upaniṣads*

Hindu doctrines. The gods of the Vedas, the ancient scriptures of India (*c.* 1200 BC), represented for the most part natural forces. Exceptions were the gods Prajāpati (Lord of Creatures) and Puruṣa (Supreme Being or Soul of the Universe), whose competition for influence provided, in its outcome, a possible explanation of how the Indian tradition came to be one of pantheism rather than of Classical Theism. By the tenth book of the Ṛgveda, Prajāpati had become a lordly, monotheistic figure, a creator deity transcending the world; and in the later period of the sacred writings of the *Brāhmaṇas* (*c.* 7th century BC), prose commentaries on the Vedas, he was moving into a central position. The rising influence of this Theism was later eclipsed by Puruṣa, who was also represented in Ṛgveda X. In a creation myth Puruṣa was sacrificed by the gods in order to supply (from his body) the pieces from which all the things of the world arise. From this standpoint the ground of all things lies in a Cosmic Self, and all of life participates in that of Puruṣa. The Vedic hymn to Puruṣa may be regarded as the starting point of Indian pantheism.

In the *Upaniṣads* (*c.* 600–300 BC), the most important of the ancient scriptures of India, the later writings contain philosophic speculations concerning the relation between the individual and the divine. In the earlier *Upaniṣads*, the absolute, impersonal, eternal properties of the divine had been stressed; in the later *Upaniṣads*, on the other hand, and in the *Bhagavadgītā*, the personal, loving, immanentistic properties became dominant. In both cases the divine was held to be identical with the inner self of each man. At times these opposites were implicitly held to be in fact identical—the view earlier called identity of opposites pantheism. At other times the two sets of qualities were related, one to the unmanifest absolute Brahman, or supreme reality (sustaining the universe), and the other to the manifest Brahman bearing qualities (and containing the universe). Thus Brahman can be regarded as exclusive of the world and inclusive, unchanging and yet the origin of all change. Sometimes the manifest Brahman was regarded as an emanation from the unmanifest Brahman; and then emanationistic pantheism—the Neoplatonic pantheism of the foregoing typology—was the result.

The views of Śaṅkara, Rāmānuja, and Radhakrishnan

Śaṅkara, an outstanding nondualistic Vedāntist and advocate of a spiritual view of life, began with the Neoplatonic alternative but added a qualification that turned his view into what was later called acosmic pantheism. Distinguishing first between Brahman as being the eternal Absolute and Brahman as a lower principle and declaring the lower Brahman to be a manifestation of the higher, he then made the judgment that all save the higher unqualitied Brahman is the product of ignorance or nescience and exists (apparently only in men's minds) as the phantoms of a dream. Since for Śaṅkara, the world and individuality thus disappear upon enlightenment into the unmanifest Brahman, and in reality only the Absolute without distinctions exists, Śaṅkara has provided an instance of acosmism.

On the other hand, Rāmānuja, a prominent southern Brahmin who held to a qualitied monism, argued strenuously against Śaṅkara's dismissal of the world and of individual selves as being mere products of nescience. In place of this acosmism he substituted the notion of world cycles. In the unmanifest state Brahman has as his body only the very subtle matter of darkness, and he decrees "May I again possess a world-body"; in the manifest state all of the things of the world, including individual selves, are part of his body. The doctrine of Rāmānuja approaches panentheism; he has certainly advanced beyond emanationistic pantheism. There are two aspects to the single Brahman, one absolutistic and the other relativistic. As in panentheism, the beings of the world have freedom. The only qualification is that, although it is Brahman's will to support the choices of finite beings, he has the power to prohibit any choice that displeases him. This power to prohibit indicates a preference for the absolute in Rāmānuja's thought, which is reflected in many ways: although God is the cause of the world, for example, and includes the world within his being, he is never affected by that world, and his motive in world creation is simply play. In sum, since the absolutistic categories were given the greater emphasis in his thought, Rāmānuja is representative of a relativistic monistic pantheism.

The presence in the Hindu tradition of both absolutistic and relativistic descriptions of the divine suggests that genuine panentheism might well emerge from the tradition; and, in fact, in the former president of India, S. Radhakrishnan, also a religious philosopher, that development did occur. Although Radhakrishnan had been influenced by Western philosophy, including that of A.N. Whitehead, later discussed as a modern panentheist, the sources of his thought lie in Hindu philosophy. He distinguishes between God as the being who contains the world and the Absolute, who is God in only one aspect. He finds

that the beings of the world are integral with God, who draws an increase of his being from the constituents of his nature.

Buddhist doctrines. Some 600 years after Buddha, a new and more speculative school of Buddhism arose to challenge the 18 or 20 schools of Hīnayāna (Sanskrit "Lesser Vehicle") Buddhism then in existence. One of the early representatives of this new school, which came to be known as Mahāyāna (Sanskrit "Greater Vehicle") Buddhism, was Aśvaghoṣa. Like Śaṅkara (whom he antedated by 700 years), Aśvaghoṣa not only distinguished between the pure Absolute (the Soul as "Suchness"; *i.e.*, in its essence) and the all-producing, all-conserving Mind, which is the manifestation of the Absolute (the Soul as "Birth and Death"; *i.e.*, as happenings), but he also held that the judgment concerning the manifest world of beings is a judgment of nonenlightenment; it is, he said, like the waves stirred by the wind—when the quiet of enlightenment comes the waves cease, and an illusion confronts a man as he begins to understand the world.

Whereas Aśvaghoṣa treated the world as illusory and essentially void, Nāgārjuna, the great propagator of Mahāyāna Buddhism who studied under one of Aśvaghoṣa's disciples, transferred *Śūnya* ("the Void") into the place of the Absolute. If Suchness, or ultimate reality, and the Void are identical, then the ultimate must lie beyond any possible description. Nāgārjuna approached the matter through dialectical negation: according to the school that he founded, the Ultimate Void is the Middle Path of an eightfold negation; all individual characteristics are negated and sublated, and the individual approaches the Void through a combination of dialectical negation and direct intuition. Beginning with the Middle Doctrine School, the doctrine of the Void spread to all schools of Mahāyāna Buddhism as well as to the Satyasiddhi (Sanskrit: "perfect attainment of truth") group in Hīnayāna Buddhism. Since the Void is also called the highest synthesis of all oppositions, the doctrine of the Void may be viewed as an instance of identity of opposites pantheism.

In the T'ien-t'ai school of Chinese Buddhism founded by Chih-i, as in earlier forms of Mahāyāna Buddhism, the elements of ordinary existence are regarded as having their basis in illusion and imagination. What really exists is the one Pure Mind, called True Thusness, which exists changelessly and without differentiation. Enlightenment consists of realizing one's unity with the Pure Mind. Thus, an additional Buddhist school, T'ien-t'ai, can be identified with acosmic pantheism.

Indeed, although a mingling of types is discernible in the Hindu and Buddhist strands of Oriental culture, acosmic pantheism would seem to be the alternative most deeply rooted and widespread in these traditions.

Ancient Near East doctrines. Just as the early gods of the Vedas represented natural forces, so the Canaanite deities known as Baal and the Hebrew God Yahweh both began as storm gods. Baal developed into a Lord of nature, presiding with his consort, Astarte, over the major fertility religion of the Near East. The immanentism of this nature religion might have sustained the development of pantheistic systems; but, whereas the pantheistic Puruṣa triumphed in India, the Theistic Yahweh triumphed in the Near East. And Yahweh evolved not into a Lord of nature but into a Lord of history presiding first over his chosen people and then over world history. The requirement that he be a judge of history implied that his natural "place" was outside and above the world; and he thus became a transcendent deity. Through much of the history of Israel, however, the people accepted elements from both of these traditions, producing their own highly syncretistic religion. It was this syncretism that provided the occasion that challenged certain men of prophetic consciousness to embark upon their purifying missions, beginning with Elijah and continuing throughout the Old Testament period. In this development, the absoluteness and remoteness of Yahweh came to be supplemented by qualities of love and concern, as in the prophets Hosea and Amos. In short, the categories of immanence came to supplement the categories of transcendence and, in the New Testament period, became overwhelmingly important. The transcendent Yahweh, on the other hand, had fitted more naturally into the categories of absoluteness. And, in the Christian West, it was the transcendent God who appeared in the doctrines of Classical Theism, while pantheism stood as a heterodox departure from the Christian scheme.

Baal, Yahweh

PANTHEISM AND PANENTHEISM IN ANCIENT AND MEDIEVAL PHILOSOPHY

Early Greek religion contained among its many deities some whose natures might have supported pantheism; and certainly the mystery religions of later times stressed types of mystical union that are typical of pantheistic systems. But in fact the pantheism of ancient Greece was related almost exclusively to philosophical speculation. For this reason it is more rationalistic, possessing a style quite different from the Pantheisms thus far examined.

Ancient Greco-Roman doctrines. The first philosophers of Greece, all of whom were 6th-century-BC Ionians, were hylozoistic, finding matter and life inseparable. The basic substances that they identified as the elements of reality—the water proposed by Thales, the boundless infinite suggested by Anaximander, and the air of Anaximenes—were presumed to have the motive force of living things and thus to be a kind of life, a position here called hylozoistic pantheism.

Pre-Socratic views

Impressed by the absolute unity of all things, the adherents of another philosophic position, that of Eleaticism (*q.v.*), so-named from its centre in Elea, a Greek colony in southern Italy, found it impossible to believe in multiplicity and change. The first step in this direction was taken by Xenophanes, a religious thinker and rhapsodist, who, on rational grounds, moved from the gods and goddesses of Homer and Hesiod to a unitary principle of the divine. He believed that God is the supreme power of the universe, ruling all things by the power of his mind. Unmoved, unmoving, and unitary, God perceives, governs, and apparently contains, or at least he "embraces," all things. So interpreted, Xenophanes provides an instance of monistic pantheism, inasmuch as, in this view, the Absolute God is united with a changing world, while the reality of neither is attenuated. This paradox may have encouraged Parmenides, possibly one of Xenophanes' disciples (according to Aristotle), to accept the changeless Absolute, eliminating change and motion from the world. Reality thus became for him a unitary, indivisible, everlasting, motionless whole. This position is basically that of absolutistic monistic pantheism in that it views the world as real but changeless. Insofar as the change and variety of the world are only apparent, Parmenides also approaches acosmic pantheism.

A third fundamental position is that of the Ephesian critic Heracleitus, among whose cryptic sayings were many that stressed the role of change as the basic reality. Heracleitus continued the hylozoistic tendencies of the Ionian philosophers. Fire, his basic element, is also the universal *logos*, or reason, controlling all things; and since fire not only has a life of its own but exercises control to the boundaries of the universe as well, the system is more complex than hylozoistic pantheism. In view of the circumstance that everything is either on the way from, or to, fire, this basic element is actually or incipiently everywhere. Since the divine works here from within the universe, indeed from within a single, but basic, aspect of it, the system is an instance of immanentistic pantheism.

The philosopher Anaxagoras, one of the great dignitaries at Athens in the golden age of Pericles, approached the problem somewhat in the manner of Heracleitus. *Nous* (or Mind) he held to be the principle of order for all things as well as the principle of their movement. It is the finest and purest of things and is diffused throughout the universe. This, like the preceding system, is an instance of immanentistic pantheism.

From the standpoint of the typology here employed, Plato may be regarded as the first Western philosopher to treat the problem of the absoluteness and the relativity in God with any degree of adequacy. In the *Timaeus* an absolute and eternal God was recognized, existing in

Plato's dualism

changeless perfection in relation to the world of forms, along with a World-Soul, which contained and animated the world and was as divine as a changing thing could be. Although the material can be variously interpreted, panentheists hold that Plato has adopted a dual principle of the divine, uniting both being and becoming, absoluteness and relativity, permanence and change in a single context. To be sure, he envisioned the categories of absoluteness as situated in one deity, and those of relativity in another; but the separation seems not to have pleased him, and in the tenth book of the *Laws*, by invoking the analogy of a circular motion, which combines change with the retention of a fixed centre, he explained how deity could exemplify both absoluteness and change. Plato thus may be viewed as a quasi-panentheist.

Aristotle, on the other hand, with his exclusivistic, transcendent God, exemplifying only the categories of absoluteness, anticipated the absolute God of Classical Theism, existing above and beyond the world.

Stoic pantheism, Neoplatonism

Stoicism, one of the foremost of the post-Aristotelian schools of thought, represents an immanentistic pantheism of the Heracleitean variety. First of all, the Stoics accepted the decision of Heracleitus that an indwelling fire is the principal element entering into all transformations and is also the principle of reason, the *logos*, ordering as well as animating all things, but that, second, there is a World-Soul, which is diffused throughout the world and penetrates it in every part. Rather than approximating Plato's spiritual World-Soul, the Stoic World-Soul is more like the *Nous* of Anaxagoras. The Stoics were Materialists, and their diffuse World-Soul is, thus, an extended form of subtle matter. That everything is determined by the universal reason is an unvarying theme in Stoicism; and this fact suggests that Stoic pantheism, despite its immanentism, stresses the categories of absoluteness rather than those of relativity in the relations holding between God and the world.

The life of reason brings man into harmony with God and with nature and helps him to understand his fate, which is his place in the universal system. Although the view is an amalgam of several types of pantheism, this particular mixture has retained its identity. It is therefore useful to call this position, or any similar combination of themes, by the name Stoic pantheism.

Plotinus, the creator of one of the most thoroughgoing philosophical systems of ancient times, may be taken to represent Neoplatonism, an influential modification of Plato's attempt to deal with absoluteness and relativity in the divine. Plotinus' system consists of the One—the absolute God who is the supreme power of the system—the intermediate *Nous*, and the World-Soul (with the world as its internal content). His World-Soul follows the Platonic model. The system really blends pantheism with Classical Theism, since the categories of absoluteness apply to the One, and the relativistic categories apply to the World-Soul. The doctrine of emanation, whereby the power of the One comes into the world, is a clear attempt to bridge the gap between absoluteness and relativity. For Plotinus, as for Classical Theism, there is immanent in man an image of the divine, which serves as well to relate man to God as does the divine spark in Stoic pantheism. Even Classical Theism may thus contain a touch of immanentistic pantheism. This view, or any similar combination of themes, is an instance of emanationistic or Neoplatonic pantheism.

Medieval doctrines. Though Scholasticism, with its doctrine of a separate and absolute God, was the crowning achievement of medieval thought, the period was, nonetheless, not without its pantheistic witness. Largely through Jewish and Christian mysticism, an essentially Neoplatonic Pantheism ran throughout the age.

The only important Latin philosopher for six centuries after St. Augustine was John Scotus Erigena. Inasmuch as, in his system, Christ's redemptive sacrifice helps to effect a Neoplatonic return of all beings to God, Erigena can be said to have turned Neoplatonism into a Christian drama of fall into sin and redemption from its power. When Erigena said that, even in the stage of separation from God, God in his superessentiality is identical with all things, he advanced beyond a strictly Neoplatonic pantheism to some stronger form of immanentistic or monistic pantheism.

In the two principal writings of the esoteric Jewish movement called the Kabbala, known for its theosophical interpretations of the Scriptures, a mystically oriented system of ten emanations is presented. A Spaniard, Avicebrón, a Jewish poet and philosopher, similarly presented a Neoplatonic scheme of emanations. And in Spain, Averroës, the most prominent Arabic philosopher, represented an Aristotelian tradition heavily overladen with Neoplatonism. For Averroës, the active intellect in man is really an impersonal divine reason, which alone lives on when man dies.

The German Meister Eckehart, probably the most significant of philosophical mystics, developed a markedly original theology. From his Stoic pantheism there arose his most controversial thesis—that there resides in every man a divine, uncreated spark of the Godhead, making possible both a union with God and a genuine knowledge of his nature. But Eckehart also distinguished between the unmanifest and barren Godhead and the three Persons who constitute a manifest and personal God. Thus, the system has similarities to both Stoic and Neoplatonic pantheism.

Cardinal Nicholas of Cusa, whose broad scholarship and scientific approach anticipated the coming Renaissance, continued the tradition into the 15th century. The "learned ignorance," in which a man separates himself from every affirmation, can have positive results, in Nicholas' view, because man is a microcosm within the macrocosm (or universe), and the God of the macrocosm is thus mirrored in all of his creatures. He also held that, in reference to God, contradictions are compatible —his "coincidence of opposites" doctrine, in which God is at once all extremes. Clearly, Nicholas wished to ascribe to God both the categories of transcendence and those of immanence without distinction. But in fact he displayed some preference for the categories of the absolute, insisting, for example, that the creatures of the world can add nothing to God since they are merely his partial appearances. Despite this bias toward absolutism, and even to acosmism, Nicholas can be appropriately viewed as espousing an identity of opposites pantheism.

PANTHEISM AND PANENTHEISM IN MODERN PHILOSOPHY

Renaissance and post-Renaissance doctrines. The humanism of the Renaissance included an enlarged interest in Platonism and in its historical carrier, Neoplatonism, as well as influences from Aristotle and from Kabbalistic sources. The view of man as a microcosm of the universe was widespread. Marsilio Ficino, one of the first leaders of the Florentine Academy, found the image and reflection of God in all men and anticipated the divinization of man and the entire cosmos. The humanist and syncretistic philosopher Pico della Mirandola, also a leading figure in the Academy, substituted for creation a Neoplatonic emanation from the divine.

The views of Ficino, Bruno, and Böhme

The most famous scholar of the Italian Renaissance was Giordano Bruno. Combining Copernican astronomy with Neoplatonism, Bruno thought of the universe as an infinite organism with monads as its ultimate constituents and world-systems as its parts. The universe, he held, is in a continual process of development and is infused with the divine life. Accepting Nicholas of Cusa's doctrine of the identity of opposites, he taught that contradictory ascriptions apply equally to God in particular and that claims concerning his immanence and transcendence are equally valid. More open to the categories of relativity than Nicholas, Bruno, however, exemplified a neatly balanced instance of identity of opposites pantheism.

The next great innovator of mystical religious thought was Jakob Böhme, who, in developing the concept of the divine life, took a decisive step beyond mere absoluteness. God goes through stages of self-development, he taught, and the world is merely the reflection of this process. Böhme anticipated Hegel in claiming that the divine self-development occurs by means of a continuing dialectic, or tension of opposites, and that it is the negative qualities

of the dialectic that men experience as the evil of the world. Even though Böhme, for the most part, stressed absoluteness and relativity equally, his view that the world is a mere reflection of the divine—apparently denying self-development on the part of creatures—tends toward acosmic pantheism.

Spinoza's rational pantheism

In the 17th century the foremost pantheist was a Jewish rationalist, Benedict Spinoza, whose training in the history of philosophy included both medieval Jewish philosophy and the Kabbala. He championed a rational rather than a mystical pantheism, so much so that all that remained of mysticism, in fact, was his concept of the intellectual love of God. The rationality of the system is suggested by Spinoza's argument that, since God is the infinite being, he must be identical with the world; for otherwise, God-and-world would be a greater totality than God alone. Also, since God is a necessary being and is identical with the world, the world must also be necessary in all its parts. It follows from this that human freedom is an impossible idea; and the sense that man has of such freedom is based on his ignorance of the causes that have determined him. Spinoza distinguished between God and the world in three ways: first, by stressing God's activity in the active sense of *natura naturans* ("the nature that [creates] nature"; *i.e.*, God) compared to the passive sense of *natura naturata* ("the nature that [is created as] nature"; *i.e.*, the world); second, he related God to eternity and the world to time; and third, he distinguished God as self-existing substance, the whole, from the world, which he conceived as the attributes and modes of that substance. In terms of the present classification, Spinoza represents a monistic pantheism tending toward absolutism.

Goethe, the incomparable German litterateur, claimed that he was a follower of Spinoza. In fact, however, his beliefs were rather different inasmuch as Goethe championed man's individuality; opposed mechanical necessity; and held a hylozoistic, or vitalistic, position in which nature was organic, a living unity. His personalistic pantheism mixes hylozoistic and Stoic types with a touch of relativism added to the mixture.

19th-century doctrines. During the 19th century, pantheism and panentheism were sustained by various kinds of Idealism that developed during the period. In these systems the categories of relativity gained in prominence; God was conceived as entering history and as being more intimately related to processes of change and development.

German Idealism. Although the philosophy of the German patriot J.G. Fichte, an immediate follower of Kant, began in the inner subjective experience of the individual, with the "I" positing the "not-I"—*i.e.*, feeling compelled to construct a perceived world over against itself—it turns out eventually that, at a more fundamental level, God, as the universal "I," posits the world at large. The world, or nature, is described in organic terms; God is considered not alone as the Universal Ego but also as the Moral World Order, or Ground of ethical principles; and since every man has a destiny as a part of this order, man is in this sense somehow one with God. In the moral world order, then, man has a partial identity with God; and in the physical order he has membership in the organic whole of nature. It is not clear, however, whether in Fichte's view God as Universal Ego includes all human egos, and the organic whole of nature. Should he do so, then Fichte would be a representative of dipolar Panentheism, since in his final doctrine the Universal Ego imitates an Absolute deity who is simply the divine end of all activity, serving equally as model and as goal. In this interpretation God is conceived both as absolute mobility and absolute fixity. It is not entirely clear whether the doctrine is to be understood as referring to two aspects of a single God, the panentheistic alternative, or to two separate gods, the alternative imbedded in Plato's quasi-panentheism. In either case, Fichte has enunciated most of the themes of panentheism and deserves consideration either as a representative or precursor of that school.

A second early follower of Kant was F.W.J. von Schelling, who, in contrast to Fichte, stressed the self-existence of the objective world. Schelling's thought developed through several stages. Of particular interest to the problem of God are the final three stages in which his philosophy passed through monistic and Neoplatonic pantheism followed by a final stage that was panentheistic.

In the first of these stages, he posits the Absolute as an absolute identity, which nonetheless includes, as in Spinoza, both nature and mind, reality and ideality. The natural series culminates in the living organism; and the spiritual series culminates in the work of art. The universe is, thus, both the most perfect organism and the most perfect work of art.

In his second, Neoplatonic, stage he conceived the Absolute as separated from the world, with a realm of Platonic ideas interposed between them. In this arrangement, the world was clearly an emanation or effect of the divine.

In the final stage of his thought, Schelling presented a theophany, or manifestation of deity, involving the separation of the world from God, and its return. In appearance this was quite like the views of Erigena or like the unmanifest and manifest Brahman of Indian thought. But, since the power of God continues to infuse the world and there can be no real separation, the entire theophany is clearly the development of the divine life. The Absolute is retained as the pure Godhead, a unity presiding over the world; and the world—having in measure its own spontaneity—is both his antithesis and part of his being, the contradiction accounting for progress. The positing within God of eternity and temporality, of being-in-itself and of self-giving, of yes and no, of participation in joy and in suffering, is the very duality of Panentheism.

It was a disciple of Schelling, Karl Christian Krause, who coined the term panentheism to refer to the particular kind of relation between God and the world that is organic in character.

Hegel's notion of Absolute Spirit

The third, and most illustrious, early post-Kantian Idealist was G.W.F. Hegel, who held that the Absolute Spirit fulfills itself, or realizes itself, in the history of the world. And in Hegel's deduction of the categories it is clear that man realizes himself through the attainment of unity with the Absolute in philosophy, art, and religion. It would appear, then, that God is in the world, or the world is in God, and that, since man is a part of history and thus a part of the divine realization in the world, he shares in the divine life; it would seem, too, that God is to be characterized by contingency as well as necessity, by potentiality as well as actuality, by change as well as permanence. In short, it would seem at first that the panentheistic dipolarity of terms would apply to the Hegelian Absolute. But this is not quite so; for Hegel's emphasis was on the deduction of the categories of logic, nature, and spirit, a deduction that provided the lineaments of Spirit-in-Itself (the categories of the intrinsic logic that the world, as Spirit, follows in its development), Spirit-for-Itself (nature as existing oblivious to its own context), and Spirit-in-and-for-Itself (conscious spiritual life, natural, and yet aware of its role in the developing world). This deduction, moving from the most abstract categories to the most concrete, is partly logical and partly temporal; it cannot be read either as a sheerly logical sequence or as a sheerly temporal sequence. As a logical sequence, it has the appearance of a Neoplatonic scheme turned on its head, since the Absolute Spirit that emerges from the deduction includes all of the steps of the preceding rich and multifarious deduction. As a temporal sequence, the system would seem to be a species of Stoic (*i.e.*, Heracleitean) pantheism, qualified by a clear Parmenidean motif (see above), which appears in its stress on an absoluteness that, from the eternal standpoint, cancels out time. This Parmenidean quality is to be found not only in Hegel but in most of the Idealists who were influenced by him. Time is real, on this view, and yet not quite real, having already eternally happened. And when Hegel spoke of the Absolute Spirit, this phrase held the internal tension of a near contradiction, for spirit, however absolute, must surely be relative to what is around it, sensitive to and dependent on other spirits. The fact that Hegel wished to give something like equal emphasis, however, both to absoluteness

and to relativity in the divine being or process suggests that his goal is identical with that of the panentheists, even though he is perhaps more fairly regarded as a Pantheist of an ambiguous type.

Monism and panpsychism. It is impossible for one to leave the 19th century without mention of the pioneering experimental psychologist Gustav Fechner (1801–87), founder of psychophysics, who developed an interest in philosophy. Fechner pursued the themes of panentheism beyond the positions of his predecessors. A panpsychist with an organic view of the world, he held that every entity is to some extent sentient and acts as a component in the life of some more inclusive entity in a hierarchy that reaches to the divine Being, whose constituents include all of reality. God is the soul of the world, which is, in turn, his body. Fechner contends that every man's volitions provide impulses within the divine experience, and that God gains and suffers from the experiences of men. Precisely because God is the supreme being, he is in process of development. He can never be surpassed by any other, but he surpasses himself continually through time. He, thus, argues that God can be viewed in two ways: either as the Absolute ruling over the world, or as the totality of the world; but both are aspects of the same Being. Fechner's affirmations comprise a complete statement of panentheism, including the dipolar deity with respect to whom the categories of absoluteness and relativity can be affirmed without contradiction.

Fechner's panentheism

Twentieth-century doctrines. The 20th century marks a decisive break with absolutism. In the first half of the century, panentheism gained in authority. The position of the Russian ex-Marxist Nikolay Berdyayev, a religious metaphysician, with his emphasis on divine and human freedom, is a manifesto of panentheism. Even more impressive was the work of the eminent British-American philosopher, Alfred North Whitehead. As in the case of Fechner, Whitehead came to philosophy from science and held an organismic view of the structure of the world. In Whitehead's view God has two natures: his primordial nature is abstract; his consequent nature is concrete and includes within itself the total history of the world. Whitehead was also a panpsychist and believed that feeling is present in some degree at every level of the world process. Whether or not he was, then, also a panentheist is in dispute. He held that the possible future and the total past are in God—in his primordial and consequent natures; but for Whitehead the present moment is relative, and contemporaries exclude each other. In the present moment of any entity, since it is the present of *that* entity, it is appropriate to say that God is in that entity, part of the data on which it acts; thus the Stoic spark of divinity has here a modern application. From the standpoint of God, on the other hand, all entities are part of God; they come from him and return to him in the passage of time, but they are not in God in the sense that their independence in the present moment is prejudiced.

The views of Whitehead and Hartshorne

It was left to Charles Hartshorne, one of Whitehead's followers, to provide the definitive analysis of panentheism. It is Hartshorne's suggestion that the organismic analogy, present in Whitehead as well as in many earlier thinkers, be taken seriously. For Hartshorne, God includes the world even as an organism includes its cells, thus including the present moment of each event. The total organism gains from its constituents, even though the cells function with an appropriate degree of autonomy within the larger organism.

CRITICISM AND EVALUATION OF PANTHEISM AND PANENTHEISM

Panentheism is then a middle way between the denial of individual freedom and creativity characterizing many of the varieties of pantheism and the remoteness of the divine characterizing Classical Theism. Its support for the ideal of human freedom provides grounds for a positive appreciation of temporal process, while removing some of the ethical paradoxes confronting deterministic views. It supports the sacramental value of reverence for life. At the same time the theme of participation with the divine leads naturally to self-fulfillment as the goal of life.

Many pantheistic and Theistic alternatives claim the same advantages, but their natural tendency toward absoluteness may make justification of these claims in some cases difficult and, in others, some argue, quite impossible. It is for this reason that a significant number of contemporary philosophers of religion have turned to panentheism as a corrective to the partiality of the other competing views.

BIBLIOGRAPHY. CHARLES HARTSHORNE and W.L. REESE, *Philosophers Speak of God* (1953), offers an extensive historical exploration of pantheism, panentheism, and Classical Theism. The fundamental basis of panentheism is discussed not only in the epilogue of the above volume, but in many other works by CHARLES HARTSHORNE, including *The Divine Relativity* (1948) and *The Logic of Perfection* (1962). For the relation of mysticism to pantheism, see W.T. STACE, *Mysticism and Philosophy* (1960). For information concerning any of the philosophers mentioned, reference may be made to their individual entries in *The Encyclopedia of Philosophy*, 8 vol., ed. by PAUL EDWARDS (1967); the *Enciclopedia filosofica*, 6 vol., 2nd ed., ed. by G.C. SANSONI (1967); or the *Diccionario de filosofía*, 2 vol., ed. by JOSE FERRATER MORA (1965).

(W.L.Re.)

Papacy

The papacy is the system of central government of the Roman Catholic Church, the largest of the three major branches of Christianity, presided over by the bishop of Rome, who is recognized by Roman Catholics as the pope and the successor of St. Peter, the "chief of the Apostles" of Jesus Christ.

This article is divided into the following sections:

NATURE AND SIGNIFICANCE

The word papacy (Latin *papatia*, derived from *papa*, pope; *i.e.* father) is of medieval origin. In its primary usage it denotes the office of the pope (of Rome), and, hence, the system of ecclesiastical and temporal government over which he directly presides. It is with the word in this latter sense that this article is concerned.

The papal office. The multiplicity and variety of papal titles themselves indicate the complexity of the papal office. In the *Annuario Pontificio*, the official Vatican directory, the pope is described as bishop of Rome, vicar of Jesus Christ, successor of the prince of the Apostles, supreme pontiff of the universal church, patriarch of the West, primate of Italy, archbishop and metropolitan of the Roman province, sovereign of the state of Vatican City, servant of the servants of God. In his more circumscribed capacities as bishop of Rome, metropolitan of the Roman province, primate of Italy, and patriarch of the West, the pope is the bearer of responsibilities and the wielder of powers that have their counterparts in the other episcopal, metropolitan, primatial, and patriarchal jurisdictions of the Roman Catholic Church. What differentiates his particular jurisdiction from these others and renders his office unique is the Roman Catholic teaching that the bishop of

Titles of the pope

Rome is at the same time successor to St. Peter, prince of the Apostles. As the bearer of the Petrine office, he is raised to a position of lonely eminence as chief bishop or primate of the universal church. The precise lineaments of this belief have undergone considerable development over the centuries, but as defined by the first Vatican Council in 1870 (and reaffirmed by the second in 1964) it involves these fundamental assertions: that Christ singled out St. Peter among the Apostles, conferring upon him not only a pre-eminence among them but also the leadership or primacy in that visible community or church which he established after his Resurrection; that Christ intended this primacy to be one not merely of honour but of true jurisdiction, and one exercised not merely by Peter himself but also by his successors down through the ages to the end of time; and that the bishops of Rome are to be recognized as these successors.

Papal primacy and functions

Because of the development of this belief, the popes have come over the course of time to wield supreme legislative, executive, and judicial powers (jurisdictional functions) in the church—issuing authoritative statements on matters of doctrine (the magisterial or teaching function), creating and suppressing church laws, establishing dioceses, appointing bishops, controlling missions, acting as a court of first instance as well as of appeal, and performing, either by deputy or in person, a host of other functions. Also, because of this belief, they had assumed already in the 5th century the title of supreme pontiff (*summus pontifex*) that had earlier been borne by the pagan Roman emperors as heads of the college of priests. During the Middle Ages the popes also began to monopolize the title of "vicar [*i.e.*, representative] of Christ," which, like the very name of "pope" for that matter, the early Latin Church had customarily accorded to bishops in general. The title of sovereign of the state of Vatican City refers to the pope's position as temporal ruler of the tiny sovereign state in Rome created in 1929 by Italy in accordance with the terms of the Lateran Treaty. The last title, "servant of the servants of God," an essentially pastoral designation, is the title that popes themselves have very often chosen to emphasize when called upon to issue solemn pronouncements of great importance for the whole church.

The instruments of papal government. In the day-to-day exercise of his primatial jurisdiction the pope relies on the assistance of the Roman Curia, a name first used of the body of papal assistants in the 11th century. The Curia had its origins in the local body of presbyters (priests), deacons (lower order of clergy), and notaries (lower clerics with secretarial duties) upon which, like other bishops in their own dioceses, the early bishops of Rome relied for help. By the 11th century, this body had, on the one hand, been narrowed down to include only the leading (or cardinal) presbyters and deacons of the Roman diocese, while, on the other hand, being broadened out to embrace the cardinal-bishops (the heads of the seven neighbouring, or "suburbicarian," dioceses). From this emerged the Sacred College of Cardinals, a corporate body possessed, from 1179 onward, of the exclusive right to elect the pope. This right it still possesses, as it does the right to govern the church in urgent matters during a vacancy in the papal office. Recent popes have extended the size of the Sacred College beyond the traditional limit of 70 and have attempted, with growing success, to broaden its national complexion and to make its membership more faithfully representative of the church's international character.

The Roman Curia

During the Middle Ages, the cardinals played an important role as a corporate body, not only during papal vacancies, as today, but also during the pope's lifetime. In the 12th century, the Roman councils that popes had hitherto convoked when urgent matters were at hand were replaced by the assembly of the cardinals, or consistory, which thus became the most important collegial (corporate) body advising the pope and participating in his judicial activity. Eventually it began to make oligarchic claims to a share in the powers of the Petrine office, and attempted with sporadic success, to bind the pope to act on important matters only with its consent. During the 16th century, however, with the final establishment of the Roman Congregations (administrative committees), each charged with the task of assisting the pope in a specific area of government, the significance of the consistory began to decline, and with it the importance of the cardinals as a corporate body. At the same time, there was an increase in the power and influence of the "curial" cardinals—those cardinals who did not administer local dioceses but served as the pope's representatives in important foreign affairs or resided permanently in Rome, holding responsibilities in the curial congregations, tribunals, and offices that proliferated in the course of the next three centuries.

By the early 20th century, the growth of the Roman Curia had produced a bewildering tangle of administrative and judicial bodies, in which neither temporal and ecclesiastical functions nor executive and judicial powers were clearly demarcated. The reforms of Pius V (reigned 1566–72) and Benedict XV (reigned 1914–22) clarified and streamlined the work of the Curia, introducing a measure of order into its maze of overlapping jurisdictions. But in the wake of the complaints about abuses of curial power that were voiced at the second Vatican Council (along with requests for an internationalization of curial staff and a modernization of curial functions and procedures), Paul VI pledged himself to act.

Modern reforms of the Curia

Though Paul VI made some changes in detail, his reforms leave intact the basic curial structure created by Pius V, with its tripartite division of the various curial bodies: the Roman Congregations composed of cardinals nominated by the pope (*e.g.*, the Congregation for the Doctrine of the Faith, which was the former Holy Office and direct descendant of the Roman Inquisition); the tribunals, three in number, which compose the judicial branch of the Curia and one of which, the Rota, handles matrimonial cases; a group of offices, councils, and secretariates, the most important of which is the Secretariat of State, presided over by the cardinal secretary of state, who now emerges as the pope's "prime minister." To promote a higher degree of coordination among the various jurisdictions, provision was made for regular meetings of department heads, summoned and presided over by the cardinal secretary of state. Similarly, to prevent bureaucratic empire-building, most curial appointments were to be made for an initial term of five years. Finally, in response to Vatican II's request, some diocesan bishops were to be present at the plenary sessions of the congregations, efforts were to be made to internationalize the curial staff, and there was to be some attempt to consult the laity. These reforms came into effect in January 1968.

HISTORY

The early bishops of Rome to Leo I. Historical evidence concerning the early Christian congregation at Rome is not extensive; that concerning its first bishops is extremely scanty. That St. Paul preached at Rome and was probably put to death there during Nero's reign (*c.* AD 67) is not contested. Somewhat less certain is the claim that St. Peter, too, visited Rome and was martyred there about the same time. Few scholars would endorse the claim that the excavations (1940s–60s) under St. Peter's Basilica have culminated in the discovery of Peter's tomb. Most would agree at the very least, however, that the excavations reveal that some Roman Christians of the early 3rd century believed that Peter's grave was on that site. This adds further weight to the already weighty tradition of Peter's presence at Rome—a tradition attested to by several important pieces of literary evidence from *c.* 96 onward, emanating from Africa and Asia Minor as well as from Rome itself, challenged by no rival tradition and contested by no contemporary witness.

St. Paul and St. Peter in Rome

Though none of the evidence indicates anything definite about the nature of Peter's position at Rome, many exegetes (interpreters employing critical techniques), both Roman Catholic and Protestant, agree that Christ singled out Peter among the Apostles and conferred upon him a certain pre-eminence. There is no necessary deduction, however, from Peter's primacy among the Apostles to the primacy of those early bishops of Rome who may have claimed to be his successors—the less so in the absence of any historical evidence permitting the assertion that

Peter exercised an office of episcopal nature at Rome. Not much light is shed on the question of the primacy by materials from the early post-apostolic era, either by the first lists of the bishops of Rome or by such important and much exploited pieces of evidence as I Clement, a letter sent by the church of Rome to the church of Corinth *c.* 96, or the letter that Ignatius of Antioch sent to the Romans *c.* 110. The interpretation of these and of similar data has been disputed. If read in historical context, what they actually state constitutes a less than persuasive witness to the primacy of the Roman bishops in the universal church. And what one reads into their silences depends very much upon one's theological presuppositions. Clement's letter is striking evidence, however, for the prominence of the Roman Church at the end of the 1st century, and, taken together, these early data indicate that in doctrinal matters, at least, the Roman Church was accorded a certain pre-eminence even among the handful of churches that claimed apostolic foundation and were therefore regarded as embodying the apostolic teaching in its full purity. They indicate, too, that this pre-eminence was intimately connected with the impressive apostolic credentials of the Roman Church and is not explicable solely by the fact that Rome was the imperial capital.

Early claims to primacy. By the 3rd century the Roman bishops were claiming for themselves a primacy of authority in the universal church comparable to that of Peter's primacy among the Apostles. This claim was asserted on doctrinal matters and did not go unopposed, a fact attested to by the vigorous resistance of St. Cyprian of Carthage to Pope Stephen I (reigned 254–257). Only in the 4th and 5th centuries was this claim transformed into the more sweeping claim to a primacy of jurisdiction. This was a response of the Roman bishops to the challenge posed by the then growing pre-eminence and mounting claims of the see of Constantinople (the seat of the patriarch of the capital of the Eastern Roman, or Byzantine, Empire). It also reflected their increasing tendency, as leading dignitaries of a civic church supported by the coercive force of the imperial administration, to understand their own primatial claim in juristic rather than biblical categories. Most actively responsible for this development were the popes Damasus I, Siricius, Innocent I, Boniface I, and, above all, Leo I (reigned 440–461). The gap between theory and practice, however, remained immense. Despite Leo's protests, the Council of Chalcedon in 451 decreed that the New Rome (Constantinople) was "to enjoy the same primacy" in the East as the Old Rome enjoyed in the West, revealing thereby the impact of the enhanced political status of Constantinople in the empire and also what was destined to be an enduring Eastern inability to understand the Petrine theory in the same terms as it was understood in the West. Although in his own day—in a Western patriarchate ravaged by barbarian invasion and the political and ecclesiastical disorganization that went with it—Leo's theoretical claims were generally conceded, it was to be centuries before they were effectively put into practice. In the meantime, it was the emperors residing in Constantinople who—summoning general councils (composed of bishops and theologians to decide doctrinal and ecclesiastical matters) as they had done since the time of Constantine (who convened the Council of Nicaea in 325)—could make the stronger claim to be the functioning supreme leaders of the Christian world, even in spiritual matters.

Opposition to the primacy of the bishop of Rome

Gregory the Great. The popes who succeeded Leo I, though they were able to play a significant, if tortuous, role in the great Christological controversies (concerning doctrines about the divine and human nature of Christ) of that period, were to a very considerable degree at the mercy of events. Threatened by the disintegration of the Western Empire, weakened as patriarchs of the West by the loss of the African, Spanish, and Italian provinces to rulers of heretical persuasion, their relationship with the emperors ruling from Constantinople continued to be one of subservience, except for Gelasius I (reigned 492–496). Gelasius, however, enjoyed the support of the Ostrogothic king Theodoric, and his pontificate predated that intensified imperial surveillance of papal activity that followed Justinian's reconquest of Italy (555). As it was, imperial influence survived even the Lombard invasion of Italy (568) and the actual demise of imperial protection.

In this respect, Gregory I, called the Great (reigned 590–604), whose fate it was to face a Lombard threat to Rome itself with no support from the emperor in Constantinople and to purchase Rome's freedom through tribute, in no way broke with his predecessors. Like them, he insisted that his own jurisdiction was universal; like them he sought to resist the growing pretensions of the patriarchs of Constantinople. Again, like his predecessors', and as befitted a former civil servant of Rome, his loyalty to the empire was almost instinctive. His writings reveal him to have been a man whose preoccupations were overwhelmingly pastoral, and it is altogether appropriate that he should have adopted the title of "servant of the servants of God." His supervision of the Italian and African churches, his efforts to convert the Lombards and Anglo-Saxons, to sponsor reform in the Frankish Church, and to re-establish links with the newly converted Visigoths in Spain were, by his own scale of values, the tasks truly congruent with his office. He took it upon himself to supervise the civil administration of Rome and to negotiate with the Lombards. This he did as the principal surviving civilian official at a time when the imperial government was clearly no longer effective and as the one who, by his reorganization of the vast landholdings of the Roman Church throughout Italy and Sicily, alone possessed the resources needed to feed the Romans, to pay the soldiers, and to buy off the invaders.

The claims and accomplishments of Gregory I

The early medieval papacy. To the inhabitants of the Roman world, Gregory I remained a leading dignitary in what was still an imperial church. To the barbarians of the West he personified the majesty of Rome, exercising over them influence rather than jurisdiction. Thus, despite his commitment of the papacy to a role of positive leadership in the Western regions (*e.g.*, Italy, Gaul, Spain, etc.) that escaped the Eastern Roman imperial orbit, it is perhaps more appropriate to regard Gregory I as a transitional figure than as "the first of the medieval popes." Only in the 8th century, after Islām had dealt shattering blows to the Eastern (Byzantine) Empire by conquering Byzantine-controlled areas, such as Egypt, Syria, the Holy Land, etc., and when the Lombards had renewed their drive to the south in Italy, did the popes finally turn to the West. The outcome of their connection with the Frankish monarchy was to be of the utmost importance. Though the Donation of Constantine (a document connected with later Pseudo-Isidorian decretals and forged in the mid-8th century presumably by pro-Roman clerics) furnished the popes with a persuasive though spurious title to the central Italian territory known later as the Papal States, it was the Frankish armies that forced the transfer of these nominally Byzantine territories from Lombard into papal hands. With the advent of Charlemagne, who exercised authority in matters ecclesiastical as well as civil with Byzantine impartiality, and who in 774 seized the Lombard crown for himself, it soon became clear that the papacy, in linking its fortunes with the Carolingians and committing itself to the exercise of temporal power in Italy, had not won the freedom of action that it had sought under Byzantine rule. But the period of Carolingian dominance was short.

Donation of Constantine

Ninth- and 10th-century developments affecting the papacy. By the latter half of the 9th century, dynastic dissension within and renewed invasion from without had so weakened the Carolingian monarchs that, for a few brief years, popes of the calibre of Nicholas I and John VIII were able not only to emphasize the primacy of honour due to the papacy but even to reaffirm with considerable force its claim to a primacy of jurisdiction, in spite of Byzantine opposition. By emphasizing the precedents that the papal coronation and anointing of emperors had repeatedly established during the earlier part of the century, the popes were also able to reinterpret that claim to involve the title to leadership of Christian society in matters temporal as well as spiritual.

In so extending their claims, these 9th-century popes were aided by factors not always of their own shaping.

Their claim to a primacy of jurisdiction drew marked theoretical support from the Frankish collection of canons later known as the Pseudo-Isidorian decretals (partially forged documents that upheld the freedom of the church from secular interference) and accepted as authentic during the Middle Ages. Again, their prestige was enhanced by the decline under Muslim rule of their ancient rivals in apostolic dignity—the patriarchates of Jerusalem, Alexandria, and Antioch. But their prestige was enhanced most by the ultimate effect of papal encouragement of evangelization and church reform in England, France, and Germany, and by the fact that to the newly converted peoples of the West the popes represented both the lost glamour of imperial Rome and the tradition of apostolic succession.

Decline of papal influence and prestige

Without this accumulated prestige and the precedents established by the 9th-century popes, the claim to primacy would have had difficulty in surviving the subsequent period of papal decadence. In the 870s the imperial government in Italy declined in influence, and the bishopric of Rome, along with other European bishoprics, was increasingly at the mercy of the local nobility, with only spasmodic interventions by the 10th-century German emperors who were powerless to effect any permanent deliverance. German "protection," however, had its own price. When the emperor Henry III descended into Italy in 1046, deposing three rival claimants to the papacy (Sylvester III, Gregory VI, and Benedict IX) and then appointing his own candidate, Clement II (and, later, several successors), the Roman Church was itself in grave danger of becoming an imperial proprietary church, similar to those multitudinous lower churches in Europe whose royal or aristocratic owners regarded them, in accordance with age-old custom, as their own private property to be disposed of at will. But with the advent in 1049 of Leo IX, the third of the popes appointed by Henry III, reform came to Rome.

The Hildebrandine papacy and the Investiture Conflict. Church reform, sponsored by two movements within monasticism in Burgundy and Lorraine and enjoying strong support from Henry III, had become very active and widespread. It was weakened, however, by the lack of unified guidance and control. Leo IX, a representative of the Lorraine movement, set out to provide the necessary guidance. He abandoned the preoccupation with local Italian politics that had characterized his immediate predecessors, and appeared in person at a whole series of synods in France and Germany to investigate, to judge, and to legislate.

Leo's reforming efforts were directed toward eliminating the corruptions of simony (the buying and selling of ecclesiastical office) and of clerical marriage (which, despite its canonical prohibition in the Latin Church, had become widespread). Both corruptions reflected the extent to which, during centuries of invasion and chaos, family and political interests had triumphed over the church's spiritual goals. At a deeper level they reflected the degree to which, in ecclesiastical and temporal realm alike, the crucial distinction that the Romans had made between the holding of office and the possession of property had become blurred. As a result, there had been a gradual extension of lay control (aristocratic, royal, or imperial) over the disposal of ecclesiastical benefices (properties), a control symbolized in the case of the higher churches by the ceremony of investiture in which the lay ruler conferred the ecclesiastical office upon his chosen nominee. Though enjoying the reforming collaboration of the emperor Henry III, Leo did not challenge this control. After many setbacks and disappointments, the more radical reformers were already beginning to doubt that the traditional objectives of moral reform could really be secured without the curtailment of lay interference. Only, however, when Henry III was gone, when a royal minority left the Roman Church without imperial protection, and when the Roman nobility seized the opportunity to try to regain control over the making of popes, did the papacy embrace a more radical approach, and only then because it was able to gain the support of new allies—hence the significance of Nicholas II's brief pontificate (1059–61). In 1059 he allied himself with the Normans of southern Italy, who thus became vassals of the papacy, pledged to support it militarily; and at a Roman synod he promulgated the historic decree placing the election of popes in the hands of the cardinal-bishops.

Reform of papal elections

The object of the decree was to destroy the power of the local aristocracy over the making of popes, but it encroached deeply upon an established imperial prerogative, nevertheless, and German opposition was immediate. Not until 1075, however, did that opposition erupt into an open battle, when the new German king, Henry IV, intent on restoring the religio-political status quo of Henry III's time but lacking his father's zeal for reform, finally clashed with Gregory VII.

Gregory VII (Hildebrand). From the early 1060s until his election in 1073, Gregory VII (Hildebrand) had been the leading radical among the reformers at Rome, and in 1075 he issued a blunt decree threatening with excommunication any prince who presumed to invest anyone with ecclesiastical office. In so doing, Gregory was not simply responding to Henry's lack of reforming zeal or to the pressure of events. Instead, it was his firm conviction that the papal primacy of jurisdiction in the universal church was no longer to be minimized. For Gregory this primacy involved an active exercise of jurisdiction in the local churches, which encroached markedly upon the cherished autonomy of the bishops. It also exalted the papacy to a position in the Christian commonwealth that involved some sort of superiority even over temporal rulers. The wholly unprecedented claim was expressed in the 12th clause of the document known as the *Dictatus Papae* ("Dictates of the Pope," 1075) and inserted in the papal register—"That he [the pope] may depose emperors."

The Investiture Conflict

On this remarkable claim Gregory acted twice during the course of the tragic conflict (Investiture Conflict) that broke out between emperor and pope late in 1075, and that outlasted both of the original protagonists. The conflict ended only in 1122 when Henry V agreed to the Concordat of Worms—a compromise settlement that failed to eliminate all vestiges of royal control over the higher churches and left lay control over the lower churches untouched.

The papacy at its height: the 12th and 13th centuries. Gregory VII has often been portrayed as an innovator who lacked both authentic ancestors and true successors. It must be affirmed, nonetheless, that the later history of the papacy, modern as well as medieval, was shaped by what he and his followers did, while the continuing disabilities characteristic of the medieval papacy owed much to what they left undone. Thus, the assimilation of the biblical notion of church office as grounded in love for others to the political notions of office as grounded in power and law—a development in process since the 4th century and earlier—reached a point of no return with Gregory. He functioned within a unified Christian society in which "state" and "church" were no longer conceived as distinct societal entities and was thus impelled by its very dynamic to assert a claim to jurisdictional supremacy even over the Christian emperor. For the next two centuries papal history was characterized by a deepening involvement, direct and indirect, in matters political. As a result, there were, under Alexander III (reigned 1159–81) and Innocent IV (reigned (1243–54), renewed clashes with the German emperors, and, under Innocent III (reigned 1198–1216), extensive and damaging papal interference in German internal affairs. What alarmed these popes was the fear that imperial policy, by encroaching upon papal territorial independence, also threatened the autonomy of papal action. But with Innocent IV, at least, such a fear was matched by his wish to vindicate, even in temporal matters, the papal claim to ultimate supremacy.

Though much of the drama of papal history in this period focuses upon these conflicts, the impact which the thoroughgoing politicization of church office had upon the nature and structure of ecclesiastical government and the pope's place in it was of more enduring significance. Here again Gregory's pontificate was something of a watershed. Any lingering belief that the pope's primacy

might be regarded primarily as one of honour was now dispelled, and any hesitation about implementing the jurisdictional primacy that had supplanted it now disappeared. The need for papal leadership was so widely accepted that throughout much of the 12th and 13th centuries the demand for it came from the local churches themselves. The outcome was an acceleration in the process that had led, by the late 13th century, to a papal exercise of judicial supremacy going far beyond the mere acceptance of appeals from lower courts: to an arrogation of the wide-ranging legislative powers manifest in the *Decretals* of Gregory IX (1234), the first officially promulgated collection of papal laws, and to the system of "papal provisions" (direct papal intervention in the disposal of benefices) that was finally to be completed by Benedict XII in 1335.

The papal monarchy. Papal leadership *in* the church was eventually replaced by papal monarchy *over* the church. Positively, this transformation was evident in the reforming legislation of the fourth Lateran Council (1215). The negative aspect was to become increasingly obvious as the 13th century wore on. It was no accident that what turned out to be the permanent schism between the Latin and Greek churches had occurred in 1054 at a time when Leo IX had embarked upon a more active exercise of the papal primacy. The more his successors succeeded in establishing the fullness of their jurisdictional power (*plenitudo potestatis*) within the Latin Church, the less chance there was of healing the schism. Nor did papal sponsorship of the Crusades, however great the prestige it had brought to Urban II at the time of the First Crusade, ultimately redound to the benefit of the religious life of the church. Least of all was the administrative centralization attendant upon the exercise of the *plenitudo potestatis* justified when it was finally measured against the price that had to be paid—notably the corruption spawned by the stringent financial measures (*e.g.*, sale of indulgences, benefices, etc.) needed to support the growing army of clerical bureaucrats at Rome. And on this point one of the things left undone by the Gregorian reformers proved to be crucial. Their failure to uproot the notion of the "proprietary church" explains both the willingness of later canonists to classify the laws governing the disposition of ecclesiastical benefices under the heading not of public but of private law (law pertaining to the protection of proprietary right), and also the tendency of medieval persons in general to regard ecclesiastical office less as a focus of duty than as a source of income or an object of proprietary right. When the 13th-century popes found that direct papal taxation did not yield funds sufficient to support their bureaucrats, they adopted the practice of "providing" them to benefices all over Europe, for the law itself encouraged them to think of such benefices as sources of much needed revenue. Thus arose the characteristic abuses of pluralism (holding more than one benefice) and nonresidence against which church reformers from the mid-13th century on railed in vain. The papacy had finally come to be regarded as an obstacle rather than a spur to reform.

Negative aspects of the 13th-century reforms

The crisis of the late Middle Ages: Boniface VIII to the Council of Trent. *The beginning of the Avignon papacy.* In 1303, despite its resounding claims and its complex governmental machinery, the prestige of the papacy had fallen so low that it was possible for mercenaries in French pay and under French leadership to harass with impunity and humiliate the pope himself, Boniface VIII, at Anagni, where he was arrested in his own family (Caetani) palace.

The aftermath of this "outrage of Anagni" was the "Babylonian Captivity"—the desertion of Rome by the popes and their long residence (1309–77) at Avignon, France. The Avignonese papacy was neither as morally corrupt nor as responsive to French pressure as generally believed. It was overwhelmingly French in complexion, and this, along with its intensification of centralization, did impose severe strains. Serious problems arose when to these strains were added the tensions engendered by the oligarchic ambitions of the cardinals, who became increasingly determined to transform their traditional involvement in the shaping of papal policy into a constitutional right.

The "Babylonian Captivity" at Avignon

The Great Schism and the rise of conciliarism. In 1378, only a year after Gregory XI had brought the papacy back to Rome, a disputed papal election initiated the Great Schism of the West (1378–1417) in which there were first two and then three rival claimants to the papal throne, and the subsequent emergence of the conciliar movement. Demanding reform "in head as well as in members," this fundamentally "constitutionalist" movement took its point of departure from the belief that the fullness of jurisdictional power pertained to the whole church and the general councils representing it, as well as to the pope himself. The assumptions basic to conciliarism were by no means novelties, and support was drawn from the commentaries of the medieval canon lawyers themselves. However, neither Urban VI nor Clement VII—the two claimants, Roman and Avignonese, who emerged from the disputed election—nor their respective successors proved capable of commanding the allegiance of the whole of Christendom or of displacing their rivals. It was their prolonged obduracy that finally, in 1409, rallied widespread support behind a conciliar solution. But even then the efforts of the Council of Pisa were unsuccessful and the notorious John XXIII emerged, not as sole pope, but as the representative of a third, or "Pisan," line of papal claimants. Only after the Council of Constance (1414–18) was the schism terminated. In its decree *Sacrosancta* (1415), it had affirmed the superiority of its authority even to that of the pope in matters pertaining to the faith, the ending of schism, and the reform of the church; the council also ended the schism of the church, deposed John XXIII and his Avignonese rival, accepted the resignation of the Roman claimant, and elected Martin V.

The decision of the fathers assembled at Constance to proceed to a papal election before enacting a full-scale program of reforming legislation turned out to be crucial. For despite the passage of the decree *Frequens* (1417), by the terms of which general councils were to assemble at frequent and regular intervals, despite the far-reaching claims and activities of the conciliarist party at the Council of Basel (1431–49), and despite the survival of conciliar theory as a viable theory of church government until well into the 18th century, the Conciliar movement (as such) diminished without effecting its full reform program and without engineering the constitutional revolution in church government that it had promoted.

The effect of nationalism on the papacy. From the early 14th century onward European royalty and nobility had extended their powers over the local churches. By the early 15th century the kings of France and England in particular had become adept at marshalling national anti-papal feeling in order to coerce the papacy to concede to them sizable shares of the taxes levied on their national churches and the benefices belonging thereto. The Avignonese popes and the popes of the schism had little choice but to yield to such diplomatic blackmail—even though, by so doing, they committed the church piecemeal to a revolution that would ultimately leave to their successors little more than a theoretically supreme authority. The substance of power over the national or territorial churches passed into the hands of kings, princes, and rulers of city-states, such as Venice. The concordats that Martin V concluded in 1418 with England and other countries were no substitutes for the reforming program espoused by the conciliarists. The infamous Concordat of 1516, which effectively delivered the French Church into the hands of the monarchy, marked the peak of this development. Its concomitant was the effective transformation of the 15th-century popes into sophisticated Renaissance princes, sometimes of dubious personal morality, always preoccupied with the restoration and preservation of their Italian principality and the consolidation in their own hands of despotic power over it, conducting their diplomatic relations as equals with the secular princes of Europe, and increasingly prone to the type of dynastic ambition current among those princes. That Adrian VI should have been the last of the non-Italian popes is symptomatic of this transformation; so, too, is the apparent inability of Leo X to comprehend the significance of Luther's chal-

The Renaissance papacy

lenge (*e.g.*, that even popes and councils can err) or the strength of the demand for church reform (*i.e.*, in head and members) made by many individuals and groups. Even when the magnitude of the Protestant threat and the pressing need for reformed and reforming papal leadership was finally recognized at Rome, the complexities of the European diplomatic situation in which the papacy was so deeply involved helped delay, until the greater part of northern Europe was lost to Catholicism as a result of the Protestant Reformation, the convocation of the urgently needed and long-awaited reforming general council.

The papacy in the modern world: Trent to the first Vatican Council. The convocation finally came during the pontificate of Paul III (reigned 1534–49), who had made reform a concern of the papacy itself. He gave his approval to the new Jesuit order that was destined to be so loyal to the papacy, appointed distinguished reformers to the College of Cardinals, established a reform commission, and embarked upon the reform of the Roman Curia itself.

Council of Trent

The opening of the Council of Trent in 1545 marked no weakening in the diplomatic and national pressures with which the late medieval papacy had increasingly been forced to accommodate itself. The council's long duration and the suspensions and reconvocations that punctuated its course reflect the incessant working of these pressures. Though the Tridentine (Council of Trent) popes were able by their diplomatic skill and their tenacity in the pursuit of reform to prevent accommodations with Protestant and quasi-Protestant demands and to preserve the uniformity of Catholic religious practice, their post-Tridentine successors were reduced progressively to political impotence because of increased autonomy in the local churches and increasing monarchical control of the national clergies.

The failure of the ecclesiastical sanctions that Paul V (reigned 1605–21) imposed on the Republic of Venice because of its violation of ecclesiastical rights pointed to the humiliations to be suffered by the papacy in the 18th century. Nationalism in the form of Gallicanism, Febronianism, and Josephism (French, German, and Holy Roman imperial doctrines, respectively, advocating the restriction of papal power) progressively isolated the pope, even in the Catholic world. Clement XIV's capitulation to Bourbon (French monarchical dynasty) pressure in 1773, when he decreed the suppression of the Jesuits, and the humiliation and imprisonment of Pius VI by the French revolutionaries and Pius VII by Napoleon Bonaparte appeared to threaten the very survival of the papal office.

Recovery of papal prestige. Following the Napoleonic Wars and the Congress of Vienna (1815), however, both the reaction to monarchism and the subsequent rise of constitutional regimes served, in different ways, to promote the recovery of the papacy. The reinstated monarchs of Catholic Europe did not share their predecessors' antipathy to the papacy. Instead, they viewed it as a valuable ally in the struggle to preserve the status quo. The later shift to constitutional regimes, weakening or severing, as it did, the bonds that had chained the clergy to the policies of their national governments, freed them to respond positively to a more vigorous and extensive exercise of papal authority.

19th-century developments

The restoration of the Papal States by the Congress of Vienna and the shape assumed by Catholicism in the wake of the Council of Trent helped to shape 19th- and 20th-century papal policy. The former disposed the 19th-century popes to align themselves with the politically conservative forces of Europe, even if this meant incurring the odium of the liberal and modernizing forces of the era. The latter disposed them to see the papacy as the hinge on which turned a complete and definitively articulated system of beliefs, clearly defined in opposition to the claims of Protestantism and increasingly and self-consciously at odds with most of the main, and quite a few of the tributary, currents of modern thought. During the early sessions, the fathers assembled at Trent had shown a willingness to understand reforms as a renewal that should go some way toward meeting legitimate Protestant demands, but during its later sessions and in the years after its closure, reform increasingly was conceived as the restoration of a purified (real or imagined) medieval norm. Thus, from the later 16th century onward the popes tended to interpret the piecemeal doctrinal and disciplinary decrees of Trent in a new and unintended way—as a closed and complete system and an "ultimate rule of faith and discipline." Ecclesiastical usages that predated Trent were progressively eliminated, so that even the slender links with Christian antiquity that Gratian's *Decretum* (*c.* 1140) had preserved in the body of the medieval canon law were now snapped.

In such a context, the popes of the modern era, despite their political weakness, have made much of their own spiritual authority and especially their teaching authority. With the growth of papal prestige after 1815, this trend remained the most persistent feature of papal history down to the mid-20th century. Pius IX identified the papacy with the Ultramontane position, which favoured an intensified centralization of ecclesiastical government in Rome and a vigorous exercise of papal authority in every aspect of religious life. But even before his pontificate, centralization was being strongly encouraged. Gregory XVI achieved—for the first time in history—complete papal control over Catholic missionary activity throughout the world. This, along with rapid advances in the technology of communication, promoted that aspect of papal domination that has characterized the Roman Catholic Church in the 20th century.

From the first Vatican Council to the second Vatican Council. After 1870, when the Papal States had been forcibly annexed to the new Kingdom of Italy and papal temporal power destroyed, the popes were free to exploit their growing spiritual authority within the church. They were also able to draw out the logical consequences of the definitions of papal primacy and infallibility promulgated by the first Vatican Council (1869–70), which was forced to a premature adjournment just before the entry of Italian troops into Rome.

Papal primacy and infallibility

While the definition of infallibility (ascribing inerrancy only to papal pronouncements made solemnly *ex cathedra* on matters pertaining to faith or morals) was a much more restricted one than many Ultramontanists (upholders of the doctrine of papal supremacy and power) had desired, it had the effect of surrounding with an aura of infallibility even papal pronouncements that laid no formal claim to it. Along with its companion definition of the papal primacy it took on an exaggerated importance—partly because it was promulgated in isolation from any general decree on the structure of the church at large and the role of the episcopate within it, and partly because of the very distinction of Pius IX's successors, notably Leo XIII. The years between Vatican I and Vatican II were marked by an increasingly energetic exercise of papal spiritual authority, both magisterial and jurisdictional. Never before had popes been so active in doctrinal and moral teaching, and the great encyclicals of Leo XIII and Pius XII, especially, became determinative in shaping the development of Catholic thinking on an imposing range of topics. At the same time, assisted by the promulgation in 1917 of the Code of Canon Law, by improvements in the technology of communication, and by the growth of the missionary churches and of native clergies deeply loyal to Rome, the process of administrative centralization was intensified to an unprecedented degree. Similarly, the rectification by the Lateran pacts of 1929 of the relationship between the Vatican and the Italian state enhanced the papacy's standing in the diplomatic world and restored to it a measure of temporal independence that was something more than symbolic. Rarely in its checkered history had the papacy seemed in a more secure position than it did in the 1950s.

But there was another side to all this. Leo XIII had done much to reverse the uncompromising hostility to modern thought symbolized so dramatically by Pius IX's *Syllabus of Errors* (1864), which condemned socialism, Bible societies, liberalism, modern scientific thought, and other 19th-century liberal views and movements. By the end of the century theologians were seeking to reshape Catholic teaching in such a way as to bring it into harmony with the dominant intellectual developments of the era—espe-

cially in philosophy, natural science, history, and biblical studies. Even Pius X's condemnation of Modernism in 1907 and the control over Catholic thought subsequently exerted by the Holy Office could not utterly destroy this process. Despite the ambivalence of Pius XII's attitude toward this rapprochement, it gained considerable momentum during his pontificate, and its unexpected strength was soon revealed when, in January 1959, to the consternation of his curial officials, the new pope, John XXIII, set in motion the machinery that led to the assembly of the second Vatican Council in October 1962.

Second Vatican Council

The contemporary papacy. The decrees of Vatican II —such as the great documents on the church, the sacred liturgy, ecumenism, and Religious liberty—attest to the degree to which its convocation liberated reforming energies and stimulated the forces of theological innovation already active in the church. Similarly, the doctrinal dissension and disciplinary turmoil increasingly characteristic of Catholic life in the late 1960s and early '70s also attest to the fact that the ending of the council in 1965 did not stifle those energies or curb those forces.

At the heart of the current controversy stands the papacy. The 1960s witnessed a broadening of the "progressive" theological vision. Paul VI's attitude to this vision apparently changed from an initial neutrality to an increasing conservatism. While the council was in session, he sought on several occasions to allay the fears of the conservative minority by such actions as withdrawing from conciliar discussion and reserving to himself the questions of clerical celibacy and family limitation; he also reiterated in a theological terminology redolent rather of Trent than of Vatican II the traditional Catholic teaching on the Eucharist (*Mysterium Fidei*, September 1965), which reaffirmed the doctrine of transubstantiation. After the end of the council his public pronouncements repeatedly reflected concern about the rapid spread of heterodox opinions among Catholics, about the increasing disrespect of some church members for ecclesiastical authority in general and papal authority in particular, and about threats to church unity. His reaffirmation of the traditional ban on the use of artificial methods of birth control (*Humanae Vitae*, July 1968) served only to exacerbate dissent and to intensify existing questioning about the fundamental status of papal authority and the mode of its current exercise.

These developments have obscured the great degree to which Paul VI continued the efforts of John XXIII both to promote the growth of Christian unity and to make of the papacy a moral and humanitarian force capable of appealing to the goodwill of all men, whatever their religion. They have also raised questions about the degree to which episcopal collegiality is manifesting itself in practice—despite the 1967, 1969, and 1970 meetings of the Synod of Bishops that Paul VI had himself established in 1965 in order to provide for it an institutional expression. Papal prestige had rarely stood higher than it did in 1963 at the time of John XXIII's death, but the papacy's fortunes have since undergone decline. From being a symbol of unity and hope, it has become, for some Catholics, a focus of dissension and apprehension.

THE THEORY OF PAPAL AUTHORITY

The title deeds: the Petrine theory. Basic to the claim of the Roman bishops to a position of primacy in the church is the Petrine theory, according to which Christ, during his lifetime, promised the primacy to Peter alone, and, after his Resurrection, actually conferred that role upon him. Thus John 1:42 and, especially, Matt. 16:18 f.: "And I tell you, you are Peter, and on this rock I will build my church, and the powers of death shall not prevail against it. I will give you the keys of the kingdom of heaven, and whatever you bind on earth shall be bound in heaven, and whatever you loose on earth shall be loosed in heaven." Also John 21:15 f.: "Feed my lambs . . . Tend my sheep." Vatican I, in defining the Petrine primacy, cited these three texts, interpreting them to signify that Christ himself directly established St. Peter as prince of the Apostles and visible head of the Church Militant, bestowing on him a primacy not merely of honour but of true jurisdiction. In defining also that the Petrine primacy was, by Christ's establishment, to pass in perpetuity to his successors and that the bishops of Rome were these successors, Vatican I cited no further scriptural texts. In defining further, however, that the Roman pontiffs, as successors in the Petrine primacy, possess the authority to issue infallible pronouncements in matters of faith or morals, the council cited both Matt. 16:18 f. and Christ's promise to Peter at the Last Supper: "But I have prayed for you that your faith may not fail; and when you have turned again, strengthen your brethren" (Luke 22:32).

Primacy of Peter

Historical conceptions of papal authority within the church. Of the Petrine texts, Matt. 16:18 f. is clearly central and has the distinction of being the first scriptural text invoked to support the primatial claims of the Roman bishops. Before the mid-3rd century, however, and even after that date, some Western, as well as Eastern, patristic exegetes (Early Church Fathers who in their interpretation of the Bible used critical techniques) understood that by the "rock" Christ meant to refer not to Peter but to Christ himself or to the faith that Peter professed. Nevertheless, in the late 4th and 5th centuries there was an increasing tendency on the part of the Roman bishops to justify scripturally and to formulate in theoretical terms the ill-defined pre-eminence in the universal church that had long been attached to the Roman Church and to its bishop. Thus, Damasus I, despite the existence of other churches of apostolic foundation, began to call the Roman church "*the* apostolic see." About the same time the categories of the Roman law were borrowed to explicate and formulate the prerogatives of the Roman bishop. The process of theoretical elaboration reached a culmination in the views of Leo I and Gelasius I, the former understanding himself not simply as Peter's successor but also as his representative, or vicar. He was Peter's "unworthy heir," possessing by analogy with the Roman law of inheritance the full powers Peter himself had wielded, which he interpreted as monarchical, since Peter had been endowed with the *principatus* over the church.

The vicar of Peter

Medieval views. On the purely theoretical level the distance between the claims advanced by Leo I and the position embodied in Vatican I's primacy decree is not great. Medieval popes, such as Gregory VII, Innocent III, and Innocent IV, clarified by their practice as well as by their theoretical statements the precise meaning of that fullness of power (*plenitudo potestatis*) over the church to which, according to some scholars, Leo I himself had laid claim. In this they were aided not only by the efforts of publicists such as the Italian theologian and philosopher Aegidius Romanus (died 1316), who magnified the pope's monarchical powers in unrestrained and secular terms, but also by the massive development during the late 11th, 12th, and 13th centuries of a highly romanized canon law. Gratian's *Decretum* (*c.* 1140), the unofficial collection of canons that became *the* fundamental textbook for the medieval student of canon law, laid great emphasis on the primacy of the Roman see, accepting as genuine certain canons that were the work of 6th- and 9th-century forgers—such as two principles that the 1917 Code of Canon Law restates: "that there cannot be an ecumenical council which is not convoked by the Roman Pontiff" and that "the First See is under the judgment of nobody."

The prevalence of such ideas and the absence of a formidable challenge to papal primatial claims during the High Middle Ages explains the lack of any conciliar definition of the Roman primacy at the great "papal" general councils of that period. Hence it took the (abortive) attempt at reunion with the Orthodox Church at the Council of Florence in 1439 to evoke the first solemn conciliar definition of the Roman primacy. This definition was included in the decree of union with the Greeks (*Laetentur Coeli*), and it went as follows:

The first conciliar definition of Roman primacy

> . . . We define that the Holy Apostolic See and the Roman Pontiff hold the primacy over the whole world, that the Roman Pontiff himself is the successor of Peter, prince of the Apostles, that he is the true vicar of Christ, head of the whole

church, father and teacher of all Christians, and [we define] that to him in [the person] of Peter was given by our Lord Jesus Christ the full power of nourishing, ruling and governing the universal church; as it is also contained in the acts of the ecumenical councils and in the holy canons.

Early modern and modern views. This decree was the basis for the solemn definition that Vatican I promulgated in 1870 as part of its dogmatic constitution (*Pastor Aeternus*). Having asserted as a matter of faith the primacy of Peter and the succession of the popes in that primacy, and having quoted in full the Florentine definition, the constitution clarified what is to be understood by "the full power of nourishing, ruling, and governing" the church, which, according to that definition, inhered in the pope's primacy. Unlike Florence, *Pastor Aeternus* specified this to include the pope's judicial supremacy, insisting that there is "no higher authority," not even an ecumenical council, to which appeal can be made from a papal judgment.

This definition marked the culmination of a development reaching back at least to the 4th and 5th centuries. But the doctrinal development that culminated in Vatican I's definition of papal infallibility cannot lay claim to a comparable antiquity. There has always been much discussion about the meaning of the prerogative of infallibility and what it implies about the status of individual doctrinal pronouncements of the church's teaching authority. The notion that the church (conceived as the community of the the faithful) is by virtue of Christ's own promise infallible—in the sense that it cannot totally deviate from the truth—is clearly scriptural in foundation and was not questioned even by the Protestant Reformers. Similarly, the notion that a pre-eminent authority attached to the doctrinal pronouncements of the Roman Church and its bishops was of great antiquity, long predating the extension of papal jurisdictional claims by the 4th- and 5th-century popes. But the combination of these two notions—*i.e.*, the identification of the supreme teaching authority of the universal church with that of the pope, and the claim that the infallibility promised to the church itself was possessed also by the pope acting as its head to guarantee the inerrancy even of his individual doctrinal pronouncements—is essentially a modern theological development and one characteristic primarily of the Roman or Ultramontane (pro-papal) theological school. This school rose to prominence in the 16th and 17th centuries; one of its most distinguished representatives was Cardinal Robert Bellarmine (died 1621). Though it drew from earlier materials—notably from the Pseudo-Isidorian decretals and from the writings of such medieval theologians as St. Thomas Aquinas (died 1274), Aegidius Romanus (died 1316), and Augustinus Triumphus (died 1328)—the Ultramontane school derived much of its initial strength from the papalist reaction that followed in the wake of the conciliar movement, and it was shaped very much in opposition to the claims that the conciliarists and their Gallican successors made on behalf of the general council. This is evident in the solemn definition of the doctrine promulgated by Vatican I, with its insistence that the *ex cathedra* definitions of the pope (those made from "the chair," or papal throne), "are irreformable of themselves and not by virtue of the consent of the Church." The conciliar debates indicate that this sentence was intended to exclude the Gallican notion that a papal definition could not claim infallibility unless, subsequently or concomitantly, it received episcopal assent. Despite the maximalist (extremist) tendencies both of subsequent Catholic apologists and of their Protestant critics, the sentence apparently was not intended to restrict the church's infallible teaching authority to the pope alone or to suggest that the pope was free to define doctrine without making every effort to take into account the mind of the church.

Ultramontanism

Nevertheless, after 1870, when the memory of the heated conciliar debates had faded away, maximalist interpretations became prominent. In particular, there was a marked tendency to stress the absolute and unlimited nature of papal jurisdictional power and to end in favour of the papacy the hitherto unresolved question of the source of episcopal jurisdiction. In response to this development, Vatican II, in its dogmatic constitution, *De Ecctesia* (1964), while endorsing Vatican I's teaching on papal primacy and infallibility, also focussed on the nature of episcopal authority. It insisted that bishops "are not to be regarded as vicars of the Roman Pontiff, for they exercise an authority which is proper to them," since, "by divine institution . . . [they] . . . have succeeded to the place of the apostles as shepherds of the Church" and are themselves, in fact, "the vicars and ambassadors of Christ." Also, "Just as, by the Lord's will, St. Peter and the other apostles constituted one apostolic college, so in a similar way the Roman Pontiff as the successor of Peter and the bishops as the successors of the apostles are joined together." This college, "together with its head, the Roman Pontiff, and never without this head" is "the subject of supreme and full power over the universal Church," a supreme authority that it can exercise in more than one fashion but "in a solemn way through an ecumenical council." The supreme authority in the church can be exercised not only personally by the pope himself but also in a collegial fashion by the *whole* episcopate, which of necessity includes the bishop of Rome as its head.

Episcopal collegiality

In so emphasizing the doctrine of episcopal collegiality, Vatican II was responding to the findings of modern New Testament and patristic scholarship concerning the nature of the primitive and ancient church, and it insisted that it was restoring an ecclesiological emphasis of great antiquity. Recent medieval scholarship indicates that this emphasis persisted into the Middle Ages and survived, in the writings of canonists and theologians, side by side with the more prominent concern with the papal primacy. The great Conciliarists active at the Council of Constance made an unsuccessful attempt to effect a stable balance between these two emphases, and even in the modern period, despite the growing prominence of Ultramontane views and their eventual triumph at Vatican I, the collegial concern was never fully displaced. It was not lost sight of by the Gallican theologians of the 16th, 17th, and 18th centuries who, however much their subservience to the exigencies of royal policy may have damaged their credibility, apparently are now recovering in Catholic eyes, at least, a certain measure of esteem.

Eastern Orthodox views and critiques. The recovery of this ecclesiological emphasis has an importance outside Roman Catholic theology. It has never ceased to dominate in the churches of Eastern Orthodoxy, with their stress on episcopal equality, their respect for the autonomy of the national or regional churches, and their insistence that the supreme teaching authority in the universal church resides (if anywhere) in the collegial decisions of the bishops assembled together in an ecumenical council. Up to the 11th century, Byzantine churchmen and theologians certainly accorded some sort of primacy to the church of Rome and its bishop. But with the growth of papal claims to a universal jurisdictional power, with the growing conviction that the Roman Church had fallen into heresy, and above all with the disastrous crusading onslaught on Byzantium in 1204, the attitude of Orthodox churchmen to Rome underwent an understandable shift. Though Byzantine theologians rarely questioned the fact of Peter's primacy among the Apostles, they concluded that their own fundamentally collegial ecclesiology necessitated the rejection of the primatial claims advanced on behalf of those who claimed to be uniquely his successors. The very attempt by the bishops of a single local church to claim a monopoly on the Petrine succession was regarded as something of a deviation, in that all bishops, insofar as they professed the faith of Peter, were to be understood as his successors.

Rejection of Roman primatial claims

In the modern period, then, the Eastern Orthodox churches have been unanimously adamant in their rejection of the papal claims to primacy and infallibility. Orthodox theologians are often careful to insist that what they are rejecting is not the notion of primacy itself but rather that actual primacy of jurisdiction as it was conceived in the Latin Middle Ages and as it has been exercised by Rome in the modern period—with its apparent corollary that all power in the church is to be regarded as proceeding outward from the primatial office and its concomitant tendency to stifle independent life in the

local churches. The original primacy of honour, which these theologians argue, was one accorded to the Roman bishops by emperors and ecumenical councils, they clearly regard as a different matter altogether. Given this fact, and also the common ground shared in ecclesiological matters by the Roman Catholic and Orthodox churches, Vatican II's affirmation of episcopal collegiality may soften the edges of the Orthodox rejection of the papal primacy.

Protestant views and critiques. The impact of that doctrine on Protestant thinking is more difficult to predict. Historically, the Protestant rejection of papal claims has been much less qualified than that of the Orthodox. Thus, in the modern theologian Karl Barth's view, Vatican I's definition of papal infallibility completed the process by which the Roman Catholic Church abandoned the Christian belief in the unique character of divine revelation, identified itself instead with that revelation, and made the pope's teaching "the infallible revelation for the present age." Philipp Melanchthon, the Lutheran author of the Augsburg confession of 1530, may have been willing to admit that a truly evangelical pope had a certain superiority over other bishops; however, even if one de-emphasizes Luther's denunciations of the pope as the Antichrist, the rejection by the major Reformers and their successors of the Petrine theory and of papal primacy by divine institution is absolute. Peter, they argued, exercised no primacy. The powers communicated to him were the powers communicated to all the Apostles. By the "rock" Christ meant himself; upon him the church is founded, and in Matt. 16:17 f. Peter stands only as the type or figure of the Christian faithful who believe in Christ as the sole "rock."

Unlike such medieval predecessors as the Waldensians (Italian reform group founded in the 12th century) or the political philospher Marsilius of Padua (died *c.* 1342), who had likewise attacked the Petrine theory, the Reformers did not base their attack upon the historical argument that Peter had never visited Rome. This argument was embraced by many a liberal Protestant theologian of the 19th century, but in the 20th it has lost most of its appeal.

Modern Protestant reinterpretation of Roman primatial claims

In the mid-20th century some Protestant theologians shifted toward the Roman Catholic understanding of the status and meaning of Matt. 16:17 f. According to Oscar Cullmann (the French Protestant biblical critic and theologian), any sound exegesis of the relevant scriptural texts points to the conclusion that Peter enjoyed a pre-eminence among the disciples even during Christ's lifetime; that the "rock" in Matthew refers not to Christ or to the faith of Peter but to his person; that Christ promised him, therefore, the leadership of the church; and that after the Resurrection Peter actually exercised that leadership. Though Cullmann argued that Peter did so only for a short time, being replaced in the leadership by James, other Protestant scholars have disagreed and have claimed for Peter a more enduring role. All, however, continue with Cullmann to distinguish sharply between conditions in the apostolic and post-apostolic church, to deny on exegetical grounds that the "primacy," or leadership, promised to Peter was intended to be passed on to any post-apostolic successors, and to insist on historical grounds that no such succession in the primacy actually occurred in the primitive church. Nevertheless, because of the degree of convergence already occurring between the Roman Catholic and Protestant exegesis of the Petrine texts, because of the re-examination of the Catholic tradition begun by Vatican II, and because of the growing Protestant sense of the need for some striking symbol of unity in the worldwide Christian community, some Protestant ecumenists in the last third of the 20th century have shown a degree of openness to the papal office that would have been unimaginable only 50 or 60 years before.

Historical conceptions of the relationship of the papacy to the world. Theories concerning the relationship of the papacy to the world at large have both reflected the established political conceptions of the day and been in tension with them. The pope has been conceived successively as a leading dignitary in an imperial church headed in effect by the emperor, as a majestic potentate possessed of a supreme and direct authority even in temporal matters, and as a primarily spiritual figure who had in temporal matters no more than an indirect power of intervention. With the post-Reformation fragmentation of Christendom, the growth of secularism, and the emergence of the unified modern state claiming within its own borders jurisdictional omnicompetence, even such attenuated claims to an indirect power became increasingly anachronistic. In the 20th century, in his relations with the world at large, the pope, while affected by the conventions regulating the relationships of heads of state with one another, possesses primarily a moral authority deriving from the dignity and prestige of his office. The strength of that authority, however, depends upon his moral standing as a person, upon the persuasive force of his cause, and upon the degree of enthusiasm it can arouse within the church.

Contemporary teaching on papal authority. After the mid-20th century, some voices were raised in Roman Catholic circles questioning both the doctrine of papal infallibility and the exercise of the papal primacy—at least as it is envisaged in the teaching of Vatican I and of the Code of Canon Law. The church's official teaching on the papal office remains that of Vatican I, solemnly reaffirmed at Vatican II. Nevertheless, the latter council's juxtaposition of the doctrine of episcopal collegiality with the existing teaching on papal primacy and infallibility created something of a dilemma in Catholic ecclesiology. Though the text of *De Ecclesia* had insisted that the doctrine of episcopal collegiality in no way impugned the pope's primacy, a minority of the council fathers remained unconvinced and were commonly said to have been won over by the explanatory note that the Theological Commission by papal authority appended to the decree as an "authentic norm of interpretation." The note is framed in much more juristic terms than is the decree itself, and, in discussing the possession by the College of Bishops of "supreme and full power over the whole Church" it insists that "there is no distinction between the Roman Pontiff and the bishops taken collectively," that "necessarily and always, the College carries with it the idea of its head" so that the bishops acting independently of the pope cannot be considered to constitute a college. At the same time, the note insists that "since the Supreme Pontiff is the head of the College, he alone can perform certain acts which in no wise belong to the bishops, for example, convoking and directing the College, approving the norms of action etc.," norms that "must always be observed."

Recent questioning of the doctrines of papal infallibility and primacy

Already in 1964 there were some who regarded this note with considerable misgiving, feeling that it withdrew from the bishops, in practical and legal terms, that supreme authority in which they had been said, on theological grounds, to be sharers. Subsequent events did little to dispel such misgivings. Despite the unquestionable vitality shown at its 1967 and 1969 meetings, the Synod of Bishops was not really allowed to function as a decision-making rather than a merely advisory body, and it was no more consulted than were the bishops as a whole when, in 1968, the pope promulgated *Humanae Vitae* (the encyclical on birth control)—considered by some observers to be the most divisive papal initiative of recent times, and one that amounted to a *de facto* negation of collegiality.

Humanae Vitae

Because of the dissent over *Humanae Vitae* and the tension engendered by the rigour of the pope's stand on the much-debated problem of clerical celibacy, attention probably will focus increasingly on the old and difficult question of the limits of papal power. Because of this, considerable importance attaches to the current revival of interest in the late medieval conciliar movement and to the assertion made by some Roman Catholic scholars (if hotly disputed by others) that a continuing dogmatic validity must be accorded to the decree *Sacrosancta*, promulgated in 1415 by the Council of Constance. This decree declared that the general council possessed an authority superior to that of the pope in matters pertaining to the faith, the ending of the schism, and the reform of the church. Those who assert this view do not always wish by so doing to cast any doubt on the dogmatic validity of

Vatican I's teaching on papal primacy and infallibility, but the efforts thus far made to demonstrate the compatibility of the respective teachings of the two councils (*i.e.*, Constance and Vatican I) remain somewhat less than persuasive.

Conclusion. Those who take a gloomy view of the papacy's future because of the current discontent may be reassured somewhat by the enduring vitality the institution has shown in the past and by its astonishing powers of recovery. But the history of past triumphs is no more an adequate guide to future developments than is that of past disasters; and there are some novel components in the current decline in papal fortunes that make predictions about its future unusually difficult. Among these may be singled out the recovery by Roman Catholic biblical scholars and theologians of a fundamentally scriptural and nonpolitical conception of the ecclesiastical office as one of service rather than of dominion, and the growing insistence with which the struggle for Christian unity has impressed itself on the consciences of Catholic and non-Catholic alike. These two developments converge to portray in a critical manner papal primacy as it has traditionally been conceived and continues currently to be exercised, and to underline the important ecumenical role that primacy might be able to fill if its traditional conception were transformed. The emphasis on the "primacy of dominion," which had its origins in the 4th and 5th centuries, was elaborated in the later medieval and modern periods, and has continued to be characteristic of the 20th-century papacy, may yet be transformed in the years immediately ahead.

BIBLIOGRAPHY. For bibliographical information on papal history, see O. CHADWICK, *The History of the Church: A Select Bibliography* (1962). For listings of theological works concerning the Petrine question and the papal office, see E. DUBLANCHY, "Infaillibilité du Pape," *Dictionnaire de théologie catholique*, vol. 7, pp. 1638–1717 (1922); G. GLEZ, "Primauté du Pape," and M. JUGIE, "Primauté dans les églises séparées d'Orient," *ibid.*, vol. 13, pp. 247–391 (1936); and for the more important recent works, H. KUNG, *Structures of the Church* (Eng. trans., 1964).

Historical: Of the general histories, J. HALLER, *Das Papsttum: Idee und Wirklichkeit*, 2nd ed., 5 vol. (1951–53), is something of a classic. Useful largely for reference are the following detailed accounts: H.K. MANN, *The Lives of the Popes in the Middle Ages*, 18 vol. (1902–32), which goes to 1304 and is continued for the period up to 1800 by L. VON PASTOR, *The History of the Popes from the Close of the Middle Ages*, 40 vol. (Eng. trans., 1891–1953). J. SCHMIDLIN, *Papstgeschichte der neuesten Zeit*, 4 vol. (1933–39), discusses the history to 1939. The standard collection of documents is C. MIRBT, *Quellen zur Geschichte des Papsttums*, 4th ed. (1924). There is much on the papacy in general church histories. The most complete of recent scholarly accounts is A. FLICHE and V. MARTIN (eds.), *Histoire de l'Église depuis les origines jusqu'à nos jours*, 24 vol. (1934 *et seq.*). The most valuable account currently appearing in English is H. JEDIN and J. DOLAN (eds.), *Handbook of Church History*, 6 vol. (Eng. trans., 1965 *et seq.*). An excellent brief introduction to papal history up to the Reformation, including a good bibliography, is G. BARRACLOUGH, *The Medieval Papacy* (1968). For the recent shift in Roman Catholic discussion of the early evidence concerning the Roman primacy, see J.F. MCCUE, "The Roman Primacy in the Second Century and the Problem of the Development of Dogma," *Theological Studies*, 25:161–191 (1964); and for the development of medieval papal claims, W. ULLMANN, *The Growth of Papal Government in the Middle Ages*, 2nd ed. (1962). For the subsequent crisis of papal authority, see F. OAKLEY, *Council Over Pope? Towards a Provisional Ecclesiology* (1969), which contains a brief history of conciliar theory and discusses much of the recent literature on the significance of the Council of Constance. E.C. BUTLER, *The Vatican Council, 1869–70*, new ed. (1962), is a satisfactory history of the first Vatican Council; reference should also be made to R. AUBERT, "L'Ecclésiologie au concile de Vatican," in B. BOTTE *et al.*, *Le Concile et les conciles: contribution à l'histoire de la vie conciliaire de l'Église* (1960). The essays gathered together in this last volume constitute a good guide to recent studies in the history of the relationship of pope to general council. W.M. ABBOTT (ed.), *The Documents of Vatican II* (1966), is an introduction to the achievement of that council.

Theological: O. CULLMANN, *Peter: Disciple, Apostle, Martyr* (Eng. trans., 1953); and O. KARRER, *Peter and the Church: An Examination of Cullmann's Thesis* (Eng. trans., 1963), are influential statements of modern Protestant and Roman Catholic thinking on the Petrine question. K. RAHNER and J. RATZINGER, *The Episcopate and the Primacy* (Eng. trans., 1962), is a fundamental analysis of the pope-bishop relationship; H. KUNG, *Infallible? An Inquiry* (Eng. trans., 1971), *The Church* (Eng. trans., 1967), and *Structures of the Church* (Eng. trans., 1964), are basic to any understanding of contemporary "liberal" Roman Catholic thinking on the papacy. For Eastern Orthodox views on the papal primacy, see F. DVORNIK, *Byzantium and the Roman Primacy* (Eng. trans., 1966); and N. AFANASSIEF *et al.*, *La Primauté de Pierre dans l'Église Orthodoxe* (1960).

(F.C.O.)

Papaverales

Papaverales, an order of flowering plants in the class Dicotyledoneae, consists of three families: Papaveraceae, Fumariaceae, and Hypecoaceae. The order as a whole contains about 44 genera and 625 species.

GENERAL FEATURES

Most plants making up the order Papaverales are herbaceous (nonwoody), though there are some shrub and small tree species and a few with a climbing growth habit. The order occurs primarily in the temperate regions of the world, extending only sparingly into the tropics. Many different species are utilized as showy garden plants, and the genus *Papaver* (poppy) is the source of edible poppy seed, poppy-seed oil, and opium. The latter is a narcotic long derived from the dried juice that exudes from the cut, unripe capsules of the opium poppy (*Papaver somniferum*). A number of substances, including morphine, codeine, heroin, and other derivatives of opium, obtained from the opium poppy are in wide medical use for the relief of pain, and many are involved in drug addiction problems.

Source of drugs

The family Papaveraceae is one of the important drug-producing families of the plant kingdom, and alkaloids are abundant in many species of the order. Some alkaloids, such as protopine, are widespread within the group but are little known outside of it. Morphine is found only in the opium poppy.

Plants of the order Papaverales do not occupy as many different types of habitats as do many orders of angiosperms (flowering plants), but at the same time they are not restricted to a narrow range of ecological situations. Although there are genera with a marked preference for woodlands and forests, and some groups can be found occupying nearly every kind of terrestrial habitat, the most favoured sites are open and well drained. Habitats range from Arctic and Alpine settings to deserts, moist woods, and stream banks. There are, however, no strictly aquatic members. A number of species are found in deserts, semideserts, and arid areas. The California poppy (*Eschscholzia*), for example, is very abundant in open valley and desert lands of southwestern North America.

Many Papaverales species are weeds and are found in fields and waste places. The great abundance of poppies in the grainfields of Europe and elsewhere is well-known; the weedy tendencies of the prickly poppies (*Argemone*) in the arid areas of the southwestern United States and in Mexico are no less marked. These plants occupy cultivated areas and spread deeply into overgrazed rangelands as well. Celandine, or swallow wort (*Chelidonium majus*), usually occurring in damp or shady places along roadsides, rock walls, and in wastelands, is a ubiquitous weed in Europe, Asia, and eastern North America.

Split distribution pattern

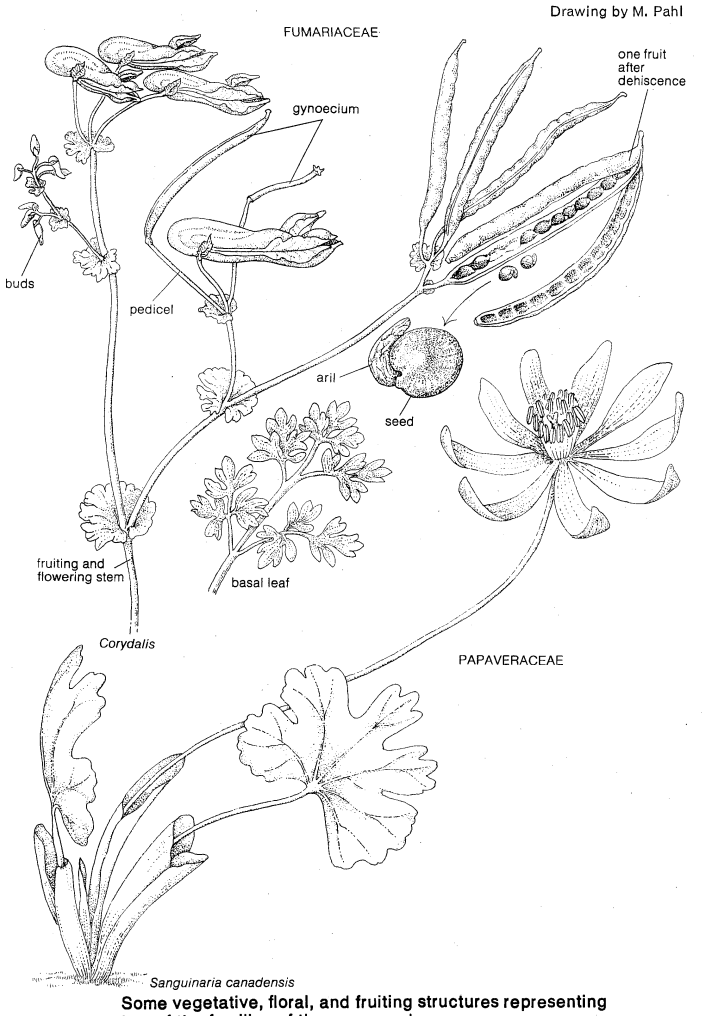
Genera of both the Papaveraceae and Fumariaceae families display the disjunct eastern Asia–eastern North America distribution pattern that was used as a geographical test of evolutionary theory by Charles Darwin and by Asa Gray. The genus *Stylophorum* of the Papaveraceae family, for example, has species found only in eastern North America and Eastern Asia, and the single species of *Adlumia* of the Fumariaceae is found in Korea and the eastern United States with a tremendous gap between.

There is a natural and practical, although not absolute, division of the order Papaverales into those plants in which the sap is watery and those in which it is milky or coloured. In plants belonging to the families Fumariaceae

and Hypecoaceae the sap is watery and acrid, and there is an abbreviated calyx (green leaflike sepals of the flower). These two characteristics serve to distinguish the two families from the Papaveraceae, in the majority of species of which the sap is milky or coloured and the calyx is well developed, fully enclosing the bud until the flower opens.

A fleshy aril (an appendage on the seed) is present, often forming a crest on the seeds of a number of species in what appears to be an adaptation to dispersal by ants. Seeds are carried away from the producing plant by the ants, and the aril is utilized as a source of food, but this does not disturb normal germination.

The widely known poppy flower is the most common flower type within the order. Even dwarf Arctic poppies and the giant matillia poppy, which is woody at the base and otherwise quite different from the common poppy, are quickly identified as poppies by the average person. At a divergent extreme are the flowers of bleeding heart (*Dicentra spectabilis*) and Dutchman's-breeches (*D. cucullaria*), which are laterally compressed and appear to be irregular in shape. Flowers of all members of the order are perfect (*i.e.*, have both pistil and stamens, the female and male flower parts) and hypogynous (the sepals, petals, and stamens attach at the base of the ovary) and have a compound ovary with parietal placentation (*i.e.*, the ovules attach to the ovary wall). An exception to the compound ovary is found in the genus *Platystemon*, in which six to 25 carpels are weakly united when young, but become free in fruit.



Some vegetative, floral, and fruiting structures representing two of the families of the poppy order.

NATURAL HISTORY

Though the breeding system of many species has not been investigated, some species of *Fumaria* and *Corydalis* are known to produce cleistogamous (closed when mature) flowers and to produce fruit without fully opening. Self-pollination is known to occur in the order, but the prevalent mode as a whole is cross-pollination, and insects are the primary carriers of pollen between plants. The open, many stamened flowers of the poppy type are inviting to many kinds of bees and similar insects; the more closed and spurred flowers of many Fumariaceae species provide for a narrower range of insect visitors. No flower type is known to be so insect-specific that only a single species is its pollinator. Even though some flowers are somewhat tubular because of the weak fusion of their petals and are more or less closed by the overlap and fusion of the inner petals, they are dependent on insects for the transfer of pollen from plant to plant to ensure a seed set. Frustrated short-tongued bees sometimes bite through the base of the corolla (flower petals) and gain direct access to the source of nectar without pollinating the flowers, however.

Pollination

There are usually several to many seeds in each fruit, but some genera have single-seeded fruits, and in a few groups the fruits break into single-seeded joints. The seeds are shed from the fruits by the opening of valves, through pores, or by the breaking of the fruit into one-seeded joints. In a few species the valves open explosively, throwing the seeds some distance from the plant. In some, the fruits are one-seeded and indehiscent (do not open along definite seams). Seed dispersal is mostly unspecialized, and seeds are spread around by water, by moving soil, or by the movements of the mature plant after it is uprooted or broken off. Ant dissemination of the arillate seeds is common in certain members of the group.

Evidently, most plants of the order are little eaten by animals, and none is known to have edible vegetative parts. The juices are either acrid and bitter or milky and often narcotic, which apparently is the basis for avoidance of the plants by both animals and man. Poppy seeds, however, are widely used to sprinkle on pastries and other foods or are ground for flour. The narcotic principle is not present in the seeds of even the opium poppy.

FORM AND FUNCTION

Taproots, rhizomes (rootlike stem structures), bulblets, and tubers make up the range of underground structures found in the order Papaverales. The stems are mostly herbaceous but may be woody and are erect or sometimes climbing. Although some leaves are entire (*i.e.*, smooth edged), they are mostly divided into pinnate (*i.e.*, the veins originate along a midrib, as in a feather) or palmately (*i.e.*, the veins radiate from a common point where the leafstalk attaches) arranged lobes or are more highly dissected. A common pattern is a group of petioled leaves arranged in a rosette at the stem base, without stem leaves or with stem leaves reduced from base to apex. The fern-shaped leaves of *Pteridophyllum* are unusual.

The inflorescences (arrangements of flowers) are diverse in form, ranging from a single scapose (stemmed) flower to a cymose or racemose (*i.e.*, clustered along a central axis on short lateral stalks, the cymose type maturing from the terminal blossoms downward, the racemose type maturing from the base upward) inflorescence in a terminal or lateral position. The flowers are of a hypogynous type or, rarely, perigynous (*i.e.*, the sepals and petals arise from a cuplike structure partially enclosing the ovary) and are fully bisexual. They are radially or bilaterally symmetrical and are complete (*i.e.*, they have all four basic flower parts, sepals, petals, stamens, and pistil), even though the two sepals present in the family Fumariaceae are much reduced and often drop early in development. In the family Papaveraceae the two to three sepals tightly enclose the buds until the flower opens, when they are shed. The sepals are usually free except in the genus *Eschscholzia*. The petals are similar to each other, in two whorls, and their number is twice as many as the sepals. In the family Fumariaceae the petals differ from each other, the four petals forming two pairs, the outer pair being somewhat saccate, or spurred, at the base.

The androecium (male complement) is made up of six stamens in two groups in the Fumariaceae family, but

Male flower parts

there are many stamens arranged in concentric whorls in the various species of Papaveraceae. *Pteridophyllum* is a special case with free petals and four stamens. Stamen development takes place centripetally (*i.e.*, from the flower margin toward the centre) which contrasts with the centrifugally developed stamens of the order Capparales, an order once held to be closely related to the Papaverales.

The styles (the upper part of the pistil), if present, may be either deciduous or persistent and sometimes become beaklike on the ultimate fruit. In many genera the style is virtually absent, and the stigmatic areas (pollen-receiving regions) are located directly upon the ovary. There is an unusually wide range of stigma types among various groups of the order Papaverales. The entire, or bilobed, type is common, but greatly expanded, recurved, crested or lance-shaped stigmas are present in *Glaucium*, among other genera, and elaborate stigmatic ridges and crests occur in a number of genera including *Cathcartia, Meconopsis, Argemone,* and *Papaver*. The basic gynoecial (female) structure is constructed of two or more fused carpels. In the family Fumariaceae, there are consistently two carpels, and the ultimate fruit produced is unilocular (*i.e.*, has only one chamber). On the other hand, the gynoecium of the family Papaveraceae, while having two or more carpels and usually being unilocular, in some taxa becomes multilocular in fruit. The fruit is two to many valved (*i.e.*, it has openings through which seeds are released).

Seeds of the order are most commonly kidney shaped, obovoid, or spherical and are frequently pitted. In a number of genera, crestlike arils are present on the seed, but true seed wings are absent. The seeds are plump to somewhat flattened and contain an endosperm (nutrient tissue for the developing embryo) that is usually oily, in addition to the small, straight to somewhat curved embryo.

Plants of the order Papaverales are widely recognized for their medicinal and poisonous properties, and more than 200 alkaloids are known in the order. Most of these are found elsewhere among the higher plants, sometimes commonly so, but in some instances only in one or a few families. At least eight alkaloids are distinctive and are known only from this order of plants.

EVOLUTION

Recorded fossils placed definitely in the order Papaverales are few and contribute little of significance to an understanding of the probable origin of the group. Based on evidence from present-day species, it is now agreed by the most recent authorities that the order is more nearly related to the order Ranunculales than to Capparales, as had been thought by earlier investigators. Within the Papaverales, the family Papaveraceae has more archaic features than the family Fumariaceae, and the Hypecoaceae family has been suggested as forming a connecting link between them.

CLASSIFICATION

Distinguishing taxonomic features. The family Hypecoaceae, with but one genus, *Hypecoum*, is distinguished by free petals and four free stamens, except that these features are also shared by the controversial monotypic (one species) genus, *Pteridophyllum*. The latter has been the basis for the family Pteridophyllaceae, but it is sometimes placed in the family Papaveraceae and at other times in the Fumariaceae. The genera *Hypecoum* and *Pteridophyllum* are distinctive, and a recent treatment gives subfamily rank to each in the family Fumariaceae.

Distinction between the two main families

The families Papaveraceae and Fumariaceae differ substantially in many characters, but the surest distinguishing features are the many regularly inserted stamens, large bud-enclosing sepals, and regular flowers of the Papaveraceae, as contrasted with the six stamens in two groups, small, scalelike sepals, the bilaterally symmetrical to irregular flowers of the Fumariaceae. Although some authorities recognize the Hypecoaceae as a family distinct from Fumariaceae and only questioningly include *Pteridophyllum* in Papaveraceae, others give subfamily rank to both *Hypecoum* and *Pteridophyllum* in the family Fumariaceae. Another possibility sometimes suggested is to place both of these genera in the family Hypecoaceae.

Alkaloid distribution supports the association of the families Papaveraceae and Fumariaceae together in the same order. Protopine, for example, is found in all members of both families so far investigated but is almost unknown elsewhere, having been reported from only one species, *Nandina domestica*, of the family Nandinaceae (order Ranunculales). The subordinate taxa (subfamilies, genera, species, etc.) of Papaveraceae mostly have their own combination of alkaloids, and different combinations appear to characterize different sections of the family. On the other hand, members of the family Fumariaceae possess a combination of isoquinoline derivatives different from anything found in Papaveraceae.

Annotated classification.

ORDER PAPAVERALES

Mostly fleshy herbs, some shrubs or small trees; sap milky (white to yellow or reddish-orange) or colourless and watery; leaves without stipules, basal or alternate, rarely opposite, flowers regular or bilaterally symmetrical; calyx falling off very early; petals 4–12, often falling early; ovary superior, of fused carpels; fruit a capsule or of 1–seeded indehiscent joints; seeds with small embryos and a fleshy, usually oily endosperm. Three families, 44 genera, and about 625 species, distributed mainly in temperate regions.

Family Papaveraceae

Herbaceous annuals or perennials, rarely shrubs or small trees; sap milky, white to yellow or reddish-orange, rarely colourless and watery; leaves alternate (rarely whorled) without stipules, entire to pinnately or palmately lobed; flowers regular, bisexual, usually solitary; sepals 2, 3, or 4, petals twice the number of sepals; stamens mostly numerous; gynoecium of 2 to several carpels; seeds usually numerous. Twenty-five genera and about 200 species.

Subfamily *Chelidonioideae*

Trichomes (plant hairs) multicellular, uniseriate (of a single row of cells); perianth (petals and sepals) 3-merous (present in groups of 3 or multiples thereof); pollen tricolpate (with 3 germination furrows) or polyporate; gynoecium 2-valved; seeds arillate. Ten genera, mostly of Asia.

Subfamily *Papaveroideae*

Trichomes multicellular, multiseriate; perianth 2- to 3-merous; pollen tricolpate; gynoecium 3-valved, but in *Papaver* with added pseudomedian traces (vascular tissue) and false partitions; seeds without an aril. Seven genera, distributed in the northern hemisphere.

Subfamily *Eschscholzioideae*

Trichomes unicellular or the plants glabrous (smooth); perianth 2-merous; pollen polycolpate; gynoecium 2-valved; seeds usually without arils. Three genera of western North America.

Subfamily *Platystemonoideae*

Trichomes multiseriate; perianth 3-merous; pollen tricolpate; gynoecium of 3 to several carpels, carpels splitting at maturity; seeds without arils. Three genera of western North America.

Family Hypecoaceae

Herbs with watery sap and mostly pinnately or palmately divided leaves; flowers small, regular; petals 4 in 2 whorls, inner petals often lobed; stamens 4; gynoecium superior; fruit articulated into many 1-seeded joints, usually tapered above into a beak. The single genus *Hypecoum*, with about 15 species, occurs from Europe and the Mediterranean area to China.

Family Fumariaceae

Herbaceous annuals, biennials or perennials, commonly glaucous (smooth, no plant hairs); sap colourless and watery; leaves in basal rosettes or along the stem or both, ternate (3-lobed) to pinnately lobed or dissected; flowers bilaterally symmetrical to somewhat irregular (rarely regular), erect to nodding, bisexual, mostly cymose or racemose; sepals 2, often bractlike (*i.e.*, small and simple in the manner of bracts, small, green appendages attached just below flowers); petals 4 in 2 series, outer often saccate or spurred; stamens 6 in two groups (4 and ungrouped in *Pteridophyllum*); gynoecium 2-carpeled; fruit 2-valved; seeds 1 to many, often lustrous or arillate or both. Nineteen genera and about 425 species.

Subfamily *Pteridophylloideae*

Herbaceous perennial; leaves basal, pinnately lobed and fernlike; inflorescence without bracts, racemose; flowers regular, petals 4, free; nonsaccate; stamens 4. A single species (*Pteridophyllum racemosum*), restricted to Japan.

Subfamily *Fumarioideae*

Inflorescences with bracts; flowers bilaterally symmetrical to irregular; outer petals saccate to spurred, inner petals joined apically; stamens 6, diadelphous (fused by their filaments into two bundles). About 18 genera and 424 species, with the greatest concentration in the Northern Hemisphere and centred in Asia.

Critical appraisal. Many of the Asiatic species of the Papaverales order have been little studied from modern points of view, and this is one of taxonomic needs in the group. Further research on *Hypecoum* and *Pteridophyllum* would help to better place these genera with respect to each other and with respect to the principal families Papaveraceae and Fumariaceae.

The great divergence of taxonomic treatments of certain genera is baffling and calls for a sound evaluation. For example, the recognition by one authority of 123 species of *Eschscholzia* (California poppy) mostly from California is not compatible with the treatment by another, in which only eight species are given for the state. A third has recently suggested that *Eschscholzia* consists of about 14 species. A similar disparity of treatments exists for the genus *Platystemon*, and this should be carefully resolved.

BIBLIOGRAPHY. W.R. ERNST, "The Genera of Papaveraceae and Fumariaceae in the Southeastern United States," *J. Arnold Arbor.*, 43:315–343 (1962), although ostensibly limited geographically, this paper offers an important review of the taxonomy of the Papaverales as a whole; a broad bibliographical coverage is included. J. HUTCHINSON, "The Genera of Fumariaceae and Their Distribution," *Kew Bull.*, 1921:97–115 (1921), is an abbreviated taxonomic treatment of the family with keys to genera and species and includes geographical information. M.A. RYBERG, "A Morphological Study of the *Fumariaceae* and the Taxonomic Significance of the Characters Examined," *Acta Horti Bergiani*, 19:122–248 (1960), includes a general discussion of the geography of the family and a critique of the taxonomy on a worldwide basis.

(R.C.R.)

Paper and Paper Production

Paper is the basic material used for written communication and the dissemination of information. In addition, paper and paperboard provide materials for hundreds of other uses, such as wrapping, packaging, towelling, insulating, and photography.

Paper has been defined as a matted or felted sheet formed on a wire screen from water suspension. The word paper is derived from the name of the reedy plant papyrus, which grows abundantly along the Nile River in Egypt. In ancient times, the fibrous layers within the stem of this plant were removed, placed side by side, and crossed at right angles with another set of layers similarly arranged. The sheet so formed was dampened and pressed. Upon drying, the glue-like sap of the plant, acting as an adhesive, cemented the layers together. Complete defibring, an indispensable element in modern papermaking, did not occur in the preparation of papyrus sheets. Papyrus was the most widely used writing material in ancient times, and many papyrus records still survive.

Manufacture of papyrus

This article is divided into the following sections:

I. The papermaking process

HISTORICAL DEVELOPMENT

Beginnings of papermaking. Papermaking can be traced to about AD 105, when Ts'ai Lun, an official attached to the Imperial court of China, created a sheet of paper using mulberry and other bast fibres along with fish nets, old rags, and hemp waste. In its slow travel westward, the art of papermaking reached Samarkand, in Central Asia, in 751; and in 793 the first paper was made in Baghdad during the time of Harun-ar-Raschid, with the golden age of Islāmic culture that brought papermaking to the frontiers of Europe.

By the 14th century a number of paper mills existed in Europe, particularly in Spain, Italy, France, and Germany. The invention of printing in the 1450s brought a vastly increased demand for paper. Through the 18th century the papermaking process remained essentially unchanged, with linen and cotton rags furnishing the basic raw materials. Paper mills were more and more plagued by shortages; in the 18th century they even advertised and solicited publicly for rags. It was evident that a process for utilizing a more abundant material was needed.

Improvements in materials and processes. *Wood pulp.* In 1800 a book was published that launched development of practical methods for manufacturing paper from wood pulp and other vegetable pulps. Several major pulping processes were gradually developed that relieved the paper industry of dependency upon cotton and linen rags and made modern large-scale production possible. These developments followed two distinct pathways. In one, fibres and fibre fragments were separated from the wood structure by mechanical means; and in the other, the wood was exposed to chemical solutions that dissolved and removed lignin and other wood components, leaving cellulose fibre behind. Made by mechanical methods, groundwood pulp contains all the components of wood and thus is not suitable for papers in which high whiteness and permanence are required. Chemical wood pulps such as soda and sulfite pulp (described below) are used when high brightness, strength, and permanence are required. Groundwood pulp was first made in Germany in 1840, but the process did not come into extensive use until about 1870. Soda pulp was first manufactured from wood in 1852 in England, and in 1867 a patent was issued in the United States for the sulfite pulping process.

Mechanical and chemical pulping

Vat sizing. A sheet of paper composed only of cellulosic fibres ("waterleaf") is water absorbent. Hence, water-based inks and other aqueous liquids will penetrate and spread in it. Impregnation of the paper with various substances that retard such wetting and penetration is called sizing.

Before 1800, paper sheets were sized by impregnation with animal glue or vegetable gums, an expensive and tedious process. In 1800 Moritz Friedrich Illig in Germany discovered that paper could be sized in vats with rosin and alum. Though Illig published his discovery in 1807, the method did not come into wide use for about 25 years.

Chlorine bleaching. Discovery of the element chlorine in 1774 led to its use for bleaching paper stock. Lack of chemical knowledge at the time, however, resulted in production of inferior paper by the method, discrediting it for some years. Chlorine bleaching is a common papermaking technique today.

Introduction of machinery. Prior to the invention of the paper machine, paper was made one sheet at a time by dipping a frame or mold with a screened bottom into a vat of stock. Lifting the mold allowed the water to drain, leaving the sheet on the screen. The sheet was then pressed and dried (see Figure 1). The size of a single sheet was limited to the size of frame and mold that a man could lift from a vat of stock.

In 1798, Nicolas-Louis Robert in France constructed a moving screen belt that would receive a continuous flow of stock and deliver an unbroken sheet of wet paper to a pair of squeeze rolls. The French government recognized Robert's work by the granting of a patent.

The paper machine did not become a practical reality, however, until two engineers in England, both familiar with Robert's ideas, built an improved version for their employers, Henry and Sealy Fourdrinier, in 1807. The Fourdrinier brothers obtained a patent also. Two years later a cylinder paper machine (described below) was devised by John Dickinson, an English papermaker.

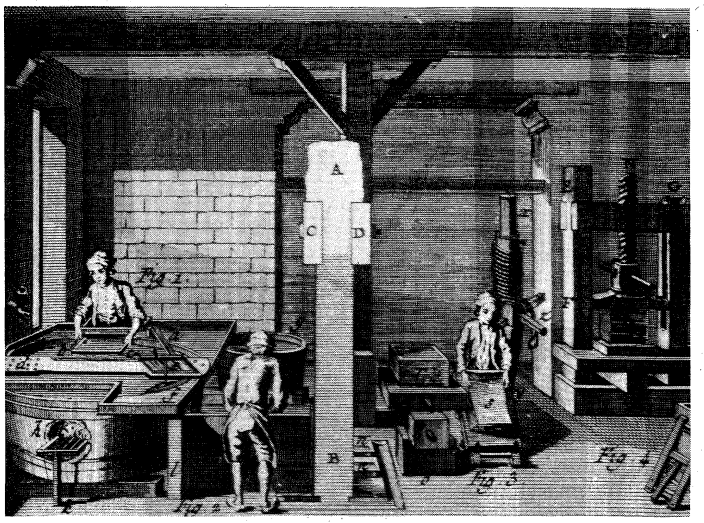


Figure 1: *Making paper by hand in an 18th-century paper mill.* The workman in "Fig. 1," the vatman, is dipping the mold into the vat of stock. The workman in "Fig. 2" is "couching" or pressing the wet sheet while it is still on the mold. The workman in "Fig. 3" is separating the sheets from the molds, making a stack to be pressed by the device in "Fig. 4." Woodcut from the *Encyclopédie*, edited by Denis Diderot and Jean Le Rond d' Alembert, 1751–72.

From these crude beginnings, modern papermaking machines evolved. By 1875 paper coated by machinery was being made for use in the printing of halftones by the new photoengraving process, and in 1884 Carl F. Dahl invented sulfate (kraft) pulp in Danzig, Germany.

Although the paper machine symbolizes the mechanization of the paper industry, every step of production, from the felling of trees to the shipment of the finished product, has also seen a dramatic increase in mechanization, thus reducing hand labour. As papermaking operations require the repeated movement of large amounts of material, the design and mechanization of materials-handling equipment has been and continues to be an important aspect of industry development.

Basic operations in papermaking

Though modern inventions and engineering have transformed an ancient craft into a highly technical industry, the basic operations in papermaking remain the same to this day. The steps in the process are as follows: (1) a suspension of cellulosic fibre is prepared by beating it in water so that the fibres are thoroughly separated and saturated with water; (2) the paper stock is filtered on a woven screen to form a matted sheet of fibre; (3) the wet sheet is pressed and compacted to squeeze out a large proportion of water; (4) the remaining water is removed by evaporation; and (5) depending upon use requirements, the dry paper sheet is further compressed, coated, or impregnated.

The differences among various grades and types of paper are determined by: (1) the type of fibre or pulp; (2) the degree of beating or refining of the stock; (3) the addition of various materials to the stock; (4) formation conditions of the sheet, including basis weight, or substance per unit area; and (5) the physical or chemical treatment applied to the paper after its formation.

FIBRE SOURCES

Properties of cellulose fibres

The cell walls of all plants contain fibres of cellulose, an organic material known to chemists as a linear polysaccharide. It constitutes about one-third of the structural material of annual plants and about one-half that of perennial plants. Cellulose fibres have high strength and durability. They are readily wetted by water, exhibiting considerable swelling when saturated, and are hygroscopic —*i.e.*, they absorb appreciable amounts of water when exposed to the atmosphere. Even in the wet state, natural cellulose fibres show no loss in strength. It is the combination of these qualities with strength and flexibility that makes cellulose of unique value for paper manufacture.

Most plant materials also contain nonfibrous elements or cells, and these also are found in pulp and paper. The nonfibrous cells are less desirable for papermaking than fibres but, mixed with fibre, are of value in filling in the sheet. It is probably true that paper of a sort can be produced from any natural plant. The requirements of paper quality and economic considerations, however, limit the sources of supply.

Wood. Pulped forest tree trunks (boles) are by far the predominant source of papermaking fibre. The bole of a tree consists essentially of fibres with a minimum of nonfibrous elements, such as pith and parenchyma cells.

Softwood and hardwood fibres

Forests of the world contain a great number of species, which may be divided into two groups: coniferous trees, usually called softwoods, and deciduous trees, or hardwoods. Softwood cellulose fibres measure from about two to four millimetres (.08 to .16 inch) in length, and hardwood fibres range from about 0.5 to 1.5 millimetres (0.02 to 0.06 inch). The greater length of softwood fibres contributes strength to paper; the shorter hardwood fibres fill in the sheet and give it opacity and a smooth surface.

When the sulfite process (see below) was the chief method of pulping in the early days of the pulp industry, spruce and fir were the preferred species. Since that time, advances in technology, particularly the introduction of the kraft process (described below), have permitted the use of practically all species of wood, greatly expanding the potential supply.

Because of the enormous and rapidly growing consumption of wood for pulp, concern regarding the depletion of forest resources has been expressed, even though yearly growth often exceeds the annual harvest. In 1962, for example, though new growth exceeded the harvest by a considerable margin, much of it was inferior in quality and less accessible than the harvested trees. Moreover, wood is now being harvested at a more rapid pace. Approximately 40 percent of the harvest is going into pulp, and that figure is expected to increase. There is also a rising public demand for withdrawal of forest land from timber production for recreational use and to prevent disturbance to the ecology of certain areas. On the other hand, application of new techniques in fertilization and genetics have brought about enormous increases in the productivity of forest lands in some areas. Universal adoption of these new methods could help to assure a future supply of wood.

Two significant trends in pulpwood utilization deserve mention. Until recently, lumbering and other wood-using industries were operated quite independently of the pulp industry. Since World War II, however, the waste from the wood-using industries, such as sawdust, has increasingly been used for pulp. In addition, more abundant and less desirable hardwoods have been used as a source of pulp. The woodyard of a pulp mill formerly stored pulpwood in the form of roundwood logs, but recently there has been a trend toward storing in the form of chips.

Rags. Cotton and linen fibres, derived from textile and garment mill cuttings; cotton linters (the short fibres recovered from the processing of cottonseed after the separation of the staple fibre); flax fibres; and clean sorted rags are still used for those grades of paper in which maximum strength, durability, and permanence are required, as well as fine formation, colour, texture, and feel. These properties are attributed to the greater fineness, length, and purity of rag fibre as compared with most wood pulp. Rag papers are used extensively for bank note and security certificates; life insurance policies and legal documents, for which permanence is of prime importance; technical papers, such as tracing paper, vellums, and reproduction papers; high-grade bond letterheads, which must be impressive in appearance and texture; lightweight specialties such as cigarette, carbon, and Bible papers; and high-grade stationery, in which beauty, softness, and fine texture are desired.

Properties of rag paper

Rags are received at the paper mill in bales weighing from 400 to 1,200 pounds (200 to 500 kilograms). After mechanical threshing, the rags are sorted by hand to remove such foreign materials as rubber, metal, and paper and to eliminate those rags containing synthetic fibres

and coatings that are difficult to remove. Following sorting, the rags are cut up, then dusted to remove small particles of foreign materials, and passed over magnetic rolls to remove iron.

The cut and cleaned rags are cooked (to remove natural waxes, fillers, oils, and grease) in large cylindrical or spherical boilers of about five-ton (4,500-kilogram) capacity. About three parts of cooking liquor, a dilute alkaline solution of lime and soda ash or caustic soda combined with wetting agents or detergents, are used with each part of rags. Steam is admitted to the boiler under pressure, and the contents are cooked for three to ten hours.

Once cooked, the rags are washed, then mechanically beaten. The beating shortens the fibre, increases the swelling action of water to produce a softened and plastic fibre, and fibrillates or frays the fibre to increase its surface area. All of these actions contribute to better formation of the paper sheet, closer contact between fibres, and the formation of interfibre bonding that gives the paper strength and coherence.

Wastepaper and paperboard. Table 1 shows the consumption of wastepaper stock in selected countries during the year 1969, both in tons and as a percentage of all paper produced in each country.

Table 1: Consumption of Wastepaper Stock in Selected Countries, 1969

	wastepaper stock used (000 tons)	percentage of paper produced
United States	10,446	19.5
Japan	4,646	26.8
West Germany	2,543	44.6
United Kingdom	2,039	37.1
France	1,464	33.3
Netherlands, The	545	31.0
Sweden	255	5.7
Finland	145	3.2

Source: *Pulp and Paper*, June 25, 1970.

By using greater quantities of wastepaper stock, the need for virgin fibre is reduced, and the problem of solid waste disposal is minimized. The expansion of this source is a highly complex problem, however, because of the difficulties in gathering wastepaper from scattered sources, sorting mixed papers, and recovering the fibre from many types of coated and treated papers.

Wastepaper may be classified into four main categories: high-grades, old corrugated boxes, printed news, and mixed paper. High-grade and corrugated stocks originate mainly in mercantile and industrial establishments. White paper wastes accumulate in envelope and printing plants, while tabulating cards are supplied by large offices. Much magazine stock comes from newsstand returns, but some comes from homes. Corrugated waste is supplied by manufacturing plants and retail stores. Printed news is derived from newsstand returns and home collections. Mixed paper comes from waste baskets of office buildings and similar sources. In recent years there has been considerable interest in wastepaper recycling in the interest of ecology.

Converters of paper and paperboard have also increasingly turned to new materials combined with paper and paperboard to give their products special characteristics. Although these new materials have broadened the market for paper, their presence has posed new problems in reusing paper stock. The most common new ingredients are asphalt, synthetic adhesives, metal foils, plastic and cellulose-derivative films and coatings, and some printing inks.

Some objectionable materials can be sorted from wastepaper, and packers generally try to remove them completely. If the producer of wastepaper knows the materials he is using, he can usually segregate trouble-causing substances at the source. Much depends on good cooperation and communication among the papermaker, dealer, packer, and producers so that all may understand what is and what is not acceptable.

There are two distinct types of paper recovery systems: (1) recovery based upon de-inking and intended for printing-grade or other white papers, accounting for about 5 to 6 percent of the total; and (2) recovery without de-inking, intended for boxboards and coarse papers, accounting for the remainder.

Paper recovery systems

Recovery with de-inking. In the de-inking recovery process, the bales of wastepaper are opened, inspected, and fed into a pulper, a cylindrical tank with capacity ranging from one to several tons of stock and provided with agitator blades that circulate and agitate the stock. Hot water and various chemicals help the agitator separate and disperse the fibres.

The amount and type of chemicals used vary considerably from mill to mill. Caustic soda is by far the most generally used, but it is often supplemented with soda ash, silicate of soda, phosphates, and surfactants (wetting agents). The temperature range is from 150° to 190° F (65° to 90° C).

The pulpers are aided in the collection and separation of large pieces of trash by special devices. After the stock leaves the pulper, it is screened to remove finer trash particles and washed to remove the dispersed ink and chemicals. In some instances the stock is bleached with hypochlorite to improve its whiteness.

Recovery without de-inking. In pulping paper stock where de-inking is not necessary, the equipment is similar to that already described. Hot water is also used in the pulper, but the chemicals for dissolving and dispersing the ink are not needed. The stock is screened and washed to remove trash and dirt.

The use of paper stock in the paper mill presents difficulties because of the presence of foreign materials. Miscellaneous trash has always required operators to be watchful, and its presence depends on the source of the waste and the care with which the paper is prepared for market.

Natural fibres other than wood. Since cellulose fibre is a major constituent of the stems of plants, a vast number of plants represent potential sources of paper; many of these have been pulped experimentally. A rather substantial number of plant sources have been used commercially, at least on a small scale and at various times and places. Indeed, the use of cereal straws for paper predates the use of wood pulp and is widely practiced today throughout the world, although on a relatively small scale of production. Because many parts of the world are deficient in forests, the development of the paper industry in these areas appears to depend to a considerable degree upon the use of annual plants and agricultural fibres.

Agricultural fibres

Nonwoody plant stems differ from wood in containing less total cellulose, less lignin, and more of other materials. This means that pulps of high cellulose content (high purity) are produced in relatively low yield, whereas pulps of high yield contain high proportions of other materials. Papers made from these pulps without admixture of other fibre tend to be dense and stiff, with low tear resistance and low opacity.

The morphology (form and structure) of the cells of annual plants also differs considerably from wood. Whereas the nonfibrous (parenchyma) cells of coniferous wood comprise a minor proportion of the wood substance, in annual plants this cell type is a major constituent. As hardwoods also often contain considerable amounts of nonfibrous cells, there is a closer resemblance between hardwood pulps and pulps from annual plants.

The preferred pulping reagents for nonwood plants are the alkalis: caustic soda, lime and soda ash, and kraft liquor (caustic soda and sodium sulfide). A characteristic of the pulping of annual plants, compared with wood, is the milder treatment necessary to produce pulp. Straw, for example, may be pulped with milk of lime in a spherical digester at a steam pressure of 25 pounds per square inch (about two kilograms per square centimetre) and a cooking time of eight to ten hours. The amount of lime used is about 10 percent of the amount of dry fibre.

Pulping reagents

Cereal straws. In the United States straw pulp was formerly used extensively for corrugating medium (*i.e.*, sheet fluted to form the inner ply of corrugated board). Since then, the use of straw pulp for corrugating medium

has been replaced by semichemical hardwood pulp. Straw pulp is still made in several European and Asiatic countries on a small scale.

Bagasse. The residue from the crushing of sugarcane, called bagasse, contains about 65 percent fibre, 25 percent pith cells, and 10 percent water solubles. An essential element in the conversion of bagasse to a satisfactory paper is the mechanical removal of a substantial proportion of the pith prior to the pulping operation. Pulping may be carried out either with soda or with kraft cooking liquor and by batch or continuous systems. Bagasse fibre averages 1.5 to two millimetres (0.06 to 0.08 inch) in length and is relatively fine.

The use of bagasse is substantial in several Latin-American countries and in the Middle East. The utilization of bagasse for paper in all the sugar-producing countries that are deficient in forest resources is a practical step.

Esparto. A desert plant of the Mediterranean area, especially in southern Spain and northern Africa, esparto grass has a higher cellulose content than most nonwood plants, with greater uniformity of fibre size and shape. The use of esparto for papermaking was developed in Great Britain in 1856. Consumption rose steadily until the mid-1950s, but since has steadily declined.

Esparto held its own against the competition with wood pulp for some time because of its favourable papermaking properties. The stock forms well on a paper machine because of free drainage and uniform fibre length, compared with rag or wood pulp. Esparto printing papers possess good resilience in contact with the printing plate, have good opacity and smoothness, and are relatively lint-free. Another important characteristic of papers made from esparto is dimensional stability with changes in moisture content.

Bamboo. Botanically, bamboo is classified as a grass, even though it attains a considerable size and the stems or culms resemble wood in hardness and density. It was demonstrated many years ago that satisfactory pulp could be made from bamboo.

Because of the abundance of bamboo in Southeast Asia, where increased production of paper is greatly needed, much interest has been displayed in bamboo pulp development. The growing cycle of bamboo is favourable, for the culms can be harvested without destroying the root system. Under ideal conditions of soil fertility and moisture, an established stand of bamboo probably would produce more fibre per acre (or hectare) per year than any other plant. Wild bamboo, however, is difficult to harvest and transport economically; so far, the interest in it has not been translated into any large-scale production. Pulp mills make use of bamboo in India, Thailand, and the Philippines. Considerable quantities of bamboo pulp are said to be made in China, but details are lacking.

Flax, hemp, jute, kenaf. These plants are characterized by a high proportion of long, flexible bast fibres that are readily separated and purified from the other materials in the plant. Consequently, such fibres have long been used for textiles and rope-making. Most of this fibre reaching the paper industry in the past has been secondary or waste fibre. It has been highly prized because of the strength and durability it imparts to such products as tags, abrasive paper (sandpaper), cover stock, and other heavy-duty paper. It is also used for duplicating and manifold paper, in which extremely light weight must be combined with exceptional strength. Flax is grown expressly for high-grade cigarette paper. Experimental quantities of kenaf have been grown and made into various grades of paper.

Synthetic fibres. The development and use of a great variety of man-made fibres have created a revolution in the textile industry in recent decades. It has been predicted that similar widespread use of synthetic fibres may eventually occur in the paper industry. Active interest has been evident in recent years, both on the part of fibre producers and of paper manufacturers. Many specialty paper products are currently being made from synthetic fibres.

The advantages of synthetic or man-made fibres in papermaking can be summarized as follows:

Whereas natural cellulose fibres vary considerably in size and shape, synthetic fibres can be made uniform and of selected length and diameter. Long fibres, for example, are necessary in producing strong, durable papers. There are limitations, however, to the length of synthetic fibres that may be formed from suspension in water because of their tendency to tangle and to rope together. Even so, papers have been made experimentally with fibres several times longer than those typical of wood pulp; these papers have improved strength and softness properties. Advantages of synthetic fibres

Natural cellulose fibres have limited resistance to chemical attack and exposure to heat. Because synthetic fibre papers can be made resistant to strong acids, they are useful for chemical filtration. Paper can even be made from glass fibre, and such paper has great resistance to both heat and chemicals.

The natural cellulose fibres of ordinary paper are hygroscopic; *i.e.*, they absorb water from the air and reach an equilibrium depending upon the relative humidity. The moisture content of paper, therefore, changes with atmospheric conditions. These changes cause swelling and shrinkage of fibers, accounting for the puckering and curling of papers. Synthetic fibres not subject to these changes can be used to produce dimensionally stable papers.

The cheapest man-made fibre, rayon, costs from three to six times as much as an equivalent amount of wood pulp, whereas most of the true synthetics, such as the polyamides (nylon), polyesters (Dacron, Dynel), acrylics (Orlon, Creslan, Acrilan), and glass, cost from ten to 20 times as much. This difference in cost does not preclude the use of existing synthetics, but it limits their use to special items in which the extra qualities will justify the additional cost. The cost factor is increased by the absence in most synthetic fibres of the bonding property of natural cellulose fibres. When beaten in water, natural fibres swell and cement together as they dry. Paper made from synthetics must be bonded by the addition of an adhesive, requiring an additional manufacturing step. Cost of wood pulp and synthetic fibres

There is a distinct similarity between synthetic fibre "papers" and the class of sheet materials known as nonwovens. As a step in the manufacture of yarn, staple fibres are carded (*i.e.*, separated and combed) to form a uniform, lightweight, and fragile web. Subsequently, this web is gathered together to form a strand or sliver, which is drawn and spun into yarn. If several of these flat webs, however, are laminated together and bonded with adhesive, a nonwoven fabric that has properties resembling both paper and cloth results. In this area it is difficult to draw a clear distinction between what is paper and what is cloth. Processes are now available to form sheet material both by the dry forming method and by the water forming or paper system. When textile-type fibres are formed into webs by either of these processes, the resulting products have properties that enable them to compete in some fields traditionally served by textiles.

PROCESSES FOR PREPARING PULP

Mechanical or groundwood pulp is made by subjecting wood to an abrading action, either by pressing the wood against a revolving grinding stone or by passing chips through a mill. The wood fibres are separated and, to a considerable degree, fragmented.

Chemical wood pulp is made by cooking wood chips with chemical solutions in digesters operated at elevated temperature and pressure. The chemicals used are: (1) sulfite salts with an excess of sulfur dioxide, and (2) caustic soda and sodium sulfide (the kraft process). The lignin of the wood is made soluble, and the fibres separate as whole fibres. Further purification can be accomplished by bleaching. Chemical wood pulp that is purified both by bleaching and by alkaline extraction is called alpha or dissolving pulp. It is used for specialty papers, for rayon and cellulose film production, and for cellulose derivatives, such as nitrate and acetate.

Semichemical pulp is made by treating wood chips with sulfite or alkali in amounts and under conditions that soften the lignin but dissolve only part of it. The softened chips are then defibred.

Mechanical or groundwood pulp. Pulpwood may arrive at the mill as bolts four feet (1.2 metres) in length or as full-length logs. The logs are sawn to shorter length, and the bolts are tumbled in large revolving drums to remove the bark. The debarked wood is next sent to grinders, where its moisture content is important for ease of grinding and quality of pulp. Moisture content should be at least 30 percent and preferably 45 to 50 percent. Wood of low moisture content is presoaked in a pond or sprayed with water.

Importance of moisture content

Early grinders employed round slabs of natural sandstone 27 inches (69 centimetres) wide and 54 inches (137 centimetres) in diameter, often directly connected to water wheels, to produce five or six tons (4,500–5,400 kilograms) of pulp per day. The wood was hand-loaded into the grinders.

Today's much larger pulp grinders are usually powered by electric motors and automatically loaded. In a recently built mill, each grinder is gear connected to a 10,000 horsepower motor; the pulpstone, at 360 revolutions per minute, can handle wood 60 to 64 inches (1.5–1.6 metres) long. Hydraulic cylinders produce a pressure of 200 pounds per square inch against the stone face. Pulp production from each stone is 130 to 150 tons (120,000–135,000 kilograms) every 24 hours.

The first artificial grinding stone was produced in 1924; since that time, artificial stones have replaced natural sandstone. Silicon carbide and aluminum oxide are the abrasives used in the manufacture of pulpstones. The abrasive material is broken down into a mixture of sizes that is screened to give fractions of uniform grain size. The abrasive grains are mixed with binder and fired at high temperature (2,300° C or 4,200° F) in the form of segments that are assembled to form the abrasive surface of the pulpstone.

The pulp stock flows from the grinder pit to a series of rifflers and screens, which separate the heavy foreign material and pieces of unfibred wood (shives), knots, bark, and the like.

Most groundwood pulp flows directly to an adjacent paper mill for use as stock. When shipped, it is formed into a sheet on a cylindrical vacuum filter. The sheets are pressed in a hydraulic press to a moisture content of about 50 percent, and the pressed sheets are formed into bales.

An important test to control the quality of groundwood pulp is freeness: the readiness with which water drains from and through a wet pad of pulp. Groundwood pulps are much less "free" than chemical wood pulps.

Freeness characteristic

In groundwood pulp, the fibres are fragmented, and there is considerable debris (fines). Also, groundwood contains all the chemical constituents of wood, including lignin, hemicellulose, resin, and various colouring materials. This means that papers containing groundwood are subject to discoloration (yellowing) upon exposure to light and heat and after aging. The yellowing of newspaper and much book paper is an example of this. Because groundwood fibres are relatively short and have only a moderate ability to bond to each other, papers containing them do not have high strength. On the other hand, papers containing groundwood have good opacity; they are bulky and have good printing qualities.

Groundwood pulp does not have a high whiteness, being limited in this quality by the colour of the wood from which it is made. Though often bleached with peroxide or hydrosulfite to improve whiteness, it does not equal pure cellulose.

Chemical wood pulp. The effect of sulfurous acid (H_2SO_3) in softening and defibring wood was observed by B.C. Tilghman, a U.S. chemist, as early as 1857. Several years later he renewed his experiments and, in 1867, was granted a patent for making paper pulp from vegetable material. He used high temperature and pressure and observed that the presence of a base such as calcium was important in preventing burned or discoloured batches of pulp. His work, however, did not result in commercial use of the process.

The sulfite process. During the 1870s the sulfite process for pulping wood was the subject of experimental work in Sweden, England, Germany, and Austria. Within a few years the process was in commercial operation both in Europe and in North America. For many decades, the sulfite process was the leading process for the pulping of wood. Since 1940, however, the kraft process has taken a predominant position, and sulfite mills are no longer being constructed.

Sulfite cooking liquor, as it is pumped to the digester at the start of a "cook," consists of free sulfur dioxide dissolved in water at a concentration of 4 to 8 percent, together with from 2 to 3 percent in the form of bisulfite. Sulfite digestion is normally carried out as a batch process in a pressure vessel, a steel shell with an acid-resistant lining of ceramic tile set in acid-proof cement or stainless steel. A common digester measures 16 feet (five metres) in diameter and 50 feet (15 metres) in height, with a domed top and a conical bottom. It has a capacity of 12 to 15 tons (11,000–13,500 kilograms) of pulp per batch. Digesters of up to 35 tons (32,000 kilograms) capacity have been constructed. Pulp mills normally have a series of digesters arranged in a digester building.

Sulfite digestion

After the blow valve is closed at the bottom, the wood chips are allowed to flow into the top opening and are distributed to fill the digester completely. Hot acid from the accumulator is pumped into the digester unit, completely filling it and replacing the air. Steam provides the heat.

At the end of the cook, the contents of the digester are blown to a blowpit by rapid opening of the bottom valve. The violence of the blow defibres the cooked chips.

From 1 to 6 percent of the digester charge is undesirable material such as knots, uncooked chips, dirt, bark, fibre bundles, and shives. The screen room separates the unwanted particles from accepted fibre, normally on the basis of particle size; there is an increasing use of the centrifugal principle, which separates particles on the basis of density.

The sulfite cooking liquor does not "cook out" or disintegrate bark and other foreign material to the same degree as kraft liquor (described below), and hence more care must be used in selecting and cleaning wood chips for sulfite.

In the conventional sulfite cook using softwood, the typical yield is 44 to 46 percent, based on wood and with a lignin content of 2 to 5 percent. At that point, a relatively light-coloured pulp with good strength properties is obtained, suitable for use in the unbleached state, especially in mixture with groundwood for a variety of printing papers. For pulps in which high brightness (whiteness) is desired, the residual lignin is removed by bleaching.

The soda process. In 1851 paper pulp was experimentally produced from wood by cooking it with caustic soda at elevated temperature and pressure. Though this soda process attained commercial importance, soda pulp was of relatively low strength; and use of the process was limited to manufacturing filler pulps from hardwood, which were then mixed with a stronger fibre for printing papers. Because this process consumed relatively large quantities of soda, papermakers devised methods for recovering soda from the spent cooking liquor; recovery has remained an integral part of alkaline pulping ever since.

The kraft process. In 1884 a German chemist, Carl F. Dahl, employed sodium sulfate in place of soda ash in a soda pulping recovery system. This substitution produced a cooking liquor that contained sodium sulfide along with caustic soda. Pulp so produced was stronger than soda pulp and was called "kraft" pulp, so named from the German and Swedish word for "strong." The process has also been termed the sulfate process because of the use of sodium sulfate (salt cake) in the chemical makeup. Sulfate, however, is not an active ingredient of the cooking liquor.

Many soda mills were converted to kraft because of the greater strength of the pulp. Kraft pulp, however, was dark in colour and difficult to bleach; for many years the growth of the process was slow because of its limitation to papers in which colour and brightness were unimportant. In the 1930s, bleached kraft became commercially important with the discovery of new bleaching tech-

niques. The availability of pulp of high whiteness and the expanding demand for unbleached kraft in packaging resulted in rapid growth of the process, making kraft the predominant wood-pulping method.

Paper produced by the kraft process is particularly strong and durable. Acceptable pulp can be produced in the kraft process from many species of wood not suitable for sulfite. The various pines, for example, especially southern yellow pine, contain large amounts of wood resin or pitch. Chemically altered and dissolved in the kraft process, this material is removed from the pulp and becomes a valuable by-product. The wood pitch is not removed to the same degree in the sulfite process, and hence high-resin woods, such as pine, are not suitable.

The kraft cooking operation

In the cooking operation, wood chips are prepared and fed to the digesting equipment by methods previously described. The cooking vessels are still widely used as batch digesters. In the past 25 years, however, continuous digesters have been developed and are being widely adopted by the kraft industry. These huge cylindrical towers, over 200 feet (60 metres) in height, have a number of zones or compartments. Wood chips and cooking liquor are fed into the top and injected into successive zones of high pressure and temperature, where impregnation and cooking takes place as the chips progress downward. Additional zones wash the spent liquor from the chips. Continuous digesters are capable of producing 600 tons (or about 540,000 kilograms) of pulp per day.

In batch cooking, after the digester is charged with chips, a mixture of "black liquor," the spent liquor from a previous cook, and "white liquor," a solution of sodium hydroxide and sodium sulfide from the chemical recovery plant, is pumped in. The digester is heated either by direct injection of steam or by the circulation of the cooking liquor through a heat exchanger.

Figure 2 shows a flow diagram of a kraft pulping operation. After completion of the cook, the spent cooking liquor is washed from the pulp; the latter is then screened and sent to the bleach plant or directly to the paper mill if it is to be used unbleached. Some of the spent liquor (black liquor) is used for an admixture with white liquor to charge new cooks; the remainder is sent to the recovery plant to reconstitute cooking chemicals.

All the sodium used for digestion is contained in the spent liquor, mostly in the form of sodium salts and sodium organic derivatives. The amount of sodium present is such that its re-use is economically necessary.

Semichemical pulp. For semichemical pulping wood preparation and chipping are essentially the same as that for other wood-pulping processes. The chips are steeped and impregnated with inorganic chemical solutions similar to those used for full chemical pulping, but in smaller amounts and with less severe conditions. Probably the most common is the solution of sodium sulfite in the neutral range, between acidity and alkalinity. Other agents used in some cases are acid sulfite, caustic soda, and kraft cooking liquor.

After the impregnation operation, the chips are fed into one or more disk refiners (described below) in series. The attrition action of refiners reduces the softened chips to pulp. The yield of semichemical pulp based on wood is 66 to 90 percent. The higher fibre yield pulps are usually termed chemimechanical pulps.

The semichemical pulps have chemical and strength properties intermediate between softwood, groundwood, and full chemical pulps. These are used in a fairly wide range of papers and boards. The major tonnage of semichemical pulps goes into the light board, termed corrugating medium, which is fluted to serve as the interior layer of corrugated boxboard in heavy-duty containers. Stiffness and adequate strength are the important properties. Semichemical pulp is used extensively in a variety of low-cost printing papers.

Bleaching and washing. The use of calcium and sodium hypochlorites to bleach paper stock dates from the beginning of the 19th century. In the early days of sulfite pulp manufacture, a single-stage treatment of pulp at low consistency, using calcium hypochlorite (chlorinated lime), satisfied most requirements.

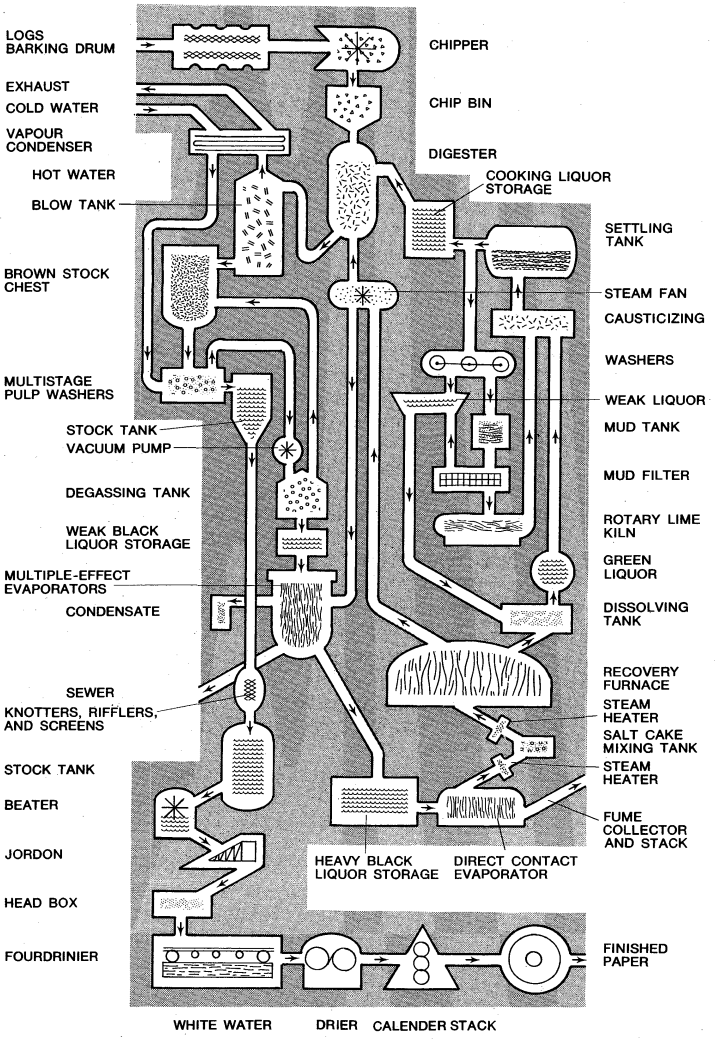


Figure 2: Stages in the kraft pulping process.
Drawing by D. Meighan

This simple bleaching treatment, however, is not practical for kraft that is difficult to bleach, nor can it retain maximum pulp strength. Accordingly, multistage bleaching systems have evolved in which various sequences of chemical treatment are employed, depending upon the type of unbleached pulp and special requirements.

During the normal first stage in a modern bleach plant, the unbleached pulp is chlorinated. Three to 4 percent of gaseous chlorine is rapidly mixed with pulp at a temperature of 70° to 80° F (21° to 27° C); the mixture is quite acid due to the acidity of the chlorine. Chlorine is absorbed largely by reaction with the non-carbohydrate components of pulp, with no brightening effect and with only slight dissolution of lignin.

Alkaline extraction

In the following stage an alkaline extraction with dilute caustic soda dissolves chlorinated compounds, which are then washed out.

In its simplest sequence the final stage consists of a treatment with a very alkaline hypochlorite to neutralize the solution, followed by a final wash.

In recent years the compound chlorine dioxide (ClO_2) has become available for on-site preparation; it is too unstable to be shipped for wood pulp bleaching. By the use of small amounts of ClO_2 in later bleaching stages, it is possible to achieve high degrees of purification and brightness without the degradation of cellulose.

The brightness of paper and other materials is determined by special reflection meters containing photoelectric cells that measure the amount of light of selected wavelength reflected by the surface. Freshly prepared pure magnesium oxide is considered to be 100 on the brightness scale. On this scale unbleached sulfite and groundwood cover the range from about 50 to about 62;

peroxide bleached groundwood, 66 to 72; single-stage hypochlorite sulfite, 80 to 85; multistage bleached pulp, 85 to 88; and multistage with chlorine dioxide, 90 to 94.

MANUFACTURE OF PAPER AND PAPERBOARD

Preparation of stock. Mechanical squeezing and pounding of cellulose fibre permits water to penetrate its structure, causing swelling of the fibre and making it flexible. Mechanical action, furthermore, separates and frays the fibrils, submicroscopic units in the fibre structure. Beating reduces the rate of drainage from and through a mat of fibres, producing dense paper of high tensile strength, low porosity, stiffness, and rattle.

The Hollander beater. An important milestone in papermaking development, the Hollander beater consists of an oval tank containing a heavy roll that revolves against a bedplate. The roll is capable of being set very accurately with respect to the bedplate, for the progressive adjustment of the roll position is the key to good beating. A beater may hold from 300 to 3,000 pounds (135 to 1,350 kilograms) of stock, a common size being about 24 feet (7 metres) long, 12 feet (4 metres) wide, and about 3.5 feet (1 metre) deep. A centre partition provides a continuous channel.

Pulp is put into the beater, and water is added to facilitate circulation of the mass between the roll and the bedplate. As the beating proceeds, the revolving roll is gradually lowered until it is riding full weight on the fibres between it and the bedplate. This action splits and mashes the fibres, creating hairlike fibrils and causing them to absorb water and become slimy. The beaten fibres will then drain more slowly on the paper machine wire and bond together more readily as more water is removed and the wet web pressed. Much of the beating action results from the rubbing of fibre on fibre. Long fibres will be cut to some extent.

Sizing, fillers, and dyes

The beater is also well-adapted for the addition and mixing of other materials, such as sizing, fillers, and dyes. By mounting a perforated cylinder that can rotate partially immersed in the beater stock, water can be continuously removed from the beater, and the stock can be washed.

Though many design modifications have been made in the Hollander beater over the years, the machine is still widely used in smaller mills making specialty paper products. For large production modern mills have replaced the beater by various types of continuous refiners.

Conical refiner. In mills that receive baled pulp and use refiners, the pulp is defibred in pulpers. While there are a number of variations in basic design, a pulper consists essentially of a large, open vessel, with one or more bladed, rotating elements that circulate a pulp–water mixture and defibre or separate fibres. The blades transform the pulp or wastepaper into a smooth mixture. Unlike beaters and refiners, pulpers do not reduce freeness and cause fibrillation in the fibres. A typical pulper has a capacity of 2,000 pounds (900 kilograms) of fibre in 6 percent solution and requires 150 horsepower to drive it.

The original continuous refiner is the Jordan, named after its 19th-century inventor. Like the beater, the Jordan has blades or bars, mounted on a rotating element, that work in conjunction with stationary blades to treat the fibres. The axially oriented blades are mounted on a conically shaped rotor that is surrounded by a stationary bladed element (stator).

Disk refiner. Like other refiners, the disk refiner consists of a rotating bladed element that moves in conjunction with a stationary bladed element. The disk refiner's plane of action, however, is perpendicular to the axis of rotation, simplifying manufacture of the treating elements and replacement. Since the disk refiner provides a large number of working edges to act upon the fibre, the load per fibre is reduced and fibre brushing, rather than fibre cutting, may be emphasized.

Sizing. Sizing has been described above as the treatment given paper to prevent aqueous solutions, such as ink, from soaking into it. A typical sizing solution consists of a rosin soap dispersion mixed with the stock in an amount of 1 to 5 percent of fibre. Since there is no affinity between rosin soap and fibre, it is necessary to use a coupling agent, normally alum (aluminum sulfate). The acidity of alum precipitates the rosin dispersion, and the positively charged aluminum ions and aluminum hydroxide flocs (masses of finely suspended particles) attach the size firmly to the negatively charged fibre surface.

Filling or loading. Paper intended for writing or printing usually contains white pigments or fillers to increase brightness, opacity, and surface smoothness, and to improve ink receptivity. Clay (aluminum silicate), often referred to as kaolin or china clay, is commonly used, but only in a few places in the world (Cornwall, in England, and Georgia, in the United States) are the deposits readily accessible and sufficiently pure to be used for pigment. Another pigment is titanium dioxide (TiO_2), prepared from the minerals rutile and anatase. Titanium dioxide is the most expensive of the common pigments and is often used in admixture with others.

Kaolin pigment

Calcium carbonate, also used as a filler, is prepared by precipitation by the reaction of milk of lime with either carbon dioxide (CO_2) or soda ash (calcium carbonate, Na_2CO_3). Calcium carbonate as a paper filler is used mainly to impart improved brightness, opacity, and ink receptivity to printing and magazine stocks. Specialty uses include the filling of cigarette paper, to which it contributes good burning properties. Because of its reactivity with acid, calcium carbonate cannot be used in systems containing alum.

Other fillers are zinc oxide, zinc sulfide, hydrated silica, calcium sulfate, hydrated alumina, talc, barium sulfate, and asbestos. Much of the filler consumed is used in paper coatings (see below).

Since most fillers have no affinity for fibres, it is necessary to add an agent such as alum to help hold the filler in the formed sheet. The amount of filler used may vary from 1 to 10 percent of the fibre.

Colouring. The most common way to impart colour to paper is to add soluble dyes or coloured pigment to the paper stock. Many so-called direct dyes with a natural affinity for cellulose fibre are highly absorbed, even from dilute water solution. The so-called basic dyes have a high affinity for groundwood and unbleached pulps.

Interfibre bonding agents. Various agents are added to paper stock to enhance or to modify the bonding and coherence between fibres. To increase the dry strength of paper, the materials most commonly used are starch, polyacrylamide resins, and natural gums such as locust bean gum and guar gum. The most common type of starch currently used is the modified type known as cationic starch. When dispersed in water, this starch assumes a positive surface charge. Because fibre normally assumes a negative surface charge, there is an affinity between the cationic starch and the fibre.

Cationic starch

The natural cellulose interfibre bonding that develops as a sheet of paper dries is considered to be due to interatomic forces of attraction known to physical chemists as hydrogen bonding or van der Waals forces (see CHEMICAL BONDING). Because these attractive forces are neutralized or dissolved in water, wet paper has practically no strength. Though this property is convenient for the recovery of wastepaper, some papers require wet strength for their intended use. Wet strength is gained by adding certain organic resins to the paper stock that, because of their chemical nature, are absorbed by the fibre. After formation and drying of the sheet, the resins change to an insoluble form, creating water-resistant bonds between fibres.

Formation of paper sheet by machines. In a paper machine, interrelated mechanisms operating in unison receive paper stock from the beater, form it into a sheet of the desired weight by filtration, press and consolidate the sheet with removal of excess water, dry the remaining water by evaporation, and wind the travelling sheet into reels of paper. Paper machines may vary in width from about five to over 25 feet (1.5–eight metres), in operating speed from a few hundred feet to over 3,000 feet (900 metres) per minute, and in production of paper from a few tons per day to over 300 tons (270,000 kilograms)

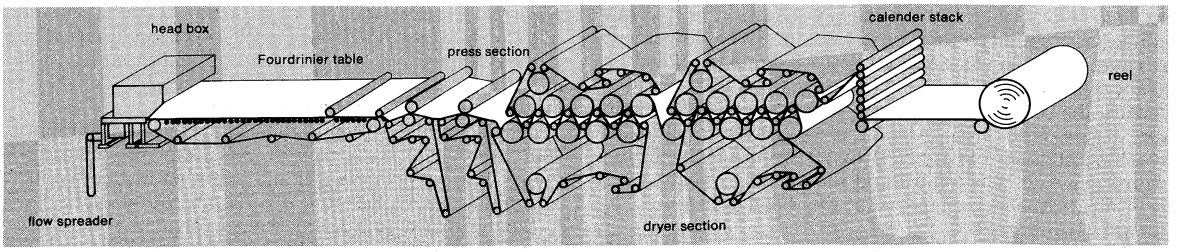


Figure 3: Elements of a Fourdrinier-type paper machine.
From K.W. Britt (ed.), *Handbook of Pulp and Paper Technology*, 2nd ed.; copyright © 1964 by Litton Educational Publishing, Inc., reproduced by permission of Van Nostrand Reinhold Co.

per day. The paper weight (basis weight) may vary from light tissue, about ten grams per square metre (0.03 ounce per square foot), to boards of over 500 grams per square metre (1.6 ounces per square foot).

Cylinder and Fourdrinier machines

Traditionally, paper machines have been divided into two main types: cylinder machines and Fourdrinier machines. The former consists of one or more screen-covered cylinders, each rotating in a vat of dilute paper stock. Filtration occurs by flow action from the vat into the cylinder, with the filtrate being continuously removed. In the Fourdrinier machine, a horizontal wire-screen belt filters the stock. In recent years a number of paper machines have been designed that depart greatly from traditional design. These machines are collectively referred to as "formers." Some of these formers retain the travelling screen belt but form the sheet largely on a suction roll. Others eliminate the screen belt and use a suction cylinder roll only. Still others use two screen belts with the stock sandwiched between, with drainage on both sides.

Figure 3 shows a typical modern Fourdrinier machine. The various functional parts are the headbox; stock distribution system; Fourdrinier table, where sheet formation and drainage of water occur; press section, which receives the wet sheet from the wire, presses it between woolen felts, and delivers the partially dried sheet to the dryer section; dryer section, which receives the sheet from the presses and carries it through a series of rotating, steam-heated cylinders to remove the remaining moisture; size press, which permits dampening the sheet surface with a solution of starch, glue, or other material to improve the paper surface; calendar stock, for compressing and smoothing the sheet; and the reel.

The headbox. The function of the headbox is to distribute a continuous flow of wet stock at constant velocities, both across the width of the machine and lengthwise of the sheet, as stock is deposited on the screen. Equal quantities of properly dispersed stock should be supplied to all areas of the sheet-forming surface. The early headbox, more commonly called a flowbox or breastbox, consisted of a rectangular wooden vat that extended across the full width of the machine behind the Fourdrinier breast roll. The box was provided with baffles to mix and distribute the stock. A flat metal plate extending across the machine (knife slice) improved dispersion of the fibre suspension, providing distribution of flow across the machine, and also metered the flow to produce a sheet of uniform weight. To accommodate increased speed in modern headboxes, the knife slice is designed to develop a jet of liquid stock on the moving wire. Modern headboxes are enclosed, with pressure maintained by pumping.

Forming Section. The Fourdrinier table section of a paper machine is a large framework that supports the table rolls, breast roll, couch roll, suction boxes, wire rolls, and other Fourdrinier parts. The wire mesh upon which the sheet of fibre is formed is a continuous rotating belt that forms a loop around the Fourdrinier frame. The wire, not a permanent part of the machine, is delicate and requires periodic replacement. It is a finely woven metal or synthetic fibre cloth that allows drainage of the water but retains most of the fibres. The strands of the Fourdrinier wire are usually made of specially annealed bronze or brass, finely drawn and woven into a web commonly in the range of 55 to 85 mesh (strands per inch). Even finer wires are used for such grades as cigarette paper, coarser wires for heavy paperboard and pulp sheets. Various types of weave are used to obtain maximum wire life and to reduce wire marking on the wet sheet.

The table rolls, in addition to supporting the wire, function as water-removal devices. The rapidly rotating roll in contact with the underside of the wire produces a suction or pumping action that increases the drainage of water through the wire.

The dandy roll is a light, open-structured unit covered with wire cloth and placed on the wire between suction boxes, resting lightly upon the wire and the surface of the sheet. Its function is to flatten the top surface of the sheet and improve the finish. When the dandy roll leaves a mesh or crosshatch pattern, the paper is said to be "woven." When parallel, translucent lines are produced, it is said to be "laid." When names, insignia, or designs are formed, the paper is said to be "watermarked." Paper watermarks have served to identify the makers of fine papers since the early days of the art. A watermark is actually a thin part of the sheet and is visible because of greater transmission of light in its area compared with other areas of the sheet. Because light transmission can be varied by degrees, it is possible to produce watermarks in the form of portraits or pictures.

Watermarks in paper

The final roll over which the formed sheet passes, before removal from the Fourdrinier wire, is the "couch roll." Prior to the transferring operation, the couch roll must remove water from and consolidate the sheet to strengthen it. In modern machines the couch roll is almost always a suction roll.

The press section. The press section increases the solids content of the sheet of paper by removing some of the free water contained in the sheet after it is formed. It then carries the paper from the forming unit to the dryer section without disrupting or disturbing sheet structure and reduces the bulk or thickness of the paper.

The first two functions are always desirable and necessary. Pressing always results in compaction, and this may or may not be desirable depending upon the grades being made.

Paper machine felts

Felts for the press section act as conveyor belts to assist the sheet through the presses, as porous media to provide space and channels for water removal, as textured cushions or shock absorbers for pressing the moist sheet without crushing or significant marking, and as power transfer belts to drive nondriven rolls or parts.

Woven felts of wool, often with up to 50 percent synthetic fibres, are made by a modified woolen textile system. Selected grades of wool are scoured, blended, carded, and spun into yarn. The yarn is woven into flat goods, leaving a fringe at each end. The ends are brought together and joined to produce an endless, substantially seamless belt.

Paper machine felts have a limited life ranging from about a week to several months. Their strength and water-removal ability is gradually lost through wear, chemical and bacterial degradation, and by becoming clogged with foreign material.

Press rolls must be strong, rigid, and well-balanced to span the wide, modern machines and run at high speed without distortion and vibration. Solid press rolls consist of a steel or cast iron core, covered with rubber of various hardnesses depending upon the particular service required. Suction press rolls consist of a bronze or stainless steel shell two inches (five centimetres) or more in thickness and usually covered with one inch of rubber.

The dryer. Paper leaving the press section of the machine has a solids content or dryness of 32 to 40 percent. Because of the relatively high cost of removing water by evaporation, compared with removing it by mechanical means, the sheet must be as dry as possible when it enters the dryers. The dryer section of a conventional paper machine consists of from 40 to 70 steam-heated drying cylinders. After passing around the cylinders, the sheet is held in intimate contact with the heated surfaces by means of dryer felts.

Until recently, relatively heavy, rather impermeable cloths composed of wool, cotton, asbestos, or combinations of these materials covered the dryer portion of the paper machine. Such cloths are termed dryer felts, though felting or fulling process is rarely used in their manufacture. Recently, relatively lightweight, highly permeable cloths called dryer fabric have come into use.

For conventional dryer felts, cotton is still the most commonly used fibre, although it is seldom used alone. The main difference between the conventional dryer felt and the open-mesh dryer fabric is air or vapour permeability. High permeability is desirable because it allows the escape of the water vapour from the sheet.

Vapour permeability

For every ton of paper dried on the paper machine, approximately two tons (1,800 kilograms) of water are evaporated into the atmosphere. About 50 to 60 tons (45,000 to 54,000 kilograms) of air are required to remove the water vapour, with about 6,000 pounds (2,700 kilograms) of steam required by the dryers.

Finishing and converting. The rolls of paper produced by the paper machine must still undergo a number of operations before the paper becomes useful to the consumer. These various operations are referred to as converting or finishing and often make use of intricate and fast-moving machinery.

There are two distinct types of paper conversion. One is referred to as wet converting, in which paper in roll form is coated, impregnated, and laminated with various applied materials to improve properties for special purposes. The second is referred to as dry converting, in which paper in roll form is converted into such items as bags, envelopes, boxes, small rolls, and packs of sheets. A few of the more important converting operations are described here.

Coating. Paper has been coated to improve its surface for better reproduction of printed images for over 100 years. The introduction of half-tone and colour printing has created a strong demand for coated paper. Coatings are applied to paper to achieve uniformity of surface for printing inks, lacquers, and the like; to obtain printed images without blemishes visible to the eye; to enhance opacity, smoothness, and gloss of paper or paperboard; and to achieve economy in the weight and composition of base paper stock by the upgrading effect of coating.

The chief components of the water dispersion used for coating paper are pigment, which may be clay, titanium dioxide, calcium carbonate, satin white, or combinations of these; dispersants to give uniformity to the mixture or the "slip"; and an adhesive binder to give coherence to the finished coating. The latter may be a natural material such as starch or a synthetic material such as latex.

Equipment installed between dryer sections on the paper machine can apply the coating (on-machine coating), or it can be done by a separate machine, using rolls of paper as feed stock (off-machine coating).

Extrusion coating

The extrusion coating process, a relatively new development in the application of functional coating, has gained major importance in the past 20 years. The process is used to apply polyethylene plastic coatings to all grades of paper and paperboard. Polyethylene resin has ideal properties for use with packaging paper, being waterproof; resistant to grease, water vapour, and gases; highly stable; flexible in heat sealing; and free from odour and toxicity.

In the extrusion-coating machine, the polyethylene resin is melted in a thermoplastic extruder that consists of a drive screw within an electrically heated cylinder. The cylinder melts and compacts the resin granules and extrudes the melt in a continuous flow under high pressure. The resin is discharged through a film-forming slot die. The die has electric heaters with precision temperature controls to give uniform temperature and viscosity to the plastic melt. The slot opening can be precisely adjusted to control film uniformity and thickness.

The hot extruded film is then stretched and combined with paper between a pair of rolls, one of which is a rubber-covered pressure roll and the other a water-cooled, chromium-plated steel roll. The combination takes place so rapidly that a permanent bond is created between the plastic film and the paper before they are cooled by the steel roll.

Corrugating. The most widely used package for commodities and manufactured products is the corrugated shipping container. A corrugated box consists of two structural elements: the facings (linerboard), and the fluting structure (corrugating medium).

Linerboard facings are of two general types: the Fourdrinier kraft liner is made of pine kraft pulp, usually unbleached, in an integrated mill as a continuous process from the tree to the paper web, and the cylinder liner is made from reprocessed fibres, generally from used containers, providing a content of about two-thirds kraft.

The operation begins by unwinding the single-face liner and corrugating medium from holders, threading the medium into the fluting rolls, applying adhesive to the tips, and bringing the medium in contact with the liner to form a single-face web. Next, the single-face web passes another glue roll that applies adhesive to the exposed flute tips of the medium. The second face liner is brought in contact with the single-face web, and the combined board travels through a hot plate section between belts to set the bond, to a cooling section, and then to a slitter–scorer.

THE WORLD PAPER INDUSTRY

The paper industry tends to be concentrated in those countries that are industrially advanced and have abundant supplies of fibrous raw material, especially wood. There is a large-scale international trade in wood pulp, pulpwood, and paper flowing from those countries with large forest resources to those countries with less or that are as yet undeveloped. The world production of paper and paperboard in the early 1970s was about 140,000,000 tons with the leading countries as shown in Table 2.

Table 2: World Production of Paper, Paperboard, and Pulp, 1970
(in 000 tons)

	paper and paperboard	pulp
U.S.	47,599	38,298
Japan	12,973	8,801
Canada	11,655	16,117
U.S.S.R.	6,704	6,714
West Germany	5,516	1,803
U.K.	4,979	—
Sweden	4,359	8,148
Finland	4,258	6,222
France	4,134	1,813
China	3,750	2,700
Italy	3,451	891
Norway	1,421	2,205

Source: *Pulp and Paper*, June 30, 1971.

The world production of pulp in 1969 was estimated to be 114,112,000 tons. The difference between the tonnage of paper and that of pulp is accounted for by the utilization of wastepaper and the use of various nonfibrous additive materials in paper manufacture. The production of pulp is widespread, occurring in practically all countries that have appreciable paper production. The abundance of suitable wood species in North America and the Scandinavian countries has established these areas as the chief pulp-producing regions. The leading pulp-producing countries are also listed in Table 2.

II. Paper properties and uses

Used in a wide variety of forms, paper and paperboard are characterized by a wide range of properties. In the

thousands of paper varieties available, some properties differ only slightly and others grossly. The identification and expression of these differences depend upon the application of standard test methods, generally specified by industry and engineering associations in the papermaking countries of the world.

SUBSTANCE AND QUANTITY MEASUREMENT

Weight or substance per unit area, called basis weight, is a fundamental property of paper and paperboard products. From the first uses of paper in the printing trades, it has been measured in reams, originally 480 sheets (20 quires) but now more commonly 500 sheets (long reams). The term ream weight commonly signifies the weight of a lot or batch of paper. Since the printing trades use a variety of sheet sizes, there can be numerous ream weights for paper having the same basis weight.

Measurement in reams

Table 3 gives basis weight ranges for some common papers.

Table 3: Basis Weight Ranges for Common Papers

paper grade	basis weight*
Carbonizing tissue†	5¾–25
Facial tissue	9–10
Manifold tissue‡	10–20
Tea-bag tissue	8–12
Blotting paper	114–266
Bond	25–60
Book, uncoated	27–91
Book, coated two sides	45–109
Bristol, postcard	135
Envelopes	16–40
Greaseproof (glassine)	20–50
Newsprint	30–35
Multiwall bag stock	40–70
Kraft wrapping, most common	25–80
Cover stock, file folders	66–216

*Basis weight in pounds per 480 sheet ream, cut to 24 by 36 inches. †Paper to be impregnated with wax and pigment to make carbon paper. ‡Employed for making several carbon copies.

To determine basis weight, the sample is brought to equilibrium under standard conditions (75° F or 24° C; 50 percent relative humidity). The paper specimens must consist of at least ten sheets with a total area of not less than 100 square inches (about 600 square centimetres). Since the properties of paper change with moisture content, all tests are conducted under standard conditions.

The caliper (thickness) of paper or paperboard in thousandths of an inch is measured by placing a single sheet under a steady pressure of seven to nine pounds per square inch between two circular and parallel plane surfaces, the smaller of which has an area of 0.25 square inch (1.6 square centimetres).

The density or specific gravity of paper is calculated from the basis weight and caliper and may vary over wide limits. Glassine, for example, may be 1.4 grams per cubic centimetre and creped wadding, used for packaging breakables, only 0.1 gram per cubic centimetre. Most common papers are in the range of 0.5 to 0.7 gram per cubic centimetre.

STRENGTH AND DURABILITY

The strength of paper is determined by the following factors in combination: (1) the strength of the individual fibres of the stock; (2) the average length of the fibre; (3) the interfibre bonding ability of the fibre, which is enhanced by the beating and refining action; and (4) the structure and formation of the sheet.

Resistance to rupture when subjected to various stresses is an important property in practically all grades of paper. Most papers require a certain minimum strength to withstand the treatment received by the product in use; but even where use requirements are not severe, the paper must be strong enough to permit efficient handling in manufacture. Tensile strength is the greatest longitudinal stress a piece of paper can bear without tearing apart. The stress is expressed as the force per unit width of a test specimen.

Since the weight of the paper and the width of the test specimen affect the force of rupture, a conventional method of comparing inherent paper strength is the breaking length; that is, the length of a paper strip in metres that would be just self-supporting. This value varies from about 500 metres (1,600 feet) for extremely soft, weak tissue to about 8,000 metres (26,000 feet) for strong kraft bag paper, and to about 14,000 metres (46,000 feet) for sheets of paper made under ideal laboratory conditions.

Breaking length

Because some paper products such as towels, sanitary tissues, and filter paper are subjected to wetting by water in their normal use, wet tensile testing has become important. This test is essentially the same as that for dry tensile strength, except that the specimen is wetted. Paper that has not been specifically treated to produce wet strength possesses from about 4 to about 8 percent of its dry strength when completely wetted. By treating paper as described above, wet strength can be raised to about 40 percent of the dry strength.

Bursting strength. One of the oldest and most widely used strength tests for paper and paperboard is the bursting test, or Mullen test. It is defined as the hydrostatic pressure (caused by liquids at rest) necessary to cause rupture in a circular area of a given diameter. Paperboard has a requirement of several hundred pounds per square inch. Other strength tests for which standard methods exist are tearing strength and folding endurance.

Stiffness and softness. The resistance of paper to a bending force is evident in the various operations of its manufacture and in its many uses. The range in this property extends from very soft, flexible tissues to rigid boards. Thicker and heavier sheets tend to be stiff, whereas soft, flexible sheets are light and thin. Even at the same weight there is a considerable difference in stiffness, chiefly due to the compactness and the amount of bonding of the sheet.

Liquid penetration. Because paper is composed of a randomly felted layer of fibre, the structure has a varying degree of porosity. Thus, the ability of fluids, both liquid and gaseous, to penetrate the structure is a property both highly significant to the use of paper and capable of being widely varied by the conditions of manufacture.

Sizing paper with vegetable materials and rosin-like substances has already been described. When paper began to be used for wrapping, consumers demanded sizing treatments that could protect the contents of the package from the effects of fluid transfer through the paper wrapping. In some instances complete impermeability was required. In another direction the use of paper as an absorbent medium for wiping up liquids, for filtering, and for saturating has created a demand for maximum wettability and permeability toward water and other fluids.

In certain types of packaging, paper must resist grease and oil penetration. The resistance of paper to the penetration of water can be increased by treatment of fibre with materials that lack affinity for water, with little effect upon sheet porosity, but the penetration of oil materials is little affected by such treatment. Oil and grease resistance is attained, in fact, by reduction in porosity. So-called greaseproof paper is made by beating an easily hydrated pulp to extremely low freeness, which results in a dense sheet with very little void space.

Absorbent papers such as towelling, sanitary tissue, and blotting and filter paper are normally made from lightly beaten stock. Since cellulose is naturally hydrophilic (*i.e.*, has a strong affinity for water), absorbent papers have a minimum of foreign materials associated with the fibre. Of particular importance are the wood rosins that may be present in pulp and produce a self-sizing effect, especially upon aging.

Absorbent papers

OPTICAL PROPERTIES

The most important optical properties of paper are brightness, colour, opacity, and gloss.

The term brightness has come to mean the degree to which white or near-white papers and paperboard reflect

the light of the blue end of the spectrum (*i.e.*, their reflectance). This reflectance is measured by an instrument that illuminates paper at an average angle of incidence of 45° and a wavelength of 457μ (microns). Brightness measured in this way is found to correlate closely with subjective estimates of the relative whiteness of paper.

Opacity is one of the most desired properties of printing and writing papers. Satisfactory performance of such papers requires that there be little or no "show-through" of images from one side of the sheet to the other. Satisfactory opacity in printing papers requires that white mineral pigments be incorporated with the paper stock or applied as a coating.

The terms gloss, glare, finish, and smoothness are used in describing the surface characteristics of paper. The broad term finish refers to the general surface characteristics of the sheet. Smoothness refers to the absence of surface irregularities under either visual or use conditions. Gloss refers to surface lustre and connotes a generally pleasing aspect. Glare is used for a more intense reflection and a more unpleasant effect. Calendering and coating are important paper-treating methods that affect gloss. Gloss of paper is determined by measuring percent reflectance at a low angle of incidence, 15 degrees (75 degrees from the perpendicular).

PAPER GRADES

Bond paper. Bond is characterized by a degree of stiffness, durability for repeated handling and filing, resistance to the penetration and spreading of ink, bright colour, and cleanliness. There are two groups of bond papers: rag content pulp and chemical wood pulp. Rag content bond may vary from 25 to 100 percent cotton fibre content. The principal uses of bond paper are for letterhead stationery, advertising pieces, announcements, leases, deeds, writs, judgments and other legal documents, currency, certificates, and insurance policies.

Types of bond papers

Book paper. Most book papers are made of various combinations of chemical wood pulp; and for lower priced grades, groundwood, semichemical, and de-inked wastepaper are also used. In addition to pulp, the "furnish" from which book papers are made contains various amounts of sizing, fillers, and dyes.

Uncoated book paper comes in four finishes: (1) antique or eggshell, (2) machine finish, MF, (3) English finish, EF, and (4) supercalendered. Antique has the roughest surface. High bulking pulps, such as soda pulp, are used and only slightly beaten in stock preparation. The sheet is lightly calendered (pressed between rollers) to provide a degree of surface smoothness while preserving the antique or eggshell appearance. Machine finish has a medium-smooth surface obtained for this finish from a calender stock at the dry end of the machine. Machine finish book is a relatively inexpensive general utility paper. It is principally used for books, catalogs, circulars, and other matter using line etchings. Machine finish book may be used for halftones up to a 100-line screen. English finish is a step higher in the book paper scale; this finish is distinguished from machine finish by a higher degree of stack beating, by greater pressure between the rollers of a machine calender, and by calendering at a greater moisture content of the sheet. Supercalendered book is the smoothest surface that can be obtained without coating. The finish is obtained by a special calendering operation after the paper leaves the paper machine. The supercalender presses the paper between successive sets of iron and compressed fibre rolls that make a smooth, compact printing surface. It is used for books, brochures, and magazines where halftone printing in the range of a 100–120 line screen is required.

Coated book papers are produced to create surfaces suitable for the printing of fine-screen halftones. Coated book paper must be uniformly smooth, receptive to printing inks, have high brightness and gloss, and be capable of folding without cracking.

Bible paper, as the name implies, was developed for lightweight, thin, strong, opaque sheets for such books as bibles, dictionaries, and encyclopaedias, which require minimum bulk. Bible papers are pigmented (loaded) with such pigments as titanium dioxide and barium sulfate and contain long fibres and artificial bonding agents to maintain strength.

Bristol. The general term bristol refers to a group of stiff, heavy papers with thicknesses ranging from 0.006 inch (0.15 millimetre) upward. These grades are made from various combinations of chemical wood pulp. The stock is beaten to a medium degree and usually well sized to prevent penetration of moisture. Increasingly important in recent years has been the use of bristols for the punch cards used in tabulating and sorting machines.

Groundwood and newsprint papers. These are printing and converting grades containing varying amounts of groundwood pulp, together with small percentages of chemical wood pulp for strength and durability.

For many years newsprint was virtually the only use for groundwood pulp, but more recently, due to improvements in the pulping process and to the introduction of a bleaching process for this pulp, a class of printing papers of broad utility has been developed. Magazines, paperbound books, catalogs, directories, and general commercial printing consume large quantities of these papers.

Groundwood papers are noted for an even, uniform formation and a high degree of opacity. These papers tend to be bulky and are receptive to printing ink. They do not have high whiteness and tend to turn yellow when exposed to light and after long aging.

Kraft wrapping. Kraft wrapping, a heavy stock used for paper bags, is used in greater volume than all other wrapping papers combined. It is composed of wood pulp in unbleached condition made from softwoods, usually pine. It is distinguished by outstanding tensile and tearing strength. Kraft wrapping is sized to retard wetting when exposed to water. For wrapping of wet materials, the paper may be given wet strength by treatment with special resins. Multiwall sacks of kraft paper are used for shipment of bulk materials.

Paperboards. The term paperboard is a general term descriptive of products 0.012 inch (0.30 millimetre) or more in thickness, made of fibrous materials on paper machines. Paperboard is commonly made from wood pulp, straw, wastepaper, or a combination of these materials.

There are three main types of paperboard: (1) boxboards, used for such products as food board, food trays, plates, and paper boxes; (2) container boards, for the manufacture of corrugated and solid fibre shipping containers; and (3) paperboard specialties, including such items as binders board, electrical pressboard, and building boards.

Types of paperboard

Sanitary papers. The group of papers known collectively as the sanitary grades include toilet tissue, towelling, facial tissue, and napkins. These grades are made from various proportions of sulfite and bleached kraft pulps with relatively little refining of the stock to preserve a soft, bulky, absorbent sheet. This sheet is further softened by machine creping, in which the wet sheet is pressed upon a smooth drying roll and subsequently removed by running against a flat stationary metal blade (doctor blade). The sheet is piled up upon itself, thus producing a creped effect. Facial tissue is dry-creped; that is, drying is complete on the drying roll before the creping doctor blade. Other papers are creped while still partially wet. Towelling is generally of heavier weight than the tissues and is usually creped while still wet. Napkins are of somewhat heavier weight than tissues. The plastic nature of paper fibres when slightly moist permits the reproduction of surface patterns by embossing to a remarkable degree. Paper napkins are an example of this art.

Because of the soft, bulky texture of sanitary papers, they are relatively weak. Since they are often exposed to wetting in use, they are often treated with resins to increase wet strength.

BIBLIOGRAPHY. AMERICAN PAPER AND PULP ASSOCIATION, *The Dictionary of Paper*, 3rd ed. (1965), is a compilation of terms and definitions relating to paper and paper manufacture. K.W. BRITT (ed.), *Handbook of Pulp and Paper Technology*, 2nd ed. rev. (1970), contains 61 articles grouped in 9 sections describing the most important features of pulp

and paper manufacturing technology. A multivolume textbook that has gone through several editions is J.N. STEPHENSON (ed.), *Pulp and Paper Manufacture*, 4 vol. (1950–55). Each volume covers a particular area of the industry: pulp manufacture, stock preparation, paper manufacture, and finishing and converting. JAMES P. CASEY, *Pulp and Paper: Chemistry and Chemical Technology*, 2nd ed. rev., 3 vol. (1960–61), emphasizes the chemistry of papermaking and is noteworthy as a review of published literature with extensive references. A two-volume work intended as a textbook for students and prepared under the auspices of the Joint Textbook Committee of the Technical Association of the Pulp and Paper Industry is *Pulp and Paper Science and Technology*, ed. by C. EARL LIBBY (1962). S.A. RYDHOLM, *Pulping Processes* (1965), places primary emphasis on the chemical reactions of pulping and is a standard reference work in the field, of interest primarily to specialists in pulping. DARD HUNTER, *Papermaking: The History and Technique of an Ancient Craft*, 2nd ed. rev. (1957), is a classic account of handmade paper over the centuries since its invention in China early in the Christian era.

(K.W.Br.)

Papuan Languages

The Papuan languages are those languages spoken in an area centred around New Guinea and extending from the islands of Alor, Halmahera, and Timor in the west to the Santa Cruz Archipelago in the east. The group includes approximately 700 languages, used by about 2,400,000 speakers. The term Papuan was originally employed merely to distinguish these languages from the Austronesian (Malayo-Polynesian) and Australian languages, and, until recently, most Papuan languages were believed to be unrelated to each other. Intensive research by teams of linguists in the New Guinea area since the late 1950s, however, has resulted in a revolutionary change in the Papuan linguistic picture, and it is now known that about 350, and perhaps even as many as 450, of the approximately 650 identified Papuan languages are related. These are spoken by more than 1,600,000 speakers, who occupy almost three-quarters of the New Guinea mainland. Belonging to 76 to 105 language families (depending on the classification used), these languages together form the Central New Guinea macrophylum. A macrophylum is a group of languages related less closely than those of a language family or stock.

Unrelated Papuan groups and languages. Of the known Papuan languages that cannot yet be linked with the macrophylum, about 130 belong to large groups, and 26 to small individual groups. Approximately 50 languages, called isolates, seem to be unrelated either to each other or to the established groups. Further research may well show that additional Papuan languages belong to the macrophylum, especially after hitherto linguistically unknown areas in New Guinea have been studied.

Enga, the numerically largest language

Each of the individual Papuan languages is, for the most part, spoken by only a few hundred to a few thousand people, though the numerically largest one, Enga, in the Western Highlands District of Australian New Guinea, has over 130,000 speakers, and several other languages are each spoken by tens of thousands. Even if related, the languages generally show considerable diversity, especially in vocabulary. A few basic grammatical characteristics are, however, shared by many languages. In numerous instances it is difficult to determine the border line between the languages and dialects, despite the presence of marked differences between two forms of speech.

Recent extensive research into the Papuan languages has resulted in the preliminary classification of most of them. Numerous new languages have been discovered, though large areas, mainly in West Irian, remain unknown. Despite the concentration of research on discovery and classification, a number of grammatical and lexical studies have been prepared, and folklore has been collected. The Summer Institute of Linguistics, an association of Protestant missionaries specializing in studying primitive languages and involved in literacy training and Bible translation, has carried out an extensive native language literacy program in New Guinea with a measure of success.

Linguistic characteristics. Most Papuan languages show extreme grammatical complexity. Their verbs vary to reflect a wide range of numbers and other features of the subject as well as of the direct and indirect objects and the beneficiary. For example, in Kiwai, a language of the Central and South New Guinea phylum, the verb *ai-ni-mi-bi-du-mo-iauri-ama-ri-go* means "they three will certainly see us two." Similarly, in the Monumbo language of the Torricelli phylum *mbepe*$_1$*-nge*$_2$ *tsi*$_3$*-p*$_4$*-ing*$_5$*-em*$_6$ can be translated as "you gave him a taro." Literally, the parts of this utterance are: taro$_1$–singular$_2$, 2nd person singular subject (you)$_3$–give, past form$_4$–masculine class 3rd person singular indirect object (to him)$_5$–plant class singular direct object$_6$. Verbs also indicate tense, aspect, mood, and the direction and circumstances in which the action they designate is performed; *e.g.*, in Gadsup, of the East New Guinea Highlands phylum, *kùmù-à-nk-ədəd-òn-ték-əp-ón-ì-nó-ké*, "had he indeed wanted to go down for him?", has several prefixes and suffixes indicating emphasis, tense, direction, and the question status. (The accent marks indicate tones; the ə is a sound pronounced like *a* in "sofa." The dashes separate the various components of the utterances, but do not normally appear in nonlinguistic texts).

Verb forms for one or more actions

There are, basically, two major types of verb forms in many Papuan languages. One, which can be referred to as the normal type, is found in sentences in which only one action is referred to; the other, which may be referred to as the special verb form, occurs in sentences in which more than one action is mentioned. In the latter type of sentence, which may have one or more special verb forms, the normal verb form also appears with the last verb in the sentence.

Numerous Papuan languages have gender and noun class systems, some with up to ten or more classes. Bewilderingly complex variations of adjectives, numerals, demonstratives, and subject and object markers often result, since these words have special forms for each of the various classes of nouns. An example is the sentence *ame akwum kuvambakwum sumupar amenakwum salikəmba*, "I saw my two big women," from Angoram of one of the small Sepik phyla. If "women," *akwum*, is replaced by "arrows," "gardens," or "frogs," etc., the entire sentence, rather than a single word, is changed: when "arrows" is substituted, the phrase becomes *ame pwanggli kəpanggli klupar amenakanggli salikənggliya;* and when "gardens" is used, it changes to *ame konggəmbər kəvambər pələpar amenkəmbər salikəmbəra.*

Many languages are tonal, with changes of pitch in words and syllables that affect the meaning of the words. The interaction of Papuan tonal systems with patterns of stress and syllable length can be extremely intricate.

Classification and distribution. Interrelated language families are grouped in phyla, groups of languages more distantly related than those of a family or stock but more certainly or closely related than those of a macrophylum. Apart from the isolates, or unrelated languages, the number of known Papuan phyla is 21. Of these, 8 are small and consist of only 2 to 6 languages each. Of the remaining 13, 6 constitute the central New Guinea macrophylum, and 3 more can tentatively be included in it.

Central New Guinea macrophylum

The central New Guinea macrophylum comprises: (1) the East New Guinea Highlands phylum, found predominantly in the Highlands and Chimbu Districts of Australian New Guinea; (2) the Finisterre-Huon phylum in the Morobe and Madang Districts, Australian New Guinea; (3) the Central and South New Guinea phylum, located mainly in the Gulf and Western Districts of Papua in the Australian sector of New Guinea and in southern and northeastern West Irian; (4) the West New Guinea Highlands phylum of the highlands area of West Irian; (5) the South-East New Guinea phylum, in the Central, Northern, Morobe, and Milne Bay Districts of the Australian sector of New Guinea; (6) the Madang phylum in the Madang District of Australian New Guinea.

Tentative member phyla of the Central New Guinea macrophylum are: (1) the Adelbert Range phylum in the Madang District of Australian New Guinea; (2) the Middle Sepik-Upper Sepik-Sepik Hill phylum, which could also be classified as constituting three separate phyla; (3) the

Anga stock (stocks are intermediate between phyla and families) spoken in the Eastern Highlands, Morobe, and Gulf Districts of the Australian sector of New Guinea.

Large Papuan phyla that are not members of the Central New Guinea macrophylum are the Ramu phylum in the Madang and East Sepik Districts of Australian New Guinea; the Torricelli phylum in the same districts; the West Papuan phylum on the Doberai (Vogelkop) Peninsula of West Irian, and in northern Halmahera, eastern Timor, and Alor; and the Bougainville phylum on the island of Bougainville. Small isolated phyla are located mainly in the West Sepik and Gulf Districts of the Australian sector of New Guinea and in the Santa Cruz Archipelago. The highest concentration of known isolates is also in these districts and in the East Sepik District, the New Britain-New Ireland area, and the British Solomon Islands.

Some Papuan languages are difficult to classify because of strong Austronesian influence upon them. Most Papuan languages are of only regional importance, but a few have achieved some cultural significance outside their immediate area because of their use as missionary languages.

BIBLIOGRAPHY. Most earlier general studies have become obsolete. Comprehensive recent bibliographical and factual information may be found in D.C. LAYCOCK and C.L. VOORHOEVE, "History of Research in Papuan Languages," in *Current Trends in Linguistics*, vol. 8, *Linguistics in Oceania* (1971); and in S.A. WURM, "Papuan Linguistic Situation," *ibid.*, updated in *Papuan Languages of Oceania* (1972). Individual language descriptions are given in H. MCKAUGHAN (ed.), *Languages of the Eastern Family* (1971). Extensive descriptive and comparative materials on Papuan languages appear in the serial publications *Pacific Linguistics*.

(S.A.W.)

Pará

Pará is a state (*estado*) of northern Brazil through which the Lower Amazon River flows to the sea. It is bounded on the north by Guyana, Surinam, and the Territory of Amapá, on the northeast by the Atlantic Ocean, on the east by the states of Maranhão and Goiás, on the south by Mato Grosso, and on the west by Amazonas. Although it is the third-largest state in Brazil and lushly tropical, it supports but a small and scattered population. Its area is 481,872 square miles (1,248,042 square kilometres). Its population (1970) was 2,197,072. The capital and chief city is Belém.

History. The Amazon region was never very attractive to the Portuguese settlers of Brazil. Belém was founded in 1615, chiefly to keep other European nations from settling there. Spanish Jesuit missions were the first settlements upstream, as at Santarém in 1661; they were finally expelled by the Portuguese in 1710. Pará was made a captaincy in 1652, reunited with Maranhão in 1654, and re-established in 1772. It did not acknowledge the Brazilian Empire established in 1822 but yielded to force in 1823. It became a state when the new republic was founded in 1889. In the 1870s a colony of people from the southern United States, together with slaves they brought with them, settled near Santarém and tried to plant cotton and sugarcane. The colony was not successful. Between 1850 and 1910 there was a period of feverish activity as workers went out into the forests to tap the rubber trees. Rubber in large quantities was shipped out through Belém, and the city grew rapidly in size and importance. Production dropped rapidly after 1910, however. Even the Ford Motor Company plantations on the Rio Tapajóz, although technically successful, failed to produce enough rubber for the needs of the Ford industries.

The natural environment. The dominant physical feature of Pará is the outlet of the Amazon Basin, which crosses the state for about 500 miles (800 kilometres) from west to east in entering the Atlantic Ocean. The lower valley is comparatively narrow, the territory on both sides rising to the level of the ancient plateau that once covered this part of the continent. On the north is the Guiana Highlands, and on the south the country rises in forested terraces and is broken by escarpments caused by the erosion of the northern slope of the great central plateau of Brazil.

Climate

The state is crossed by the Equator, and the climate is equatorial: contrary to popular impression, the temperatures are never so high as in the central and southern Mississippi River Valley of the United States. A temperature of 100° F (37.8° C) has never been recorded. The average is 78° F (26° C), with a range between the coldest and warmest months of between 2° and 3° F (1.4° C). Rainfall and humidity are high. The year is divided into a very rainy season from January to June and a season of less rain for the rest of the year. Average annual rainfall is more than 59 inches (1,500 millimetres). Wherever there is exposure to the easterly trade winds the humidity is compensated for, and the climate is quite comfortable.

An enormous amount of water pours into the ocean through the State of Pará. The Amazon itself winds about on its floodplain, leaving a maze of abandoned channels in the form of oxbow lakes and an intricate crescentic pattern of levees and swamps, much like the Lower Mississippi in North America. At Óbidos the floodplain is scarcely a mile in width, but it opens out again downstream. The margins of the floodplain are marked by valley bluffs—steep slopes that rise abruptly about 150 to 200 feet above the river. The floodplain—which is somewhat less than 10 percent of the total area of Pará—is inundated each year during high-water stages, but the country above the valley bluffs is never flooded. The chief towns are located on the bluffs at places where the main channel of the river swings against the margin of the floodplain. The Amazon receives the water of several great tributaries. From west to east on the southern side are the Tapajóz, the Xingu, and the Tocantins. No large tributaries reach the Amazon on the northern side within Pará. At the mouth of the Amazon is the Ilha de Marajó (Island of Marajó). It is 150 miles (241 kilometres) long by 100 miles (161 kilometres) wide, with an area of 18,519 square miles (42,964 square kilometres). The northwestern part of the island is a part of the floodplain; but in the east and south the island is made up of higher ground. The Tocantins provides most of the water in the Rio Pará, southern side of the Ilha de Marajó, while the main stream of the Amazon passes to the north. The Rio Pará and the Amazon are connected by a network of side channels. Lying across the mouth of the main stream, north of the Ilha de Marajó, are two smaller islands that, like eastern Marajó, stand above the floods. They extend in a southeast to northwest direction: the Ilha Caviana is 48 miles (77 kilometres) long, and the Ilha Mexiana is 34 miles long.

Soils

The soils of much of Pará are not fertile. The alluvial material on the floodplain could yield abundant crops, especially rice, but the settlers along the Amazon have never made much use of it. Beyond the valley bluffs, where the land is never flooded and where mud and silt are never deposited, the soil has for a long time been exposed to the action of percolating rainwater. Water sinking into the soil carries the finer soil particles down far beyond the reach of shallow-rooted crops; it also dissolves and carries away the more easily soluble of the minerals. The result is a sandy, red-coloured soil made up largely of iron and aluminum compounds in which crops do poorly. Trees, with deeper roots, grow luxuriantly in the moist climate.

As for plant and animal life, except for a few patches of savanna, most of the state is covered with dense, tropical rain forest, or *selva*, with thousands of species of broadleaf evergreen trees. It is the soil under the *selva*, where little light reaches the ground, that is deeply leached and that, when the forest is cleared, quickly loses its capacity to produce crops. The native Indians used a small clearing only for a year or so before shifting to a new clearing. There are no large game animals such as are found in the tropical regions of Asia and Africa. The largest land animal of the *selva* is the tapir; there are many smaller animals, including several species of cat, and the area is rich in both kinds and number of birds, insects, and reptiles. The rivers abound with fish

and river turtles, and much of the Indians' food supply came from the Amazon itself—from fish, turtles, and turtle eggs.

The contemporary state. *Population.* The population of Pará was 2,197,072 in 1970, as compared with 1,550,935 in 1960 and 1,123,273 in 1950. Population density is about five persons per square mile, but at that is higher than that of the other states (Acre and Amazonas) and territories (Rondônia, Roraima, and Amapá) that make up the northern region of Brazil. As in the rest of the nation, the population is young.

Urban and rural settlements

The population is concentrated in the few cities and towns, which in turn are concentrated along the rivers. The largest is Belém, the capital, with 565,097 people, on the Rio Pará; others include Santarém, on the Tapajóz; Óbidos, on the Amazon; and Bragança, on the small river Caeté, near the coast. Aside from these, of which only Santarém has a population of more than 20,000, there are a few small settlements and trading posts on the principal rivers and tributaries, a few plantations, and small, scattered, tribal groups of Indians. Some are so remote and isolated that well into the second half of the 20th century they had still had no contact with modern civilization.

Ethnically the population is composed of people of European, Indian, and mixed European and Indian ancestry; and, since the 1930s, Japanese, who have settled in northern Pará.

Administration and social conditions. The state government and administration is centred in Belém, the capital. As has been the case with most of the other states of Brazil, it has had to rely heavily on the federal government for financial and other assistance. The federal government has assumed special interest and responsibility in the economic development of the Amazon Basin.

One indication of social conditions in an area is the average life-span of the inhabitants. In Pará it is 38 years. Health, education, and welfare programs in the cities are limited; outside the cities they are nonexistent.

Educational institutions

Belém is the leading educational centre of northern Brazil. Academic training is available in medicine, dentistry, pharmacy, law, engineering, and the fine arts. There is also a normal school to train teachers, an institute for research on tropical diseases, and an institute specializing in tropical agriculture.

Economy. In colonial times Pará (*i.e.*, Belém and its environs) prospered from the production of sugar and, later, of rice, cotton, and coffee, until other places in Brazil began to produce them more economically. For a while in the late 19th and early 20th centuries the economy was based on rubber. Since then it has been based primarily on the collection and export of other forest products, chiefly Brazil nuts, medicinal herbs, organic oils and insecticides, and fibres. Since World War II some plantation products have been introduced with considerable success by Japanese colonists, jute in increasing quantities along the Amazon River and black pepper just to the south of Belém and near Santarém in the north.

Transportation and communications. Transportation within the state and externally is almost entirely by water or air. The main port for Amazon River craft as well as for international and coastal shipping is Belém, and the Belém Airport is the principal air facility in northern Brazil. The only railway in the state runs from Belém to Bragança, a distance of 145 miles (233 kilometres).

The construction of the Belém do Pará–Brasília Highway, and the Transamazônica road running west from Belém to the Peruvian border, both built during the 1970s, are leading to a new tide of pioneer settlement and resource development in the hitherto most isolated parts of the Amazon Basin.

Cultural life. Cultural life is centred in Belém, to which most of the rest of the population has little access. Cultural institutions there include the Museu Paraense Emílio Goeldi, the Teatro da Paz, a classical theatre, and the Biblioteca e Arquivo Públivo do Pará (Public Library and Archives of Pará).

Future prospects. Two oil strikes in the Lower Amazon Valley have led to hopes that commercially important deposits of petroleum may be found. The Belém–Brasília Highway will provide a link with the centre of the country. Like the first Europeans to set eyes on that lush valley, in the 16th century, many 20th-century men instinctively feel that any place that is able to produce tropical growth in such rich profusion must be capable of producing wealth for man. Someday perhaps it will.

(P.E.J.)

Paracanthopterygii

The fishes that constitute the superorder Paracanthopterygii are a predatory, primarily marine group and modern one of about six major branches of the Teleostei, or bony fishes, the dominant living aquatic vertebrates. Some 1,160 living species of paracanthopterygian fishes have been described; they range in length from a few centimetres to about two metres (more than six feet).

In general body form there is considerable diversity, but ichthyologists have classed them as a discrete group, largely on the basis of a distinctive jaw musculature, on the structure of the caudal (*i.e.*, at the tail end) vertebrae, and the placement of the pelvic fins (they are usually in the midbody region or further toward the head).

The Paracanthopterygii comprises six orders: Batrachoidiformes, or toadfishes, about 45 species; Gadiformes, or codfishes, about 800 species; Gobiesociformes, or clingfishes, about 100 species; Lophiiformes, or anglerfishes, about 210 species; Percopsiformes, or trout-perches, about eight species; and Polymixiiformes, or beardfishes, three species. Most of the orders are primarily marine, with worldwide distribution; the percopsiforms, however, occur only in freshwaters of North America. Batrachoidiforms and gobiesociforms occur mainly in tropical and temperate shallow water along continental coasts and to a limited extent in freshwater. Gadiforms are represented by both shallow-water and deep-sea types. The most widely known gadiforms are the commercially important species and the only economically important paracanthopterygians: the true cods (*Gadus*); hakes (*Merluccius, Urophycis*); haddocks (*Melanogrammus*); pollocks (*Pollachius*); and whitings (*Merlangius*). All are abundant in waters of the continental shelf of the North Atlantic, where they have been commercially fished for centuries. Lophiiforms live on tropical reefs as well as in the ocean depths. Polymixiiforms occur at moderate depths in most warm seas, generally near continents.

Size range

The largest of the Paracanthopterygii are the codfishes, which grow to about two metres in length and attain weights that may exceed 90 kilograms (about 200 pounds). Certain goosefishes (Lophiiformes) reach a length of about two metres and a body weight of 35 kilograms (about 75 pounds); other lophiiforms are as small as 2.5 centimetres (about one inch) long. Batrachoidiforms grow to about 30 centimetres (one foot) in length, gobiesociforms to about eight centimetres (three inches). The largest percopsiforms are about 15 centimetres (six inches) long. Polymixiiforms reach no more than 30 centimetres in length.

NATURAL HISTORY

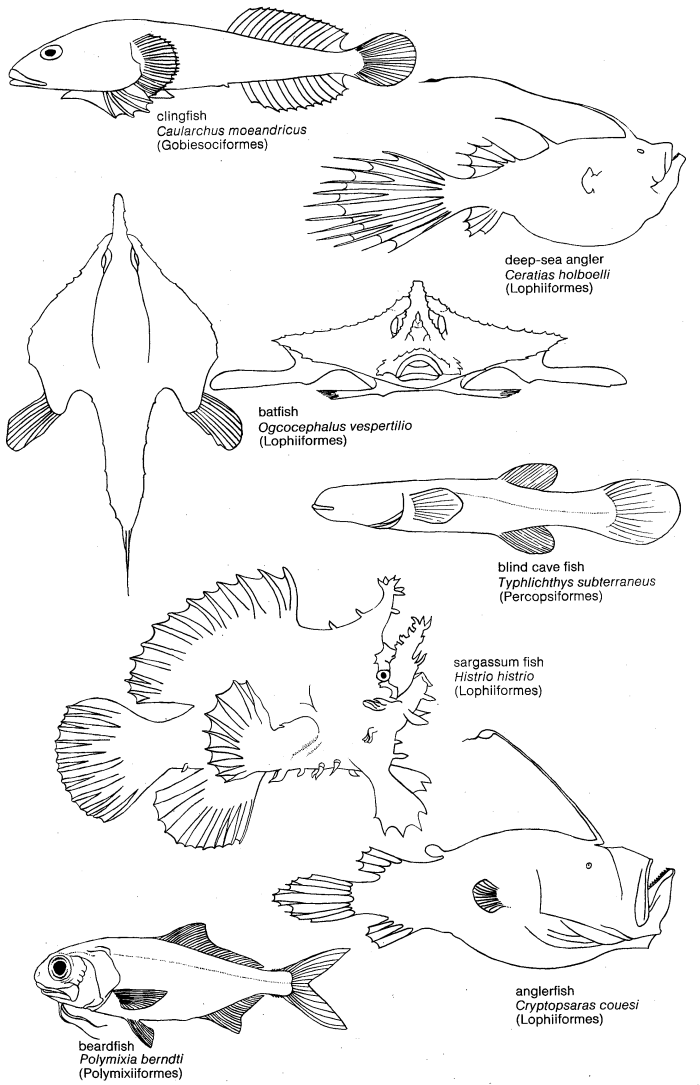
Life cycle and reproduction. Eggs of the oyster toadfish (*Opsanus tau*) of the western Atlantic—one of the most carefully studied batrachiforms—are laid in dark recesses of all sorts, including sunken tin cans and shoes. The male guards the eggs and young for about three weeks, after which the young fishes begin life on their own; some have been found living in living oysters. Luminous organs known as photophores, numbering several hundred and set in long horizontal rows, are believed to be sexual attractants in the midshipman (*Porichthys*)—so named because the organs resemble rows of bright buttons on a naval uniform. The northern midshipman (*P. notatus*), a common species on the eastern Pacific coast, spawns in shallow water, attaching its eggs to a rocky surface. The male guards the eggs. Like other batrachoidiforms, the midshipman lives and grows on the ocean bottom.

Most species of codfishes (which comprise some 70 species of Gadiformes) migrate over long distances. They gather in late winter and early spring to spawn, each species going to a particular area. The periodic movements are closely related to seasonal variations in water temperature. Fecundity of some codfish species is prodigious. The European ling (*Molva molva*) may deposit as many as 60,000,000 eggs each season. The eggs and larvae of most species are found in the plankton (*i.e.*, the aquatic organisms, collectively, suspended in the sea). Weeks or months elapse before the eggs hatch. Young codfishes are commonly found in very shallow water, but they move into deeper water as they become older. The eggs of grenadiers (family Macrouridae), a bottom-feeding group of cods, are believed to be laid near the bottom; the buoyant eggs rise part way to the surface. The larvae are known mainly from below 100 fathoms (about 180 metres, or 600 feet); older larvae occur at greater depths. In the Mediterranean pearlfish (*Carapus acus*), a member of another codlike group (family Carapidae), clumps of eggs, released by the female in late summer, appear at the surface and hatch into a specialized larva, the vexillifer, which lives amid the plankton. After attaining a length of about seven to eight centimetres (about three inches), it transforms to another larval stage, the tenuis, descends to the bottom, and becomes a parasite in a sea cucumber (*Holothuria tubulosa* or *Stichopus regalis*). The tenuis, apparently dependent upon its host for survival, undergoes a further transformation to the juvenile stage; in the process, its length decreases from 20 to 10 centimetres (eight to four inches). The Mediterranean pearlfish is believed to pass most of its life in the host. Very little is known of the general biology of reproductive habits of the brotulas and cusk eels (family Ophidiidae), also of the cod group. They are both oviparous (egg-laying) and viviparous (live-bearing). The males of some viviparous species produce spermatophores (sperm cases). The European eelpout (*Zoarces viviparus*) of the cod family Zoarcidae bears living young about five centimetres in length and numbering as many as 400. Fertilization is internal, and embryonic development occurs in the ovary of the female. Other eelpouts are believed to be live-bearing, but the ocean pout (*Macrozoarces americanus*) of the western Atlantic lays eggs that are guarded by one or both parents. The lophiiforms are primarily bottom fishes as adults, but many produce floating rafts of eggs. The eggs of the deep-sea anglerfishes (suborder Ceratioidei) are unknown; but it is believed that they float to the surface; the larvae occur in surface waters, gradually descending to deeper waters as they grow older. The females of the deep-sea anglers are from three to 13 times as large as the males. Females have an illicium, or "fishing pole," which is a modified spine of the dorsal, or back, fin that has moved forward onto the top of the head. At the tip of the illicium is a fleshy enlargement, the esca, used to lure prey within range of capture. (The illicium and esca are generally present also in male anglerfish other than in the Ceratioidei.) Commonly the esca is luminous; the female also has other light-producing organs. In 1922 a specimen of the anglerfish *Ceratias holboelli* was discovered; small specimens attached to its abdomen were thought to be its young. A few years later similar finds led to the discovery that the smaller fish were really mature males living parasitically on the female. Further investigation showed that the males, soon after their transformation from the larval state, bite onto an older, larger female, after which the female and male tissues unite; the separate circulatory systems join; and the male becomes a permanent appendage of the female.

Little is known of the reproductive habits of the gobiesociforms, percopsiforms, or polymixiiforms. Gobiesoci-

Parasitic habit of pearlfish

Parasitism of male anglerfish



Drawing by J. Helmer based on (cave fish) N.B. Marshall, *The Life of Fishes* (1965); Weidenfeld and Nicolson Co. Ltd., London; (all but *Polymixia*) David Starr Jordan, *A Guide to the Study of Fishes;* Holt, Rinehart & Winston, Inc.

Figure 1: Body plans of representative members of the orders Gobiesociformes, Lophiiformes, Percopsiformes, and Polymixiiformes.

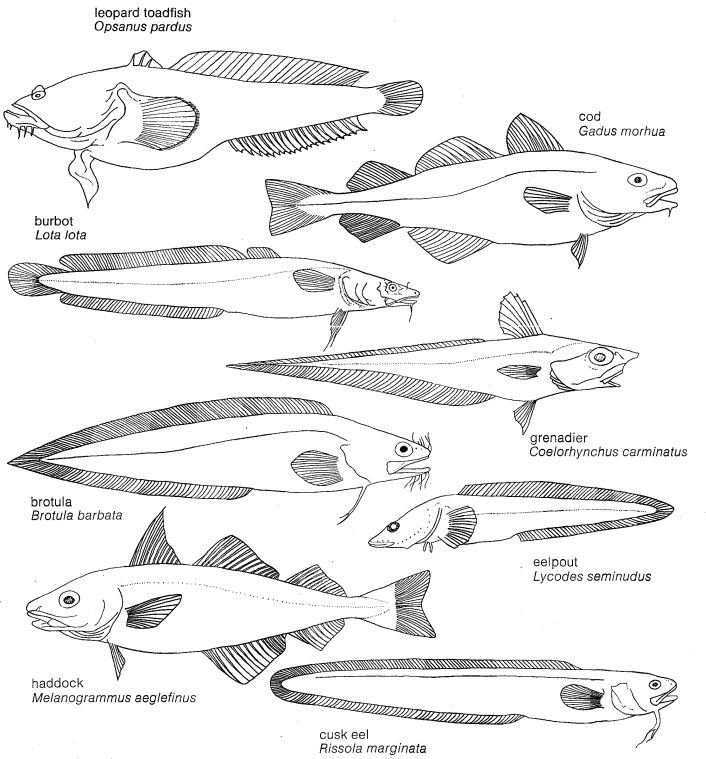


Figure 2: Body plans of representative members of the orders Batrachoidiformes (toadfish) and Gadiformes.

Drawing by J. Helmer from David Starr Jordan, *A Guide to the Study of Fishes;* Holt, Rinehart & Winston, Inc.

forms are known to lay eggs in shallow water, attaching them to rocks or plants; and percopsiforms are known to spawn in the spring of the year in shallow water.

Ecology and behaviour. All batrachoidiforms are bottom dwellers. True toadfishes (Batrachoidinae, about 25 species) occur in shallow or moderate depths along continental coasts; some ascend rivers. The oyster toadfish lives under rocks or amid debris, awaiting prey of almost any type, which is taken with a sudden snap. Venomous toadfishes (Thalassophryninae, about nine species) are restricted to the coasts and rivers of Central and South America. Because of their sluggish habits, these fishes are sometimes stepped on by man and can inflict painful wounds. Midshipmen (Porichthyinae, about 12 species), which are restricted to tropical and temperate coasts of the Americas, are unusual in being shallow-water fishes with photophores, a feature generally found in deepwater forms. Most midshipmen occur in depths of less than 50 fathoms (one fathom = six feet), and all are found in water shallower than 200 fathoms.

Gadiform fishes of the family Gadidae (about 70 species) are all marine species, except for the burbot (*Lota lota*); some, however, ascend rivers with the tides. Bottom dwellers, they occur on the continental shelves from shallow water to about 200 fathoms and, although distributed throughout the oceans, are most numerous in the eastern North Atlantic. Deep-sea cods (Moridae, about 70 species) are cold-water bottom fishes, living at greater depths along the continental slopes. Grenadiers (Macrouridae, about 300 species), typically bottom fishes, live along the continental slopes at depths of 100 to 1,000 fathoms. Few species are cosmopolitan in distribution, but the group as a whole is widely distributed in tropical and temperate latitudes. A few species of gadiforms (Muraenolepididae, four species) are confined to Antarctic seas, and, like the cods, they are bottom fishes, living at moderate depths. The pearlfishes (Carapidae, about 27 species) are marine, mainly tropical, shallow-water, eel-like fishes adapted to living inside the body of various invertebrates. They have been collected from a variety of hosts, including tunicates, oysters, and sea cucumbers. Brotulas and cusk eels (Ophidiidae, about 250 species) are mainly bottom dwellers. Some are shallow-water species with nocturnal habits, but the group as a whole is the dominant teleostean family at depths greater than 2,000 fathoms. Some have been taken at depths of about 4,000 fathoms, the greatest depth at which any form of fish life is known.

Eelpouts

Eelpouts (Zoarcidae, about 80 species) are bottom fishes, commonly occurring from shallow water to depths of 1,000 fathoms. Species are most abundant in the higher latitudes of both hemispheres, especially in the north. Eelpouts are common in shallow water of Arctic and Antarctic seas, and they constitute a significant part of the polar faunas.

Gobiesociforms (about 100 species) are mostly marine fishes, typically inhabiting the intertidal zone. Some species (*Diademichthys*) hide among the spines of sea urchins. In tropical America, four species (*Gobiesox*) are known from swift-flowing freshwater streams.

Of the lophiiforms, the ceratioids, or deep-sea anglerfishes, are the only abyssal (deep-sea) forms. They occur primarily at depths of 1,000 to 3,000 fathoms. Unlike other lophiiforms, they are midwater fishes, are uniformly black, and have no pelvic fins. They are apparently feeble swimmers, depending primarily on their light organs to attract prey. Some (*Melanocetus, Linophryne*) are known to swallow fishes several times their own length, accommodating them in a highly distensible stomach. Crustaceans and other invertebrates are also eaten. Many lophiiforms—so-called frogfishes (Antennariidae, about 60 species)—are shallow-water forms, commonly inhabiting coral reefs. Frogfishes typically have highly varied colour patterns, and some species are able to change colours. In habit they are sedentary but can use their fins to walk on the bottom and to climb over obstacles. The tropical sargassum fish (*Histrio histrio*), so called because it lives amid floating brown algae of the genus *Sargassum*, clings with prehensile (*i.e.*, adapted for seizing, or wrapping around) pectoral fins as it searches for prey, which is sucked into the mouth by the powerful jaws and expandable cheeks.

The lophiiform group known as goosefishes (Lophiidae, about 12 species) seldom occur in shallow water, preferring instead the moderate depths (10 to 500 fathoms) along the continental slopes of tropical and temperate regions. The batfishes (Ogcocephalidae, about 60 species), are mainly deepwater lophiiforms, but some (*Ogcocephalus, Halieutichthys*) are regularly found in water only a few feet deep. Like frogfishes, they walk on the bottom, using their pectoral fins. Batfishes are awkward swimmers and, when disturbed, tend to bury themselves in the bottom rather than swim away.

Fishes without eyes

Percopsiforms (about eight species) live under conditions of dim light. Cave fishes, with eyes reduced to nonfunctional rudiments, have elaborate systems of sense organs in the skin of the head, body, and tail; they live in total darkness. Because of the secretive habits of percopsiforms, little is known about species other than the trout-perch (*Percopsis omiscomaycus*), which is widely distributed in central North America and is abundant in some of the Great Lakes, where it occurs in clear water to a depth of about 35 fathoms. Polymixiiforms, numbering only three marine species, are generally found at depths of 150 to 350 fathoms.

FORM AND FUNCTION

Batrachoidiforms generally have two dorsal fins; a small anterior fin, usually with two spines; and a long posterior fin. In venomous species, the hollow fin spines form an efficient apparatus for the injection of venom. A similar spine is found on each cheek (operculum).

Among the gadiforms, the dorsal and anal fins of some deep-sea cods are distinctively arranged as three dorsals and two anals. This arrangement also occurs in some codfishes (Gadidae). The macrourids are characterized by a long, tapering tail. A tubular light organ containing luminescent bacteria is sometimes present along the ventral midline of both sexes. All but a few species have a well-developed swim bladder; in the males of some species and in some codfishes, the swim bladder is equipped with drumming muscles, indicating that sound can be produced. In the bregmacerotids and muraenolepidids there are two dorsal fins, with the anterior fin represented by a single ray. Brotulas may have the pelvic fins either present or absent, but cusk eels have them anterior in position, under the lower jaw. They are kept in continuous probing motion, as the fish swims just off the bottom, and aid in detecting food. Zoarcids (eelpouts) are elongated, eel-like fishes, with pelvic fins rudimentary or absent.

Suction disk for attachment to bottom

Gobiesociforms, with a depressed head, wide mouth, and tapering body, resemble toadfishes, but they are distinctive in having a prominent sucking disk on the ventral surface. The paired pelvic fins, thoracic in position, form part of the disk, various fleshy pads and folds forming the remainder. The disk allows clingfishes to hold fast to rocky bottoms amid the often turbulent wave action of their shallow-water environment.

Some lophiiforms are unique among teleostean fishes in having only two gills. The ogcocephalids are somewhat flattened anglers, in this respect resembling lophiids rather than the balloon-like antennariids; they are distinctive in having the illicium, when not in use, concealed in a tube (illicial cavity) between the eyes and over the mouth. Like most anglerfishes they lack typical scales but are distinctively equipped with bony tubercles (projections) and spines imbedded in the skin.

The polymixiiforms are singular in having a pair of fleshy barbels, or "whiskers," under the jaw. Each barbel is supported by three small bones.

EVOLUTION AND PALEONTOLOGY

Fossil batrachoidiforms include only material from lower Pliocene marine deposits (about 5,000,000–7,000,000 years old) of North Africa. These fossils are similar to a living species, *Batrachoides didactylus.*

Fossil gadiforms are relatively numerous and are known primarily from Tertiary marine deposits (about 2,500,000–65,000,000 years old) of the Northern Hemi-

sphere. A Paleocene fossil (54,000,000–65,000,000 years old) has been identified as a codlike fish; some Eocene fossils (38,000,000–54,000,000 years old) have been identified for the families Bregmacerotidae and Gadidae; and Oligocene-Miocene fossils (7,000,000–38,000,000 years old) for the families Bregmacerotidae, Gadidae, Macrouridae, and Ophidiidae. In addition, many fossil ear stones (otoliths) and scales, beginning with specimens from the Cretaceous (65,000,000–136,000,000 years ago), are similar to the Gadiformes. Fossil gobiesociforms are unknown. Fossil lophiiforms include two species from Eocene marine deposits of Europe and one species from Pliocene marine deposits of North Africa; one Eocene species has been identified as a goosefish (Lophiidae), the other as a frogfish (Antennariidae).

Fossil percopsiforms include three genera from Tertiary freshwater deposits of North America and one (*Sphenocephalus*) from Cretaceous marine deposits of Europe. Of the North American genera, two (*Amphiplaga, Erismatopterus* from the middle Eocene) have been identified as trout-perches (Percopsidae), and one (*Tricophanes*, Oligocene–Miocene) as a pirate perch (Aphredoderidae). The relationships of *Sphenocephalus* are obscure. Fossil polymixiiforms include a diversified group of about six genera known primarily from Cretaceous marine deposits of Europe and the Middle East; a few others are known from the Tertiary.

CLASSIFICATION

Annotated classification.

SUPERORDER PARACANTHOPTERYGII

Most with a distinctive type of jaw musculature (involving levator maxillae superioris muscle and associated structures); pelvic fins usually placed anteriorly, thoracic (midbody) or even further forward; primarily marine; worldwide distribution; about 1,600 living species.

Order Polymixiiformes (beardfishes)

Middle Cretaceous to Recent. Barbels supported by rays; spines on the dorsal and anal fins; pelvic fins subthoracic. Deepwater marine fishes; three species. Adult length about 30 cm.

Order Percopsiformes (trout-perches, pirate perches, and cave fishes)

Eocene to Recent. Mouth gape and buccal dentition reduced; median fin spines reduced or lost; head with spine ornamentation. About 8 living species, all freshwater; North America; length 8–15 cm.

Order Gadiformes (cods, cusk eels, pearlfishes, eelpouts and grenadiers)

Paleocene to Recent. Early gadiforms were similar in structure to early percopsiforms, but almost all remained marine and subsequently specialized into a variety of environments. Reduced caudal skeleton; elongate body; altered head and jaw structure. Very reduced fin spines; marine, worldwide. About 800 species. Length 7 to about 200 cm.

Order Batrachoidiformes (toadfishes)

Miocene to Recent. Bottom fishes with short, small, spinous dorsal fins; long soft-rayed dorsal fins; flat heads; about 45 species; marine, occasionally freshwater, shore fishes of tropics. Length to about 30 cm.

Order Lophiiformes (goosefishes, anglerfishes, frogfishes, and batfishes)

Eocene to Recent. Spinous dorsal fin modified as a movable lure. Some deep-sea forms with light organs and males parasitic on females. Marine, widespread; in shallow-water and deep-sea habitats. About 210 species. Length to about 200 cm.

Order Gobiesociformes (clingfishes)

Recent; flattened, depressed fishes with a ventral sucker formed of the pelvic fins and surrounding tissue; no spiny dorsal fin; about 100 species; marine and occasionally freshwater in tropics and along many temperate seacoasts.

Critical appraisal. The interrelationships of the groups listed here as paracanthopterygians are not yet well established, and the classification given here is provisional. There is considerable agreement that trout-perches (Percopsiformes) and cods (suborder Gadoidei) are closely related, and this agreement may be considered the basis of the group Paracanthopterygii. What other fishes should be included in the Paracanthopterygii is a question receiving continued study. Some ichthyologists have held that clingfishes (Gobiesociformes) are related to dragonets (Callionymidae); that eelpouts (Zoarcidae) are related to blennies (Bathymasteridae, Blenniidae, etc.); that brotulas, cusk eels, and pearlfishes (suborder Ophidioidei) are related to the river blackfish of Australia (the perchlike *Gadopsis*); that beardfishes (Polymixiiformes) are related to squirrelfishes and their relatives (Beryciformes); and that toadfishes (Batrachoidiformes) and anglers (Lophiiformes) also have their relationships within the great mass of perchlike fishes (Acanthopterygii). In addition, killifishes (Cyprinodontidae) and related live-bearers (Poeciliidae) are believed by some writers to be related to trout-perches (Percopsiformes) and by others to silversides, flying fishes, and their relatives (Atheriniformes). Thus, future study may result in the transfer of some groups from the Paracanthopterygii to the Acanthopterygii and vice versa.

BIBLIOGRAPHY

General: C.M. BREDER and D.E. ROSEN, *Modes of Reproduction in Fishes* (1966); E.S. HERALD, *Living Fishes of the World* (1961).

Faunal: J.E. BOHLKE and C.C.G. CHAPLIN, *Fishes of the Bahamas and Adjacent Tropical Waters* (1968); W.A. CLEMENS and G.V. WILBY, *Fishes of the Pacific Coast of Canada*, 2nd ed. (1961); A.H. LEIM and W.B. SCOTT, *Fishes of the Atlantic Coast of Canada* (1966); T.C. MARSHALL, *Fishes of the Great Barrier Reef and Coastal Waters of Queensland* (1964); Y. OKADA, *Fishes of Japan*, rev. ed. (1965); T.D. SCOTT, *The Marine and Freshwater Fishes of South Australia* (1962); J.L.B. SMITH, *The Sea Fishes of Southern Africa*, 4th ed. (1961); M. WEBER and L.F. DE BEAUFORT, *The Fishes of the Indo-Australian Archipelago* (1911–62), a multivolume work; A. WHEELER, *The Fishes of the British Isles and North-West Europe* (1969).

Specific: (*Batrachoidiformes*): B.B. COLLETTE, "A Review of the Venomous Toadfishes, Subfamily Thalassophryninae," *Copeia*, pp. 846–864 (1966); C.R. GILBERT, "Western Atlantic Batrachoidid Fishes of the Genus *Porichthys*, Including Three New Species," *Bull. Mar. Sci.*, 18:671–730 (1968). (*Gadiformes*): D.C. ARNOLD, "A Systematic Revision of the Fishes of the Teleost Family Carapidae (Percomorphi, Blennioidea), with Descriptions of Two New Species," *Bull. Br. Mus. Nat. Hist. (Zool.)*, 4:245–307 (1956); U. D'ANCONA and G. CAVINATO, *The Fishes of the Family Bregmacerotidae* (*Dana Rep.* 64) (1965); N.B. MARSHALL, "Systematic and Biological Studies of the Macrourid Fishes (Anacanthini-Teleostei)," *Deep Sea Res.*, 12:299–322 (1965); J.G. NIELSEN, "Systematics and Biology of the Aphyonidae (Pisces, Ophidioidea)," *Galathea Rep.*, 10:1–90 (1969); D.W. STRASBURG, "Description of the Larva and Familial Relationships of the Fish *Snyderidia canina*," *Copeia*, pp. 20–24 (1965); A.N. SVETOVIDOV, *Gadiformes* (1962; orig. pub. in Russian, 1948). (*Gobiesociformes*): J.C. BRIGGS, "A Monograph of the Clingfishes (Order Xenopterygii)," *Stanford Ichthyol. Bull.*, 6:1–224 (1955). (*Lophiiformes*): E. BERTELSEN, *The Ceratioid Fishes* (*Dana Rep.* 39) (1951); M.G. BRADBURY, "The Genera of Batfishes," *Copeia*, pp. 399–422 (1967); L.P. SCHULTZ, "The Frogfishes of the Family Antennariidae," *Proc. U.S. Natn. Mus.*, 107:47–105 (1957). (*Percopsiformes*): M.B. TRAUTMAN, *The Fishes of Ohio, with Illustrated Keys* (1957); L.P. WOODS and R.F. INGER, "The Cave, Spring, and Swamp Fishes of the Family Amblyopsidae of Central and Eastern United States," *Am. Midl. Nat.*, 58:232–256 (1957). (*Polymixiiformes*): E.A. LACHNER, "Populations of the Berycoid Fish Family Polymixiidae," *Proc. U.S. Natn. Mus.*, 105:189–206 (1955).

Systematic: W.A. GOSLINE, *Functional Morphology and Classification of Teleostean Fishes* (1971); D.E. ROSEN and C. PATTERSON, "The Structure and Relationships of the Paracanthopterygian Fishes," *Bull. Am. Mus. Nat. Hist.*, 141:357–474 (1969).

(G.J.N.)

Paracelsus

Known as Paracelsus the Great to his devoted followers and as the Medical Luther to his many enemies, Philippus Aureolus Theophrastus Bombast von Hohenheim was one of the most extraordinary and outrageous men of the Renaissance. He also had the characteristics of a natural extrasensory perceptionist and was perhaps best described in his own time as a "magician." As a physician, surgeon, and chemist, he stimulated the development of pharmaceutical chemistry by his discovery and use of

new medical remedies. These, however, were only by-products of his central teaching, which often upset his contemporaries and remains controversial today. His basic dictum, "Magick is a Great Hidden Wisdom—Reason is a Great Open Folly," best expresses his teaching, and his declaration, "Resolute Imagination can accomplish all things," reflects his magical presuppositions. He was always forthright, frequently vitriolic, scurrilous, scathing, and caustic.

Paracelsus, engraving by an unknown artist, c. 1600.
By courtesy of the Musee National Suisse, Zurich, Switz.

Paracelsus was born near the village of Einsiedeln (now in Switzerland), on November 10 (some say November 14), 1493, the only son of a somewhat impoverished German doctor and chemist. Theophrastus, as he was first called, was a small boy when his mother died; his father then moved to Villach in southern Austria. There the boy attended the Bergschule, founded by the wealthy Fugger family of merchant bankers of Augsburg, where his father taught chemical theory and practice. Youngsters were trained at the Bergschule as overseers and analysts for mining operations in gold, tin, and mercury, as well as iron, alum, and copper-sulfate ores. The young Paracelsus learned from miners' talk of metals that "grow" in the earth, watched the seething transformations in the smelting vats, and perhaps wondered if he would one day discover how to transmute lead into gold, as the alchemists sought. Thus Paracelsus early gained insight into metallurgy and chemistry that, doubtless, laid the foundations of his later remarkable discoveries in the field of chemotherapy.

In 1507, at the age of 14, he joined the many vagrant youths who swarmed across Europe in the late Middle Ages, seeking famous teachers at one university after another. During the next five years Paracelsus is said to have attended the Universities of Basel, Tübingen, Vienna, Wittenberg, Leipzig, Heidelberg, and Cologne but was disappointed with them all. He wrote later that he wondered how "the high colleges managed to produce so many high asses," a typical Paracelsian jibe.

Rejection of traditional education and medicine

His attitude upset the schoolmen. "The universities do not teach all things," he wrote, "so a doctor must seek out old wives, gipsies, sorcerers, wandering tribes, old robbers, and such outlaws and take lessons from them. A doctor must be a traveller, . . . Knowledge is experience." Paracelsus held that the rough-and-ready language of the innkeeper, barber, and teamster had more real dignity and common sense than the dry-as-dust scholasticism of Aristotle, Galen, and Avicenna, the recognized Greek and Arab medical authorities of his day.

Paracelsus is said to have graduated from the University of Vienna with the baccalaureate in medicine in 1510, when he was 17. He was, however, delighted to find the medicine of Galen and the medieval Arab teachers criticized in the University of Ferrara, where, he always insisted, he received his doctoral degree in 1516 (university records are missing for that year). At Ferrara he was free to express his rejection of the prevailing view that the stars and planets controlled all the parts of the human body. Against this astrological determinism, he declared:

> Man is a star. Even as he imagines himself to be, such he is. He is what he imagines. . . . Man is a sun and a moon and a heaven filled with stars. . . . Imagination is Creative Power. Medicine uses imagination strongly fixed. Phantasy is not imagination, but the frontier of folly. . . . *Because Man does not imagine perfectly at all times, arts and sciences are uncertain, though, in fact, they are certain and, by means of imagination, can give true results. Imagination takes precedence over all.*

He is thought to have begun using the name "para-Celsus" (above or beyond Celsus) at about that time, for he regarded himself as even greater than Celsus, the renowned 1st-century Roman physician.

Clearly a man of this type could never settle for long in any seat of learning, and so, soon after taking his degree, he set out upon many years of wandering through almost every country in Europe, including England, Ireland, and Scotland. He then took part in the "Netherlandish wars" as an army surgeon, at that time a lowly occupation. Later he went to Russia, was held captive by the Tatars, escaped into Lithuania, went south into Hungary, and again served as an army surgeon in Italy in 1521.

Ultimately his wanderings brought him to Egypt, Arabia, the Holy Land, and, finally, Constantinople. Everywhere he sought out the most learned exponents of practical alchemy, not only to discover the most effective means of medical treatment but also—and even more important—to discover "the latent forces of Nature," and how to use them. He wrote:

> He who is born in imagination discovers the latent forces of Nature. . . . Besides the stars that are established, there is yet another—*Imagination*—that begets a new star and a new heaven.

After about ten years of wandering, he returned home in 1524 to Villach to find that his fame for many miraculous cures had preceded him. When it became known that the Great Paracelsus, then aged 33, had been appointed town physician and lecturer in medicine at the University of Basel, students from all parts of Europe began to flock into the city. Pinning a program of his forthcoming lectures to the notice board of the university on June 5, 1527, he invited not only students but anyone and everyone. The authorities were scandalized and incensed by his open invitation. Ten years earlier Luther had nailed his Theses on Indulgences to the doors of the Wittenberg Schlosskirche. Later, Paracelsus wrote:

Height of his career

> Why do you call me a Medical Luther? . . . I leave it to Luther to defend what he says, and I will be responsible for what I say. That which you wish to Luther, you wish also to me: you wish us both in the fire.

Three weeks later, on June 24, 1527, surrounded by a crowd of cheering students, he burned the books of Avicenna, the Arab "Prince of Physicians," and those of the Greek physician Galen, in front of the university. No doubt his enemies recalled how Luther, just six and a half years before at the Elster Gate of Wittenberg on December 10, 1520, had burned a papal bull that threatened excommunication. Paracelsus seemingly remained a Catholic to his death, although it has been said that his books were placed on the *Index Expurgatorius*. Like Luther, he also lectured and wrote in German rather than Latin, for he loved the common tongue.

Despite his bombastic blunders—the adjective indeed derives from his surname Bombast—he reached the peak of his tempestuous career at Basel. His name and fame spread throughout the known world, and his lecture hall was crowded to overflowing. He stressed the healing power of nature and raged against those methods of treating wounds, such as padding with moss or dried dung, that prevented natural draining. The wounds must drain, he insisted, for "If you prevent infection, Nature will heal the wound all by herself." He attacked venomously many other medical malpractices of his time and jeered mercilessly at worthless pills, salves, infusions, balsams, electuaries, fumigants, and drenches, much to the delight of his student-disciples.

Paracelsus' triumph at Basel lasted less than a year, however, for he had made too many enemies. By the spring of 1528, he was at loggerheads with doctors, apothecaries, and magistrates. Finally, and suddenly, he had to flee for his life in the dead of night. Alone and penniless he wandered toward Colmar in Upper Alsace, about 50 miles north of Basel. He stayed at various places with friends. Such leisurely travel for the next eight years allowed him to revise old manuscripts and to write new treatises. With the publication of *Die grosse Wundartzney* ("Great Surgery Book") in 1536 he made an astounding comeback; this book restored, and even extended, the almost fabulous reputation he had earned at Basel in his heyday. He became wealthy and was sought by royalty.

In May 1538, at the zenith of this second period of notoriety, he returned to Villach again to see his old father, only to find that he had died four years previously. On September 24, 1541, Paracelsus himself died in mysterious circumstances at the age of 48 at the White Horse Inn, Salzburg, where he had taken up an appointment under the prince-archbishop, Duke Ernst of Bavaria.

Assessment

His medical achievements were outstanding. In 1530 he angered the city council of Nuremberg by writing the best clinical description of syphilis up to that time, maintaining that it could be successfully treated by carefully measured doses of mercury compounds taken internally, thus foreshadowing the Salvarsan treatment of 1909. He stated that the "miners' disease" (silicosis) resulted from inhaling metal vapours and was not a punishment for sin administered by mountain spirits. He was the first to declare that, if given in small doses, "what makes a man ill also cures him," an anticipation of the modern practice of homeopathy. Paracelsus is said to have cured many persons in the plague-stricken town of Stertzing in the summer of 1534 by administering orally a pill made of bread containing a minute amount of the patient's excreta he had removed on a needle point. He was the first to connect goitre with minerals, especially lead, in drinking water. He prepared and used new chemical remedies, including those containing mercury, sulfur, iron, and copper sulfate, thus uniting medicine with chemistry, as the first *London Pharmacopoeia*, in 1618, indicates. Paracelsus, in fact, contributed substantially to the rise of modern medicine, including psychiatric treatment. Carl Gustaf Jung, the psychiatrist, wrote of him that "We see in Paracelsus not only a pioneer in the domains of chemical medicine, but also in those of an empirical psychological healing science." (J.G.H.)

BIBLIOGRAPHY. *The Hermetic and Alchemical Writings of Aureolus Philippus Theophrastus Bombast of Hohenheim, Called Paracelsus the Great*, trans. by A.E. WAITE, 2 vol. (1894), contains a useful biographical preface and full text of the 28 principal works of Paracelsus—now long out of print. FRANZ HARTMANN, *The Life of Philippus Theophrastus Bombast of Hohenheim, Known by the Name of Paracelsus, and the Substance of His Teachings* (1887), is a useful biographical outline, with good translations of extracts from the main works. See also A. STODDART, *The Life of Paracelsus, Theophrastus von Hohenheim* (1911), a clear biographical outline, with a popular summary of his writings; BASILIO DE TELEPNEF, *Paracelsus: A Genius Amidst a Troubled World* (1945), a concise biographical essay and an outline of his teaching, notes, and a map of his travels, based upon research to 1945; JOHN G. HARGRAVE, *The Life and Soul of Paracelsus* (1951), an attempt to correct certain misconceptions regarding the outlook and teaching of this extraordinary genius; HENRY M. PACHTER, *Paracelsus: Magic into Science* (1951); WALTER PAGEL, *Paracelsus* (1958), an analysis of antecedents of Paracelsus, his thought and influence; and ALLEN G. DEBUS, *The English Paracelsians* (1965), a study of the influence of Paracelsus on English thought in the years after his death.

Paraguay

Paraguay is an independent republic of South America, located in the south central part of the continent. A landlocked country, it is dwarfed by the larger bordering countries of Bolivia to the northwest and north; Brazil to the east; and Argentina to the southeast, south, and west. It has an area of 157,048 square miles (406,752 square kilometres) and a population of about 2,400,000. The national capital of Asunción is located at the confluence of the Paraguay and Pilcomayo rivers. Rivers play a vital role in the nation's economic life, providing the country with access to the distant Atlantic Ocean. Indeed, the name of the country is said to derive from the word *Paraguay* which, in the Guaraní Indian language, could be defined as a place with a great river.

The many wars in which Paraguay has been engaged in the past have gained for its people a high reputation for courage. Yet the ravages of war have left their mark upon national life—Paraguay's population is now small, its economy is underdeveloped, and its politics centralized. The tensions that exist between the European and Indian populations in other Latin-American countries do not, however, exist in Paraguay. Its people are largely persons of mixed Spanish and Indian ancestry and are generally bilingual in both Spanish and Guaraní. National problems are, therefore, often more economic than social in nature. The economy is based upon agriculture and forestry and is in need of diversification and industrial development. (For history, see PARAGUAY, HISTORY OF. For related physical features, see GRAN CHACO; PARAGUAY RIVER; PARANA RIVER.)

THE LANDSCAPE

Natural features. *Relief.* The Paraguay River, which runs through the country from north to south, divides it into two distinct geographical regions—the Región Oriental (Eastern Region) and the Chaco Boreal (Northern Chaco).

Región Oriental and the Chaco Boreal

The Eastern Region, comprising about a third of the country, is an extension of the Brazilian Plateau and lies at a height of between 1,000 and 2,000 feet above sea level. The Cordillera de Amambay (Amambay Mountains) runs approximately north to south along part of the frontier with Brazil. From the northeast, the Montañas de Aracanguy extend roughly southward in the direction of Encarnación, diminishing to hills in their southern range. To the east of these mountains runs the Alto Paraná (Upper Paraná) River Valley. (The Paraná forms both the eastern and southern borders of the country.) To the west lies the broad valley of the Paraguay River. The western part of the Oriental is the most favourable to human settlement. It contains the Ypoá and Ypacaraí lakes in the south.

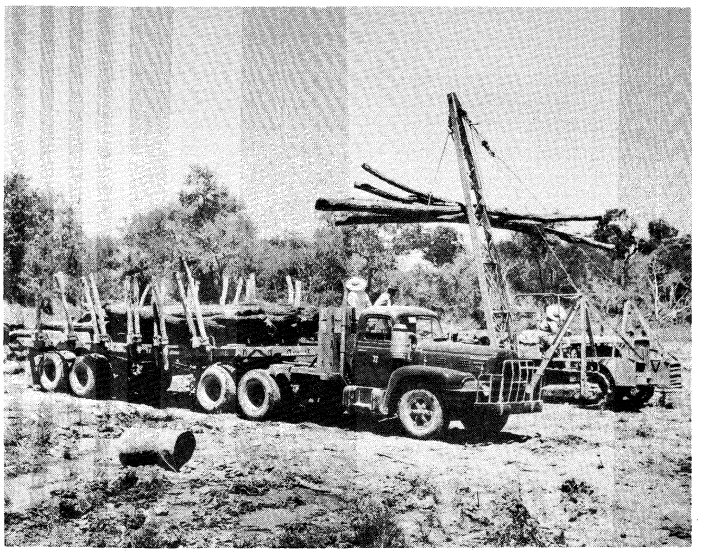
The Chaco Boreal, which covers 95,000 square miles—about two-thirds of the country—forms part of the Gran Chaco (*q.v.*), a flat and largely featureless tropical region that also extends into Bolivia and Argentina.

Drainage. Four-fifths of the country's perimeter is traced by the Paraguay, Apa, Alto Paraná, and Pilcomayo rivers. Multiple tributaries cross the eastern and central regions. The mountain ranges of Amambay and Mbaracayú form the watershed between the Paraguay and the Alto Paraná rivers that join to form the Paraná at the country's southernmost corner. Important eastern tributaries of the Paraguay River include the Jejuy, the Apa, the Tebicuary, and the Aquidabán. The only important tributary flowing from the west is the Pilcomayo, which joins the Paraguay at Asunción. It forms the southern border of the Chaco Boreal and is the main river of the area. The other rivers of the Chaco are slow and sluggish; they often drain into marshes or disappear completely in dry areas.

The river system

The largest lake in Paraguay, Ypoá is located 40 miles south of Asunción. Its waters are dispersed through various channels into the Río Tebicuary and the marshlands of the Ñeembucú. A smaller lake, Ypacaraí, is situated in a valley of the Cordillera de Los Altos, 18 miles east of the capital. It gives rise to the Río Salado, which flows into the Paraguay River.

Soils. The eastern third of the country is composed of crystalline rocks that are largely covered by red sandstones, lavas, and other volcanic rocks. In the Alto Paraná Basin there is a layer of *terra rossa*, or red ochre. The iron-bearing soils of the region contain red and yellow ochres and black clays. The forested areas are enriched by humus. The Paraguay River Basin is covered with a



Workmen loading quebracho logs obtained from the eastern Chaco Boreal region, Paraguay.
By courtesy of the Organization of American States

layer of fertile alluvial soil, interrupted by outcroppings of the underlying crystalline formations. The soils of the Chaco are composed of alluvial mud, clay, and sand brought down from the Andes.

Climate. The climate does not resemble that of other tropical countries. There is little difference between the seasons, and the days of extreme temperatures are few. During the summer months between October and March, the temperature extremes range between 77° and 104° F (25° and 40° C) whereas in the winter months from April to September the range is between 50° and 68° F (10° and 20° C). Annual average rainfall decreases gradually from east to west, from about 50 inches in the Apa River region to about 32 inches in the Chaco. In the west, the low rainfall combines with a high evaporation rate and the porosity of the soil to produce periodic droughts that are often followed by floods that are caused by the periodic rise in the waters of the main rivers.

The tropical forests

Vegetation and animal life. The far eastern region is densely forested with such tropical hardwoods as the urunday, cedar, curupay (family Mimosaceae), and lapacho (family Bignoniaceae), which are used for building materials. There is also a small supply of softwoods. Between the low eastern hills and the Paraguay River, well-watered and often marshy land is covered with tall grasses; forests grow on the uplands and along the river valleys. The river valleys of the Chaco are marshy and sporadically forested. There are scattered stands of the quebracho, or axbreaker, a tree with red wood that yields a valuable tanning extract. To the west, the land is dotted with scrub thorn trees, dwarf shrubs, and giant cacti. There are also stands of the tall caranday palm; and the *palo santo*, or "holy wood," that contains a valuable oil; as well as pineapples, which are native to Paraguay; *guaimipiré*; and sugarcane.

Medicinal plants—including the paratodo tree that yields quinine—are found in profusion. Yerba maté, or the Paraguayan tea plant, is a form of holly that grows wild in the northeast and is cultivated in the southwest.

The name Chaco means hunting ground, and the area teems with wildlife. The jaguar, locally known as *tigre*, lives among herds of wild boar, water hog, and deer. The armadillo and anteater are common, as is the otter and the coypu—a South American aquatic rodent. The varied birdlife includes ibis, herons, toucans, muscovy ducks, doves, partridges, parakeets, and parrots. Insect life is extensive and includes the deadly tarantula spider.

The animal life of the east is similar to that of Brazil and Argentina. There are various monkeys, tapirs, peccaries, deer lions, and jaguars, as well as foxes, weasels, and otters. The more than 400 species of birds include eagles, falcons, partridge, pigeons, and herons. The waters abound with crocodiles, caymans, and a large variety of fish.

Settlement patterns. *Rural settlement.* Almost the entire rural population lives within the 50-mile-wide corridor between the Paraguay River and the eastern hills. There are six principal areas of settlement within the region. These are the northern zone in the triangle between the Paraguay, Apa, and Ypané rivers; the Cordillera region associated with the Salado, Pirayú, and Caañabé rivers; the Guaireña region watered by the Caaguazú, Ybytimí, and Tebicuary rivers; the Misiones zone through which flow the Ypoá and Alto Paraná rivers; the Ñeembucú region traversed by the Paraguay, Paraná, and Tebicuary rivers; and the area around Encarnación.

Settlement in the Chaco is sparse. Indigenous tribes devote themselves to hunting and fishing, and there are several agricultural religious colonies.

Urban settlement. The largest urban concentration is found in the national capital, Asunción, and its major suburb, Lambaré. The five departmental capitals—Concepción, Villarrica, Coronel Oviedo, Ñeembucú (Pilar), and Encarnación—each have populations of about 20,000. Caacupé, east of Asunción, has over 7,500 residents, but most of the other cities are small and serve as administrative or marketing centres.

THE PEOPLE

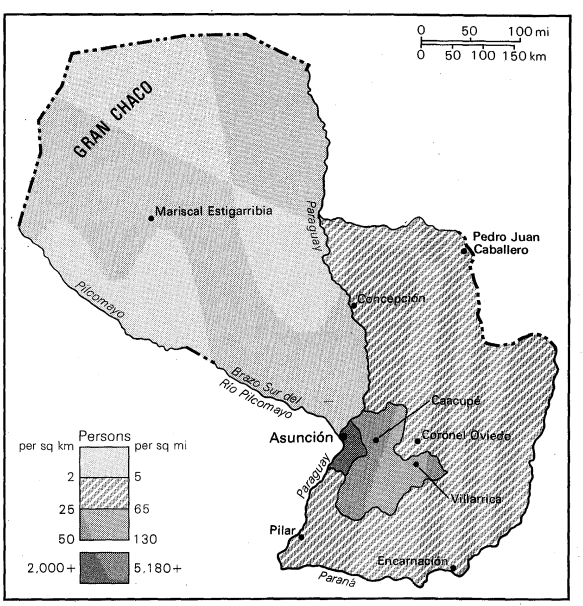
Population groups. *Ethnic composition.* Paraguay has the most homogeneous population in South America. About 75 percent of its inhabitants are native Paraguayans who are mainly of Guaraní descent, but who have at least one Spanish ancestor.

The immigrant population is small and generally assimilated. Immigrants came largely from western Europe, but Japan, Korea, and Australia are also represented. Assimilation has been resisted by the colonies of Mennonites (a Christian sect noted for its simple life-style and for its rejection of public and military service) and of Hutterites (a Christian sect that stresses communal living and communal ownership of property). These sects have been granted a special legal status exempting them from military service and awarding them their own legal jurisdiction.

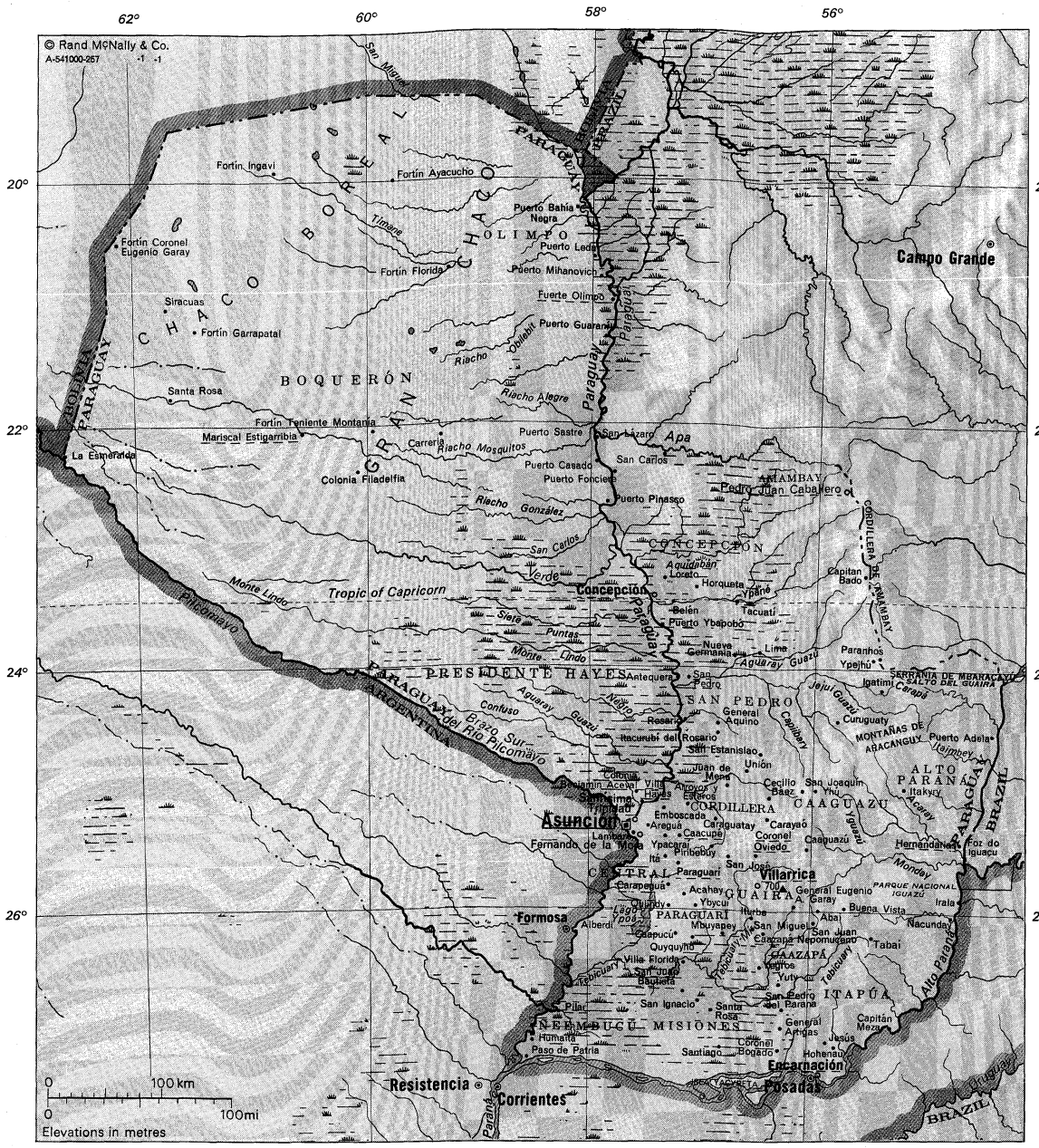
The religious colonies

There are only about 65,000 Indians living in the country. Small Indian tribes such as the Tapietés and the Ñanaguás inhabit the Chaco, as do the Matacos, Lenguas, Sanapanás, Chamacocos, Moros, Chulupies, and Macás.

Language groups. Spanish and Guaraní are the official languages of the country. Spanish is the medium of instruction, business, and government, whereas Guaraní is



Population of Paraguay.



PARAGUAY

MAP INDEX

Political subdivisions

Cities and towns

Paraguay, Area and Population

	area		population	
	sq mi	sq km	1962 census	1970 estimate
Regions				
Occidental				
Departments				
Boquerón	64,877	168,030	40,000	47,000
Olimpo	7,882	20,415	4,000	5,000
Presidente Hayes	22,579	58,480	30,000	42,000
Oriental				
Departments				
Alto Paraná	7,817	20,247	24,000	63,000
Amambay	4,993	12,933	35,000	68,000
Asunción	77	200	289,000	437,000
Caaguazú	8,345	21,613	125,000	227,000
Caazapá	3,666	9,496	92,000	106,000
Central	1,024	2,652	229,000	256,000
Concepción	6,970	18,051	86,000	111,000
Cordillera	1,910	4,948	188,000	200,000
Guairá	1,236	3,202	115,000	133,000
Itapúa	6,380	16,525	150,000	202,000
Misiones	3,025	7,835	59,000	74,000
Ñeembucú	5,345	13,868	58,000	70,000
Paraguarí	3,187	8,255	203,000	228,000
San Pedro	7,723	20,002	92,000	128,000
Total Paraguay	157,048*	406,752	1,819,000	2,396,000*

*Figures do not add to total given because of rounding.
Source: Official government figures.

mainly the language of the home. Most of the population is bilingual.

Religious groups. Roman Catholicism is the official religion, and its adherents constitute the majority of Paraguayans. There is tolerance for other Christian sects, as well as for other religions that are practiced by about 5 percent of the population.

Demography. In 1962 the population numbered about 1,800,000, and by 1970 it was estimated to have grown to 2,400,000. Most of the population is rural, and there are more women than men. Although life expectancy at birth is about 60 years of age, the population is young, and over 70 percent is under 30 years old. Infant mortality is high, with about one child out of every eight dying before it is one year old.

Population distribution

The mean density of population is 15.3 persons per square mile. Distribution is uneven, and the highest density occurs at Asunción. The Chaco Boreal has only about 1.3 persons per square mile.

The rate of emigration has been high since the 1940s. It has been estimated that between 1940 and 1970 about 400,000 Paraguayans have fled to Argentina and Brazil in search of more political freedom or better economic opportunities. Immigration has remained sparse and restricted and is far from compensating for the loss of population.

THE ECONOMY

Although there is a favourable balance of trade, with exports exceeding imports, the gross national product is small and average per capita income is low. The country has few developed mineral resources, and its economy is dependent upon agricultural and forestry products.

Distribution of resources. *Minerals.* Most mineral deposits are found east of the Paraguay River. Manganese is located near Emboscada, malachite (a green ore of copper) and azurite (a blue ore of copper) are found near Caapucú, feldspar and mica near Concepción, and talc and *pirofilita* (hard, iron-bearing flagstone) near Caapucú and San Miguel. Ochre is found in the Cordillera region, and gypsum and limestone near the Paraguay River. Other minerals include marble, iron, and salt.

Biological resources. Biological resources include hardwood forests and stands of quebracho trees. There are over 6,000,000 head of cattle, as well as horses, sheep, pigs, and goats. Fish are plentiful in the rivers and marshes and include shad, mackerel, and flounder.

Power resources. There are three main sources of hydroelectric power—the tributaries of the Paraguay River, the tributaries of the Alto Paraná River, and the Alto Paraná River itself. The tributaries of the Paraguay are largely located in the north of the eastern region. They flow down from a height of about 165 feet above the middle watermark of the Paraná and are broken by waterfalls and rapids. The Acaray and the Monday rivers, tributaries of the Alto Paraná, have a total power potential of 400,000 kilowatts, and the power station operation on the Acaray generates 90,000 kilowatts.

The hydroelectric potential

In the Alto Paraná, the Guairá Falls on the Brazilian border and the Apipé Rapids on the border with Argentina have a potential of 10,000,000 and 2,000,000 kilowatts, respectively, but are located far from the main centre of power consumption at Asunción.

Sources of national income. *Agriculture, forestry, and fishing.* Agriculture represents the most common economic activity and employs over half of the working population on 2 percent of the land. Most of the farmers are squatters on land held by absentee landlords, and they supplement their income by working on ranches or trade. Cassava and maize (corn) are the most common crops; and rice, tapioca, sugarcane, and wheat are also widely grown. Tobacco and cotton, tung oil, peanuts, and other oil seeds are grown for export. Yerba maté, coffee, and a variety of fruits are raised mainly for domestic use.

Cattle raising is a traditional activity that contributes to both domestic consumption and to the export trade. The cattle are raised in the southern district of Misiones, in the central region near Concepción, and in the eastern and northeastern Chaco. Almost all cattle are raised for slaughter; the dairy industry is little developed.

Forestry products

Forestry accounts for about one-fourth of the value of exports, and large quantities of wood are used for the generation of electricity in thermal power stations. Forestry is extensive, but further development is hampered by inadequate transportation facilities and the lack of sufficient machinery.

Fishing is carried out on the country's rivers and marshes on an individual basis. There is no organized fishing industry.

Mining and quarrying. Despite the varied mineral deposits already described, mining and quarrying are the least developed economic activities. They are hampered by the remote location of most ore deposits. Much of the existing activity is directed toward the production of clay

domestic utensils and jewelry, construction materials such as limestone and cement, and lime.

Manufacturing. Paraguay is one of the most industrially undeveloped countries of South America. Most manufacturing is located at Asunción, although some processing plants are located near sources of raw materials. Most manufacturing consists of the production of such extractive products as canned meat, tannin extract, vegetable oils, petitgrain oil (derived from the bitter-orange tree), ginned cotton, and tea. There are, however, growing textile and consumer-goods industries; and shoes, sugar, cigarettes, soap, cement, furniture, plywood, and matches are also produced.

Energy. Most electricity is produced in the wood- and oil-burning thermoelectric plants at Asunción. The hydroelectric plant on the Acaray River is also important, and in the late 1960s a total of over 200,000,000 kilowatt-hours of electricity were produced.

Financial services. The two main banking establishments are the Central Bank, which handles all monetary and foreign exchange functions, and the National Development Bank, which grants credits to large and small agricultural and industrial enterprises and conducts normal banking functions. There are also branches of other Latin-American, European, and United States banks. Foreign exchange reserves are limited, but loans are obtained from international development banks and associations.

Foreign trade. The principal exports—processed meat, forest products, vegetable oils, and tobacco—are chiefly marketed in Argentina, the United States, the United Kingdom, western Europe, and Uruguay. The main sources of imports—the United States, Argentina, West Germany, and the United Kingdom—provide Paraguay with heavy machinery, motors, transportation equipment, foodstuffs, fuel and lubricants, and iron and iron products. Paraguay is a member of the Latin American Free Trade Association (LAFTA). An elaborate system of subsidies, surcharges, and multiple exchange rates ensures a favourable balance of trade. Many goods, however, escape customs control along the country's river boundaries.

The government's role in the economy

Management of the economy. Although private initiative is encouraged, the government plays an important role in the planning of economic activities. It defines economic and social policy and participates directly in business life. The government holds a monopoly on the production of alcohol and has nationalized most of the transportation and the telecommunications services.

Taxation. Most of Paraguay's revenue is derived from import duties and from income, inheritance, and real estate taxes. Profits derived from commerce, industry, finance, pastoral farming, and real estate are taxable, but those from arable farming, the professions, and manual labour are exempt. There is little indirect taxation.

Trade unions. There are two large trade unions. The Confederation of Paraguayan Workers and the Syndicalist Movement of Paraguay are centred in Asunción. Strikes are illegal, and labour disputes are resolved by intervention of the National Council of Economic Coordination.

Economic policies and problems. Economic policy is directed toward the maintenance of a favourable balance of trade while improving agricultural and industrial production. Foreign loans are often floated, and the public debt continues to increase. Domestic inflation is serious, and the value of services has increased more rapidly than the production of goods. The country is attempting to reduce its dependence upon Argentina and Brazil, although it receives preferential treatment from LAFTA as an underdeveloped country.

Transportation. *Roads.* Economic expansion has been hampered by the lack of an adequate system of roads; the government, however, has sponsored highway construction since the early 1960s. Paraguay's link in the Pan-American Highway runs east from Asunción to Puerto Stroessner on the Alto Paraná River, where it is connected by bridge to the town of Foz de Iguacu, Brazil, and the highway to the Atlantic Ocean port of Paranagua, Brazil. The Chaco is served by a highway that runs northwestward from Asunción through Filadelfia and Palmar Ustares to the northeastern Bolivian border. Another road runs north from the national capital through Concepción to the Bolivian border at Pedro Juan Caballero.

Railways. Most of the railway system is composed of government-owned routes. The Paraguayan Central Railway operates between Asunción and Encarnación on the Alto Paraná in the southeast and connects with a train ferry to Posadas, Argentina. The Northern Railway runs a line between Concepción and Horqueta, 33 miles to the north. The Chaco is served by privately owned industrial railways.

Water transport. The chief mode of transportation is by water. The Alto Paraná is the artery for trade with Brazil and Argentina, and the Paraguay is the commercial highway for the port of Asunción. International traffic includes Paraguayan, Argentinian, Dutch, and British shipping companies.

The national merchant marine was founded in 1945 and began operations three years later. It owns cargo vessels, oil tankers, refrigerated ships, and passenger ships.

Air services. Air transport is provided by the government-owned national airline, the air force-operated military air transport, and two private companies. Because of its strategic location, Asunción is an important international junction, and ten foreign airlines use its airport. There are six domestic airports at Asunción, Pilar, Concepción, Encarnación, Coronel Oviedo, and Mariscal Estigarribia, as well as several private airstrips.

ADMINISTRATION AND SOCIAL CONDITIONS

The presidency

Structure of the government. *Constitutional framework.* The president is elected by direct vote for a five-year term. He must be a Catholic, be a Paraguayan by birth, and be over 40 years of age. The president is accorded wide powers; these include the appointment of ministers and other administrators, the conduct of foreign affairs, the declaration of a state of siege, the dissolution of the legislature, the periodic enactment of decrees, and the mobilization of troops.

The Council of State is an advisory body composed of nine members, all of which are appointed by the president. Four of its members are representatives for agriculture, manufacturing, commerce, and labour. The council's functions include the approval of measures initiated by the president.

The legislature is composed of a 60-member Chamber of Deputies and a 30-member Chamber of Senators. The legislators are chosen by popular vote for a term of five years. They are empowered to initiate legislation and may override a presidential veto or reject a budget by a vote of two-thirds majority in each house.

Local government. The country is divided into two provinces—the Occidental, to the west of the Paraguay River, and the Oriental region, to the east. Occidental Province is divided into three large departments, whereas the heavily populated Oriental Province is divided into 13 departments. The capital, Asunción, together with the suburb of Lambare, is administered separately. Each department is further divided into *partidos*, or rural districts, that are controlled by the central government. *Partidos* are composed of a number of still smaller units called *compañías* that often correspond to areas inhabited by kin groups. The *compañías* often have Guaraní names.

The political process. *Elections.* Voting is compulsory for all men and women who are 18 years of age and older. Resident aliens are also allowed to vote in municipal elections. Elections for the executive and legislative branches of the central government are, as mentioned, held every five years. If the president dissolves the legislature, elections must be held within two months.

Political parties. Paraguay is constitutionally a multiparty state. Parties are legally free to organize, but they are closely regulated by the executive branch of the government. Parties can be prohibited if they attempt to commit a crime against the regime in power, and they are not allowed to align themselves with, or accept aid from, foreign nations.

The political process is, in effect, controlled by the party of the chief executive. Other parties are recognized, however, in the composition of the legislature. The party that wins a simple majority is awarded two-thirds of the seats of both houses, and the rest of the seats are proportionately distributed among the minority parties. While giving nominal recognition to the legal existence of other organizations, this process ensures executive control of the political life of the country.

The Supreme Court

Justice. The Supreme Court is composed of five members who are chosen by the president and confirmed by the Council of State. Judges of the lower courts and magistrates are appointed by the Supreme Court. The judiciary is supervised by the Ministry of Justice, and judges can be tried by the legislature. Judges are not allowed to hold any other position except that of a teacher. Magistrates, who are often political functionaries, are granted legal immunity.

The armed forces. More than one-third of the total national budget is spent upon the police and the armed forces. Military service is compulsory for all male citizens. The armed forces are composed of an army, national and territorial guards, an air force, and river-patrol gunboats.

Government services. *Education.* Elementary education is free and compulsory for children between the ages of seven and 14. School facilities are inadequate; about 22 percent of the population is illiterate and only about 75 percent of school-age children attend classes. Almost all children speak Guaraní at home, and the schools provide their first contact with the Spanish language, which is the medium of instruction. Instruction includes reading and writing, mathematics, and science; in addition, agriculture, home economics, and health are emphasized.

Secondary education is provided by government-run *colegios* (high schools), and privately owned lycees. Primary schoolteachers are given six years of training. The Normal School at Asunción prepares students for further training as secondary schoolteachers at the National University. The secondary level provides training in business and the professions, and there are technical and vocational schools, as well as a bilingual secretarial school for women.

The two universities—the National University and the Catholic University—are located in Asunción. The larger National University is financed by the state and attendance is free. The International Higher Institute of Public Relations confers a Master of Public Relations degree that is recognized by the National University. There are also schools of agriculture and veterinary medicine, as well as bacteriological and radiology institutes.

Health and welfare services. The Social Welfare Institute serves all wage earners. Benefits include treatment of both occupational and nonoccupational diseases and accidents, maternity care, disability compensation, and life insurance covering accidental death. Railway and bank employees receive benefits through retirement and pension funds. There are state agencies for the planning and implementation of land reform and for technical, economic, and social assistance to small farmers.

Most medical facilities and physicians are located in Asunción. There are a few provincial hospitals, but most care is available in health centres. Voluntary organizations combat leprosy; inter-American cooperative services combat tuberculosis and malaria.

Housing. Except for the larger cities, housing conditions are inadequate. The most common home is an adobe hut roofed with reeds and floored with packed earth. In the cities, Spanish architecture predominates. Water supply and sewage systems are most inadequate in urban areas and nonexistent in the countryside. The cities are faced with a chronic housing shortage, due to the influx of people from the countryside and the expanding population. Crowded housing conditions are also related to the unequal distribution of land; large estates are depopulated while small plots are overcrowded. To encourage building, the Inter-American Development Bank grants 30-year mortgage credits.

The police. Police chiefs, or *comisarios*, are appointed by the president. The police in Asunción are responsible for the maintenance of dossiers on the country's citizens, for the prevention and investigation of crime, and for the maintenance of the prisons. People arrested while committing a crime can be detained for 24 hours before being placed under judicial jurisdiction.

Social conditions. *Wages and the costs of living.* Minimum wages were introduced by the Ministry of Labour in 1961. The average annual per capita income is equivalent to about $200 U.S., but this figure is deceptive because most subsistence farmers do not participate in the cash economy. Chronic inflation has resulted in a continually rising cost of living and in a decreasing value of the minimum wage.

Health conditions. Intestinal and lung diseases are the major causes of death. Inadequate shelter during the winter months is the chief factor in the prevalence of influenza, pneumonia, and tuberculosis. There is an incidence of malaria and leprosy, and, as mentioned, infant mortality is high.

Economic and social divisions. The main divisions of Paraguayan society follow economic lines. The chief groupings are the rich and the poor, the learned and the unlearned, the owners and the labourers.

CULTURAL LIFE AND INSTITUTIONS

Guaraní and Spanish traditions

The main characteristic of Paraguayan culture is its fusion of both the Guaraní and Spanish traditions. Folklore, the arts, and literature reflect this dual origin. Guaraní musicians adopted the Spanish harp and guitar, and Latin-Americanized versions of European dances are popular. Sculpture and wood carving were adapted for the embellishment of Catholic churches and missions. Paraguayan artists were receptive to the European Impressionist movement. The country's outstanding handicraft is the production of ñandutí lace, which is thought to represent a combination of 16th-century needlepoint lacemaking techniques from Europe with Guaraní traditional patterns and designs.

Cultural institutions. Almost all cultural institutions are located in Asunción. There are three learned academies concerned with Paraguayan and Guaraní history and culture, as well as various societies and research institutes. The fine arts are represented by the Normal School of Music, the Conservatory of Music, the National Academy of Fine Arts, and the Asunción Symphony Orchestra. There are museums concerned with ethnography, natural history, and military history, as well as collections of the work of Paraguayan artists and private collections of national memorabilia.

Library services are also centred upon the national capital. The largest are the National Library, the National Museum, and the National Archives; other government institutions include the Public Library of the Ministry of Defense, and the Library of the Ministry of Foreign Affairs. The American Library is attached to the Godoi Museum. Scientific materials are collected by the Library of the Museum of Natural History and Ethnography and the Library of the Paraguayan Scientific Society.

The press and broadcasting. Although Paraguay's political life is subject to control, censorship of the public media is not excessive, and moderate criticism of the government is allowed. All newspapers are published in Spanish and are issued on a daily or weekly basis. *ABC*, founded in 1967, and *La Tribuna*, founded in 1925, have the largest circulations in the country. *La Tarde* and *Comunidad* carry Catholic news items. *El Páis* is the official government organ; *Patria* supports the Colorado Party; and *El Radical, El Liberal*, and *El Enano* are all liberal newspapers.

Radio Nacional is the government radio station. Its programs are supplemented by those broadcast from the eight privately owned commercial radio stations. Television service is available only in Asunción.

Future prospects. The future of Paraguayan political life depends on the maintenance of the principles of free enterprise, on the liberty of all to work without being

obligated to belong to a certain political party, and on free elections. Economic advance is dependent upon greater and more equitable use of hydroelectric power, the suppression of smuggling, better use of foreign loans, and the attraction of emigrants back to Paraguay.

BIBLIOGRAPHY

General works: A. SCHUSTER, *Paraguay, land, volk, geschichte, wirtschaftsleben, und kolonisation* (1929), a comprehensive study, though now dated; PHILIP RAINE, *Paraguay* (1956), an introductory survey covering the history, economics, and social conditions of the country; J. HALCRO FERGUSON, *The River Plate Republics: Argentina, Paraguay, Uruguay* (1968); EDWARD A. HOPKINS, RAYMOND E. CRIST, and WILLIAM P. SNOW, *Paraguay, 1852 and 1968* (1968), with a guide to further literature. See also general texts on Latin America.

Economics: JOSEPH PINCUS, *The Economy of Paraguay* (1968), the best general reference on this subject in English; GEORGE PENDLE, *Paraguay: A Riverside Nation,* 3rd ed. (1967), a brief, informative study; INSTITUTO ITALO LATINO AMERICANO, *República del Paraguay* (1970), a succinct, accurate summary of demographic and economic–financial structures of Paraguay.

Social conditions and culture: PAN AMERICAN UNION, *Paraguay* (1965); ELMON R. and HELEN S. SERVICE, *Tobati: Paraguayan Town* (1954); and MARTIN DOBRITZHOFER, *An Account of the Abipones, an Equestrian People of Paraguay,* 3 vol. (1822; reprinted in 1 vol., 1970), two ethnological studies.

Constitution: *Constitución de la República del Paraguay* (1967; Eng. trans. issued by the PAN AMERICAN UNION, 1969), includes the constitution of 1967 and previous ones for 1844, 1870, and 1940.

(J.P.P.)

Paraguay, History of

The first Europeans known to have visited what is now Paraguay were Alejo García, in 1525, and Sebastian Cabot, in 1528. The area was colonized for Spain by Domingo Martínez de Irala, under whose orders Juan de Salazar founded Asunción in 1537. During the 17th and 18th centuries the country was chiefly noted for its *reducciones*, unique autonomous communities of Christian Indians run by Jesuit missionaries. After 1776 it was part of the viceroyalty of Río de la Plata, centred at Buenos Aires.

The establishment of the nation. Paraguay's struggle for independence was not so much a fight against far-off Spain as against nearby Buenos Aires, which, considering itself heir to the Spanish viceroyalty, sought to extend its rule over Paraguay when it declared its own autonomy in 1810. The junta of Buenos Aires sent an army to Paraguay under the command of their spokesman, Manuel Belgrano; the Paraguayans resisted Belgrano, defeating him at Paraguarí and Tacuarí. On May 14–15, 1811, a Paraguayan revolution led by Capt. Pedro Juan Caballero repudiated Spanish rule. One month later a governing junta was established, presided over by Fulgencio Yegros, hero of the recent campaigns; its principal member, however, was José Gaspar Rodríguez Francia. The new government proposed to Buenos Aires that the two provinces form a confederation, but the plan was not accepted. Instead, on October 12, 1811, a military alliance was arranged, based on Paraguayan independence. Difficulties emerged very soon, with Buenos Aires' unsuccessful attempt to recruit Paraguayan troops for the war against Spain. Asunción requested, also without success, arms to meet a Portuguese threat in the north. The alliance was torn asunder. The leader of the Banda Oriental (now Uruguay), José Gervasio Artigas, tried to involve Paraguay in his struggle against Buenos Aires, unfurling the banner of federation. Paraguay proclaimed its neutrality in the civil wars of the Plata, but Buenos Aires applied economic pressures to reduce Paraguay to submission. A delegation was sent to Asunción to stress the necessity and the advantages of annexation. Outraged, the Congress proclaimed a republic on October 12, 1813, rejecting any accord with Buenos Aires. The regime was headed at first by two consuls, in imitation of ancient Rome. José Gaspar Rodríguez Francia, one of the consuls, was an admirer of classical antiquity. The other consul was Fulgencio Yegros, whose prestige quickly waned.

The Francia dictatorship

The consulate was soon replaced (1814) by a dictatorship under Francia, a strange, misanthropic figure, an antimilitarist in a warrior nation and a follower of Voltaire and Rousseau. In the absence of a constitution and with no defined limits on his powers, Francia (officially designated "El Supremo Dictador") affirmed his will as the only law until he died. When Congress ceased to meet, he concentrated all power in himself, prohibiting any political activity and ruthlessly suppressing his opposition. The discovery of a conspiracy in 1820 brought on a reign of terror. Yegros and other leaders of the revolution were executed or committed suicide. Paraguay became cloistered in an impenetrable isolation, cut off from contact with the outside world. Foreign commerce was forbidden; no one could leave or enter the country. The French botanist Aimé Bonpland spent ten years in captivity for crossing the border. Bolívar, at the zenith of his prestige, proposed that Argentina invade Paraguay in order to free Bonpland, "a mission which would command the world's attention." Francia declared that "he wanted neither peace nor war with anyone," defending his regime as the only way to save Paraguay from the anarchy in which South America was floundering. He did in fact impose the most absolute law and order, but the people lived, however tranquilly, in a vacuum, ignorant of everything that happened elsewhere. Paraguay was "the China of America" until the dictator died a natural death on September 20, 1840, leaving no provisions for a successor.

There were tendencies to anarchy as one government rapidly succeeded another until Carlos Antonio López, a philosophy professor, came to the fore. He had remained prudently away in the countryside all during the dictatorship. López briefly re-established the consulate, sharing power with Mariano Roque Alonso, an obscure military man. In 1842 Paraguay belatedly issued a formal declaration of its independence, but the dictator of Argentina, Juan Manuel de Rosas, refused to recognize this act. López, chosen as president of the republic in accord with a constitution adopted in 1844, obtained the support of Brazil against Buenos Aires; in 1845 and 1849 he declared war on Argentina, without occasioning major repercussions. Finally, on July 17, 1852, after the fall of Rosas, the new Argentine chief of state, Gen. Justo José Urquiza, recognized Paraguayan independence. The great powers followed suit.

Paraguay's conflicts with its neighbours. Once Paraguay became active in international affairs, the country began a period of great progress, aided by hundreds of European and North American technicians. The state, nevertheless, maintained strict economic controls and continued to be the principal entrepreneur. Blast furnaces, railroads, a river merchant fleet, shipyards, arsenals, and a powerful army made Paraguay one of the foremost nations of South America, more and more in conflict with its neighbours. In 1855 boundary and navigational disputes prompted Brazil to send an unsuccessful ful naval expedition against Paraguay. The complaints of an American-backed industrial company, the only one allowed by López, and the bombardment of the gunboat USS "Water Witch" in 1855 caused the U.S. to dispatch a punitive squadron. Through the good offices of President Urquiza, the dispute was resolved by arbitration favourable to Paraguay. There were other grave conflicts with France and England, caused by López' firm opposition to European influence. Because of an unresolved boundary dispute, relations with Argentina were also poor. When Carlos Antonio López died, on September 10, 1862, he gave this last advice to his son and successor, Gen. Francisco Solano López:

> There are many disputes which you must settle; do not try to resolve them with the sword but rather with the pen, especially in the case of Brazil.

Francisco Solano López soon found himself involved in a disagreement with Argentina and Brazil, which viewed the growing power of Paraguay with uneasiness. A Uru-

guayan faction, citing a threat against their country by Argentina and Brazil, asked Paraguay's aid. López, convinced that the balance of power in the region was in danger and believing that Paraguay owed its independence to that balance, demanded clarifications, which Argentina and Brazil denied him. On August 30, 1864, López lodged an unheeded protest over a Brazilian invasion of Uruguay. On November 12, he ordered the capture of the Brazilian ship "Marqués de Olinda," and hostilities were begun. In a lightning campaign Paraguay seized the principal strongholds of Brazil's Matto Grosso region. López asked permission to cross Argentine territory for the purpose of attacking Brazilian forces in Uruguay. When the request was denied, the Paraguayan Congress declared war on Argentina on March 18, 1865. In less than a month the Paraguayans had occupied the Argentine province of Corrientes and were advancing toward Uruguay. On May 1, 1865, Uruguay (where anti-Paraguayan forces had triumphed) signed a secret alliance with Argentina and Brazil by which the three powers agreed to cooperate against López. They arranged in advance for the partition of Paraguay when he had been deposed.

The Paraguayan War

Defeated at Uruguaiana, Brazil, Paraguayan forces retreated to their own territory, to the region around Humaitá, near the confluence of the Paraná and Paraguay rivers. For two years bloody battles were fought in that sector; finally the Paraguayans had to abandon Humaitá, and López dug in around Villeta. In December 21–27, 1868, a gigantic battle took place at Lomas Valentinas, ending in the virtual extermination of the Paraguayan army. López retreated to the mountains, where he formed a new army. The final campaign, in Amambay, was the most dramatic of all. Paraguayans died not only from their wounds or from hunger but even at the hands of López himself, whom the defeat had driven desperate. He even ordered that his own mother be beaten. Only 400 starving soldiers arrived with López at Cerro Corá, his last encampment. There, on March 1, 1870, shouting "I die with my native land," he was killed in combat. Paraguay had been razed to the ground, with 1,000,000 dead. Most of the survivors were women.

Under a liberal constitution, ratified November 25, 1870, Paraguay undertook a difficult reconstruction. Although the domination of the victorious nations was complete, they nearly went to war with each other over the interpretation of the severe provisions of their secret pact. In 1872 Brazil imposed its version of the boundaries of the conquered nation, helping Paraguay contain the territorial ambitions of Argentina. Buenos Aires finally agreed to submit its border dispute with Paraguay to arbitration. Pres. Rutherford Hayes of the United States ruled in favour of Paraguay in 1877. When the former allies returned home, Paraguay was left to its own fortunes and entered a period of great political instability: in 80 years only six presidents served a full term. The two political parties that even today polarize the views of Paraguayans were founded almost simultaneously in 1887—first the opposition Partido Liberal (Liberal Party) and then the Partido Colorado, which had assumed power. In 1904, when a revolution removed the Colorados from office, the Liberales took control of the government. Efforts to reconstruct the nation became hampered by a dispute with Bolivia over possession of the Gran Chaco region in western Paraguay. Colorado administration made repeated concessions (1879, 1887, and 1894), in the hope of attracting Bolivian commerce. To contain Bolivian incursions into the Chaco, the Eligio Ayala government (1924–28) armed the nation and entered the disputed territory. A clash occurred on June 15, 1932, when Ft. Carlos Antonio López, or Pitiantuta, was captured by the Bolivians. For the next three years, Paraguay and Bolivia fought bitterly in the inhospitable Chaco. While Eusebio Ayala held civil power in Paraguay, the military commander, Gen. José Félix Estigarribia, demonstrated an intelligent strategy and by a string of victories pushed the Bolivians back almost to the edge of the disputed territory. Through the mediation of Argentina, Brazil, Chile, Peru, Uruguay, and the United States, a cease-fire was concluded June 12, 1935. A peace conference convened at Buenos Aires, signing the final treaty on July 21, 1938. The Paraguay–Bolivia boundary was submitted to the arbitration of the presidents of the six mediating nations, who handed down their decision the following October 10. Paraguay retained three-fourths of the Gran Chaco, including its frontier along the Paraguay River.

The Chaco War

Recent history. On February 17, 1936, a military coup jailed Ayala and Estigarribia, removed the Liberales from power, and abrogated the constitution of 1870. The new government was short-lived; in 1939 Estigarribia was elected president with the support of the Liberales. On September 7, 1940, a few days after dictating a new constitution along authoritarian lines, he was killed in an airplane crash. His successor was Gen. Higinio Morínigo, who relentlessly persecuted the Partido Liberal. He sided with the Allies of World War II. With the support of the Colorados he crushed a revolt in 1947, but the following year he was deposed by those same Colorados, who used that means to return to power. Six presidents succeeded each other for short terms until 1954, when Gen. Alfredo Stroessner took over the government, supported by the army and the Colorado Party.

Close cooperation by the United States allowed Stroessner to undertake important public works. The currency was stabilized; the geographic isolation of the country was broken down by modern highways stretching to the Atlantic and the Río de la Plata, and by a network of airlines. Stroessner's dictatorship was modified in 1967, when a new constitution reinstated Congress and granted some civil liberties. Still, opposition continued to be severely suppressed. The church clashed repeatedly with the government. In the early 1970s there was no guerrilla action or terrorism in Paraguay, but the standard of living was one of the lowest in Latin America, and, in spite of heavy foreign aid, progress in the struggle against underdevelopment was slight.

BIBLIOGRAPHY. General works include CECILIO BAEZ, *Resumen de la historia del Paraguay desde la época de la conquista hasta el año 1880* (1910); EFRAIM CARDOZO, *Paraguay independente* (1949) and *Breve historia del Paraguay* (1965); and HARRIS GAYLORD WARREN, *Paraguay: An Informal History* (1949), with an excellent annotated bibliography.

Well-documented studies of the independence movement are BLAS GARAY, *La revolución de la independencia del Paraguay* (1897); FULGENCIO R. MORENO, *Estudio sobre la independencia del Paraguay* (1911), an economic interpretation. These have been superseded by JULIO CESAR CHAVES, *Historia de las relaciones entre Buenos-Ayres y el Paraguay, 1810–1813* (1938). On the Francia dictatorship, JULIO CESAR CHAVES, *El supremo dictador, biografía de José Gaspar de Francia*, 2nd ed. (1946), is a masterly work. On the López period, JUAN FRANCISCO PEREZ ACOSTA, *Carlos Antonio López, obrero máximo, labor administrativa y constructiva* (1948), is well documented; and JUSTO PASTOR BENITEZ, *Carlos Antonio López: estructuración del Estado Paraguayo* (1949), provides a sociological interpretation. On the War of the Triple Alliance, see PELHAM HORTON BOX, *The Origins of the Paraguayan War*, 2 vol. (1929, reprinted 1967); and with new documentation, HAMON J. CARCANO, *Guerra del Paraguay: acción y reacción de la Triple Alianza*, 2 vol. (1941); EFRAIM CARDOZO, *Vísperas de la guerra del Paraguay* (1954), *El imperio del Brasil y el Río de la Plata* (1961), and Cardozo's journalistic diary *Hace cien años* (1964–70), a day-by-day account of the war. For the period between the two wars, see ANTONIO ZINNY, *Historia de los gobernantes del Paraguay, 1535–1887* (1887); and GOMES FREIRE ESTEVES, *Historia contemporanea del Paraquay* (1921).

On the Chaco War, the memoirs of MARSHAL ESTIGARRIBIA, *The Epic of the Chaco* (1950); and DAVID H. ZOOK, JR., *The Conduct of the Chaco War* (1960), are basic.

(E.C.)

Paraguay River

The fifth largest river in South America, the Paraguay River (Portuguese Rio Paraguai) is the principal tributary of the Paraná River and an important artery of the Río de la Plata system, which serves the interior of South America; it is 1,584 miles (2,550 kilometres) long. The source and upper course of the Paraguay River are in Brazil, where it demarcates part of the frontier with Para-

guay before entering Paraguay itself, which it traverses from north to south; it then forms the frontier between Paraguay and Argentina for the last 150 miles of its course before entering into the Paraná River, which flows into Argentina.

The "river of cockades"

The name Paraguay is taken from the Guaraní language and could be translated "river of *paraguas* (coloured, plumed birds)" or "river of cockades," an allusion to the plumed headdresses once worn by the riverine peoples.

Inaccessible to oceangoing ships, the river is used only for local traffic; steamers from Buenos Aires, capital of Argentina, ply upstream as far as Asunción, capital of Paraguay. Often flowing through country rife with malaria and other tropical diseases and subject for much of its length to seasonal flooding, the Paraguay River constitutes more of an obstacle than an aid to the development of its basin. Surveys of its upper course, however, are being conducted by the Brazilian government in conjunction with the United Nations to assess the area's economic potential. (For details on associated physical features, see GRAN CHACO; PARANA RIVER; and PLATA, RIO DE LA.)

The natural environment. *Physiography.* The Paraguay rises in Brazil, in the central plateaus of Mato Grosso, at an altitude of 980 feet above sea level. Where it becomes navigable, about 150 miles downstream, near Cáceres in Brazil, after its confluence with the Sepotuba, it is 275 feet wide and 20 feet deep. Another 20 miles downstream, where the Jauru joins it at an altitude of 400 feet above sea level, the Paraguay enters the alluvial Mato Grosso *pantanal* (floodplain); it crosses the plain's western edge over a sandy bed, flowing around the many islands in its course. During its passage through the *pantanal*, it receives such important tributaries as the Cuiabá, the Taquari, and the Miranda.

About 470 miles downstream, it flows north to south to form the boundary between Brazil and Paraguay before being joined by a tributary, the Río Apa, which flows in from the east and demarcates part of the Brazilian–Paraguayan frontier. The river then enters Paraguay, having travelled about 950 miles from its source. After flowing for more than 200 miles across Paraguay, it is joined by the Río Pilcomayo at the Argentinian border, near Asunción. It then flows south-southwest along the Argentinian–Paraguayan frontier for about 140 miles, until it is joined on its west bank by the Bermejo. Continuing along the border for another 40 miles, it then empties into the Paraná River at a short distance from the Argentinian city of Corrientes.

Hydrography. Throughout its hydrographic basin, which covers more than 380,000 square miles, altitudes rarely exceed 650 feet above sea level. Rainfall varies from 40 to 80 inches annually. Over a long distance, the gradient of the river varies from about three-quarters of an inch to an inch per mile. The various streams of the basin have low banks or natural levees, built up when silt is deposited along the slower flowing portions of the river channel during flood stage. When the river recedes, its banks thus remain elevated above the level of the neighbouring plains. During floods a continuous water table, often as much as 15 miles wide, underlies the inundated plains, and about 38,600 square miles of surface area are flooded.

The Paraguay River has varying rates of flow between its source and its mouth. Above Corumbá, in Brazil, it has a typically tropical regime—at its highest in February and at its lowest from July to August. Below Corumbá, the high point occurs in July and the low point from December to January.

Variations in the rate of flow

At Corumbá, the mean discharge rate during the driest period exceeds 30,000 cusecs (cubic feet per second); the mean annual rate, 68,000 cusecs; and the mean discharge of floodwaters, 100,000 cusecs. At Puerto Sastre, at the confluence of the Apa, the average discharge of floodwaters reaches more than 225,000 cusecs, with the maximum occurring from June through August and the minimum in January. The variation in rate of flow is not attributable to the regimes of the river's tributaries but is explained primarily by the overflowing of water onto the *pantanal* in summer and by the gentle gradient of the riverbed. The upper Paraguay (down to Concepción, Paraguay) consequently floods from December to March, while the middle Paraguay (from Concepción to Asunción) floods between May and June as a result of the delay in flow of waters from the upper stream. Along the lower Paraguay (from Asunción to the Paraná River), tributaries with tropical regimes carry their greatest volume of water from December to March, causing floods in February; whereas the contributions of the Andean tributaries, the Bermejo and the Pilcomayo, produce floods between February and June. The flood area exceeds 30,000 square miles.

From its confluence with the Apa for the 630 miles to its mouth, the Paraguay, while still navigable, runs on a shallow, broad bed, with an average width of about 2,000 feet. Its right (Argentinian) bank gradually lowers, whereas its left (Paraguayan) bank becomes elevated, forming cliffs. Along this stretch, floods develop principally on the western bank, spreading over the Argentinian plain for distances of from three to six miles. These lands form part of the Gran Chaco (*q.v.*). Downstream from the complex of low hills south of Assunção known as Lomas Valentinas, the floods spill out over both banks, inundating areas between six and nine miles wide. At this point, the riverbed has an average width of about 2,300 feet.

Climate. The predominant climate of the Paraguay Basin is of the hot and humid savanna type, characterized by dry winters from April to September and heavy rains in summer from October to March. More than 80 percent of the annual precipitation occurs in the summer months, with little or no rainfall in June and July. Annual mean temperatures are above 64° F (18° C), the absolute maximum temperature being from 104° to 107° F (40° to 42° C) and the absolute minimum temperature being about 34° F (1° C). October is frequently the warmest month, except in the south, where the warmest month is January and where the contrast between rainy summers and dry winters is more pronounced. The extreme southern portion of the basin has a humid, moderately warm climate in which the heaviest rains occur during the summer and the winters are not completely dry.

The *pantanal* complex

Vegetation and animal life. Some of the *pantanal*'s vegetation, called the "*pantanal* complex," is typical of the Mato Grosso plateau regions, while some other is typical of lowlands. Plants that thrive in water and in moist soils as well as those that flourish at moderate temperatures or are adapted to dry regions are found within the complex. As on the banks of the Niger in West Africa, both forest and grassland types occur. The water plants, found on the permanently flooded lands, are typified by the water hyacinth and by the "Victoria regia" water lily. Moisture-loving species, such as the trumpetwood and the guama, flourish over most of the floodplain.

On the savanna, after the floods, various grasses such as paspalum and knotroot bristle grass reappear. Vegetation of a more evolved type, which thrives at moderate temperatures, occupies the unflooded highland. It is represented by nut-bearing palms and by various types of laurels. In the forests of the region, the carandá, or copernicia palm (a tropical palm that yields wax), the paratudo, the muriti palm (a large fan palm), and various types of quebracho trees (South American hardwoods that are a source of tannin) predominate. In the Gran Chaco region along the west bank of the river, and in other sections where drought is more pronounced, plants adapted to dry conditions occur. In the lowlands of eastern Paraguay, forest cover and savanna grasslands alternate.

Among the fish of the Paraguay River, some of which supplement the diet of the riverine population, are the dorado (a fish resembling salmon), the piranha (a fish resembling the bluegill that travels in large schools and devours any meat that falls in the water), and the pacu (which resembles bass).

The development of the river. *Early history.* Before the arrival of the Spaniards in the 16th century, the region was inhabited primarily by the Paiaguá (a part

The original inhabitants

of the Guaycure tribe) and by the Guaraní. A subgroup of the Guaraní, the Xaraiés, lent their name to the original designation of the *pantanal*—the Lagoon of the Xaraiés. The Paiaguá and the Guató of the *pantanal* engaged in foraging, hunting, and fishing. Others, such as the Tereno and the Guiana (Arawak), practiced a rudimentary type of agriculture, building hillocks above flood level. The southern Guaraní cultivated maize and cassava (manioc). The inhabitants of the Gran Chaco—the Paiaguá, Mataco, Mascoy, and Zamuco—were nomadic fishers and hunters.

In what is now Paraguay, the Spaniards and Portuguese interbred freely with the indigenous Amerindians. Consequently, the present riverine population of the country is largely mestizo, or mixed, and Guaraní as well as Spanish is the common language. In Brazil, however, miscegenation was less general, so that some groups of indigenous peoples remain, forming isolated nuclei. Others, like the Bororo, the Tereno, the Cadiueu, and the Bacairi, constitute minorities who have adopted Christianity and Brazilian culture and who live on the fringe of the region.

Among the first Europeans to enter the region were the Spanish Jesuits who mapped the river in the 17th century. In the 19th century, the first hydrographic survey of the river was made, followed in the 20th century by a more detailed hydrographic survey.

Navigation. The Paraguay is navigable from its mouth as far upstream as Cáceres, a total distance of about 1,400 miles. Several of its Brazilian tributaries are navigable, such as the Jaurú, the Sepotuba, the São Lourenço, and the Taquari. Ships with a draft of more than six feet, however, cannot go upstream beyond Corumbá. The major ports of the river are Corumbá, Cuiabá, and Pôrto Esperança in Brazil and Asunción in Paraguay.

At Pôrto Esperança, the Northwest Brazil Railway, from São Paulo on the Atlantic coast, crosses the river over the President Dutra Bridge (6,500 feet long) and runs north to Corumbá, where it links with the Brazil–Bolivia Railway at nearby Puerto Suárez, Bolivia.

Economic resources. The Brazilian section of the Paraguay Basin is sparsely peopled, with a population that becomes slightly more numerous toward the Río Apa in the south. Livestock raising for local consumption is the principal economic activity. Subsistence agriculture is carried on by traditional nonmechanized methods; attempts to establish settlers in the region have failed. A vegetable-extract industry is based on maté (Paraguayan tea), while condensed tannin, used in leather making, is extracted from the bark of the quebracho tree. Manganese deposits also are exploited. Transport and communications are poorly developed.

Prospects. Largely because of flooding, the Paraguayan section of the basin is similarly economically underdeveloped—even in the vicinity of Asunción and Concepción, where there are larger population concentrations. Since 1967 the Brazilian government and the United Nations Special Fund have been conducting jointly a hydrological study of the upper Paraguay with a view to developing the region. Another United Nations Survey, in which the Argentinian and Paraguayan governments are participating, is investigating navigational problems and the potential for river transport between Asunción, the Paraguayan capital, and Corrientes in Argentina.

(W.F.O.)

Paraíba

Paraíba, a primarily agricultural state of northeastern Brazil, is bounded by the states of Rio Grande do Norte on the north, Ceará on the west, and Pernambuco on the south and by the Atlantic Ocean on the east. Its area is 21,765 square miles (56,372 square kilometres), and its population early in the 1970s was over 2,400,000. Its chief river, the Paraíba, rises on the Pernambuco border and flows toward the sea past the state capital, João Pessoa, 11 miles from the port of Cabedelo. The only other large city is Campina Grande, a cotton-shipping centre farther inland.

Paraíba (or, as it is known in its older variants, Parahyba and Parahyba do Norte) is a name of Tupí Indian origin formed from the words *para* and *hiba,* meaning "arm of the river." The name of the capital was itself formerly Paraíba, but it was changed to honour the memory of a former governor, João Pessoa, a reformist and vice presidential candidate whose assassination in 1930 helped spark the revolution that brought Getúlio Vargas to national power in Brazil.

Historical background

Northeastern Brazil was the first part of the country to become wealthy when, in the 16th century, the Portuguese established the world's first large-scale sugarcane plantations there with African slave labour. Founded on August 5, 1585, as the captaincy of Itamaracá (a captaincy being a kind of fiefdom granted by the Portuguese crown), Paraíba shared in the sugarcane riches of the period; and, because sugar required large investments, cheap labour, and machinery, economic and political power fell into the hands of a few wealthy families. In the 18th century, cotton, Paraíba's chief product today, began to be produced and became a significant export.

Physical features and economy

Physiographically, Paraíba in the east has a narrow coastland of sandy beaches and dunes, off which deep-sea fishermen, or raftsmen (*jangadeiros*), ride the surf on tree-trunk rafts. There is no coastal plain; from this seaboard the land rises abruptly to coastal mesas (*taboleiros*), which, together with a few inland river valleys, provide the principal wealth of the state—from cotton, sugar, and sisal, together with tobacco, corn (maize), cacao, oiticica oil, and hides. These coastal areas, enjoying dependable rainfall, were once covered by dense tropical forests but from the early period of plantations were cleared away for crops and pastures. To the west, behind the zone of coastal mesas, a hilly upland known as the Bordorema Plateau occupies most of the central part of the state. Copper, tin, and a variety of rare minerals are mined on the plateau, though their economic importance is less than that of agriculture. The plateau is a semi-arid region once covered by deciduous, thorny scrub woodland called *caatinga.* In the *caatinga* there are only small islands of abundant rains and forests on the tops of the higher mountains. Generally, though, the area is dry not so much because of a lack of measurable rain as because of the unevenness of the rainfall and the poor drainage. Rain falls in summer and autumn and evaporates quickly or, partly because of man's excessive clearing of the land, runs off the impermeable ground, leaving sandy gullies and pebble-strewn stretches of dry earth. Life in the *caatinga* country thus depends on irrigation; even though the federal government has built a number of reservoirs, however, the nature of the hilly land permits irrigated crops to be raised only around the margins of the lakes behind the dams. Finally, Paraíba's farthest western section consists of broad plains—peneplains, or semi-arid flatlands, developed by erosion, with only a few tablelands called *chapadas* remaining from an earlier era.

In the 1960s Paraíba began to industrialize; two industrial parks in João Pessoa and Campina Grande housed some 30 industries, including clothing, agricultural machinery, cellulose, plastics, soaps, synthetic fibres, stoves, and rubber shoes. Two highways crosscut the state to connect it with the rest of Brazil; in the early 1970s power lines brought electrical energy from Pernambuco and distributed it among all the towns of Paraíba.

A university, the Universidade Federal da Paraíba, located in João Pessoa, with a faculty also in Campina Grande, was founded in 1955. (C.de P.L.)

Parallax, Astronomical

Parallax is the difference in direction of an object as seen by an observer from two different positions. In astronomy, the measurement of parallax is used directly to find the distance of a celestial body from the Earth (geocentric parallax) and from the Sun (heliocentric parallax). The two positions of the observer and the position of the object form a triangle; if the base line between the two observing points is known and the direction of the object as seen from each has been measured, the apex angle (the parallax) and the distance of the object from the observer

can be found simply. In the determination of celestial distance by parallax measurement, the base line is taken as long as possible in order to obtain the greatest precision. For the Sun and Moon, the base line is the distance between two widely separated points on the Earth; for all bodies outside the solar system, it is the axis of the Earth's orbit. The largest stellar parallax is 0.76″, for Alpha Centauri; the smallest that can be directly measured is about 25 times smaller, but indirect methods discussed below permit calculation of the parallax, inversely proportional to the distance, for more and more distant objects but also with more and more uncertainty.

LUNAR AND SOLAR PARALLAX

Definitions of lunar and solar parallaxes

The parallax of the Sun or Moon is defined as the difference in direction as seen from the observer and from the Earth's centre. In Figure 1, let O be the observer on the surface of the Earth, E the centre of the Earth, and M the position of the Moon; then the angle OME is the parallax. This varies with the altitude of the Moon. If the Moon is directly overhead, the parallax is zero, and parallax is greatest when the body is on the horizon. At an angular distance z from the zenith, Z, we find from the triangle OME that $\sin p = a/r \sin z$. When $z = 90°$, $\sin p = a/r$ and this value is called the horizontal parallax or, briefly, the parallax. For all bodies except the Moon, p is so small that it does not differ appreciably from $\sin p$, and it is usually expressed in angular measure. The definitions of lunar and solar parallax must be further refined because of the spheroidal figure of the Earth. The numerical values generally given are those of the equatorial horizontal parallax. The solar parallax is usually derived from measurements of the positions of other bodies of the solar system.

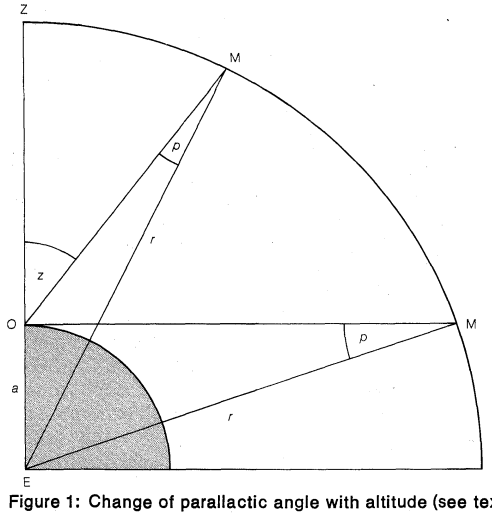


Figure 1: Change of parallactic angle with altitude (see text).

Lunar parallax. The first parallax determination was for the Moon, by far the nearest celestial body. Hipparchus (150 BC) determined the Moon's parallax to be 58′ for a distance of approximately 59 times the Earth's equatorial radius as compared with the modern value of 57′02.6″; that is, a mean value of 60.2 times. Lunar parallax is directly determined from observations (see Figure 2) made at two places, such as G, Greenwich, England, and C, the Cape of Good Hope, that are nearly on the same meridian. Angles z_1 and z_2 are observed, and other data are obtained from the latitudes of the observatories and the known size and shape of the Earth. In practice, stars near the Moon are observed also to eliminate uncertainties of refraction and instrumental errors. In this way Thomas Henderson obtained a value of 57′01.8″ for the Moon's equatorial horizontal parallax in 1837. From a series of observations of a small lunar crater (1905–10), the value 57′02.5″ was found.

Lunar landings in 1969 have brought at least the possibility of direct measurement of the Earth's equatorial angular diameter as seen from the Moon. This is slightly more than 2°, about twice the lunar horizontal equatorial parallax.

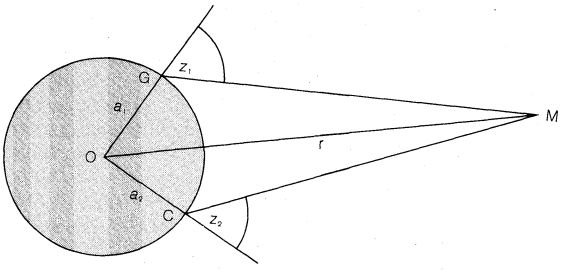


Figure 2: Measurement of parallax by observations from a northern and a southern observatory.

Another method rests on a comparison of the force of gravity at the Earth's surface with its value at the Moon. If M and m are the masses of the Earth and Moon, r the mean distance, P the sidereal period of revolution of the Moon about the Earth and k the constant of gravitation $k(M + m) = 4\pi^2 r^3/p^2$ where $\pi = 3.14$. Also, g, the value of gravity at the Earth's surface, determined accurately from pendulum observations, is equal to kM/a^2. Hence

$$\left(\frac{a}{r}\right)^3 = 4\pi^2 \frac{M}{M+m} \cdot \frac{a}{gP^2} \qquad (1)$$

As the quantities on the right-hand side are known with great accuracy, a/r is accurately determined as 57′2.7″.

Observations of lunar occultations, eclipses of stars by the Moon, also can be used to determine the lunar parallax. Observing the occultation of the same star from two or more stations may be compared to the Henderson method described above, because the same point on the Moon is observed from two or more geographic locations. In 1950 the lunar parallax was thus obtained from observing four occultations at nine stations in the United States, tying the lunar distance to the U.S. geodetic network. A value of 57′02.602″ for the parallax was obtained.

Radar measures of the distance from the Earth to the Moon have provided a recent value of the lunar parallax. Radar ranges have the advantage of being a direct distance measure, although the ranges are affected by variations in the surface topography of the Moon and require assumptions about the lunar radius and the centre of mass. The first systematic measurements were made at the U.S. Naval Research Laboratory by B.S. Yaplee, who obtained a value of 57′02.540″.

Adopted value of the lunar parallax

The International Astronomical Union, on the basis of existing observations, in 1964 adopted a value of 57′02.608″ for the lunar parallax corresponding to a mean distance of 384,400 kilometres.

Laser beams directed at retro-reflectors placed on the Moon by Apollo 11 astronauts in 1969 have led to lunar ranging with 15 centimetres precision and promise accuracies approaching 2 centimetres in the point to point measurement of the distance.

Solar parallax. *Trigonometrical methods.* In accordance with the law of gravitation, the relative distances of the planets from the Sun are known; and the distance of the Sun from the Earth can be taken as the unit of length. The measurement of the distance or parallax of any planet will determine the value of this unit. The smaller the distance of the planet from the Earth, the larger will be the parallactic displacements to be measured with a corresponding increase in accuracy of the determined parallax. The most favourable conditions are therefore provided by the observation, near the time of opposition, of planets approaching close to the Earth. The determination can be based either on simultaneous or nearly simultaneous observations from two different places on the Earth's surface, or on observations made after sunset and before sunrise at the same place, when the displacement of the place of observation produced by the rotation of the Earth provides the base line for the measurements.

The first reasonably accurate determination of the Sun's parallax was made in 1672 from observations of Mars at Cayenne, French Guiana, and Paris, from which a value of 9.5″ was obtained.

In 1877 Sir David Gill made an expedition to Ascension Island and observed the opposition of Mars, using a heliometer to measure the distance of the planet from neighbouring stars, making observations after sunset and before sunrise. A value of 8.78″ was deduced.

In 1898 the minor planet Eros was discovered; it has an orbit so elliptical that at its nearest approach Eros comes within about 24,000,000 km of the Earth. The opposition of 1900–01 (when the least distance was less than 50,000,000 km) was extensively observed and gave a value for the solar parallax of 8.804″. A more favourable opposition occurred in 1930–31 when the least distance was only 26,000,000 km. From photographic observations in both hemispheres, Sir Harold Spencer Jones derived a solar parallax of 8.790″ ± 0.001″. This parallax corresponds to a distance of 149,670,000 km.

Methods depending on velocity of light. The value of the velocity of light has been determined with very high precision and may be utilized in several different ways. A direct method is the converse of the procedure of Ole Rømer in the discovery of the velocity of light; *i.e.*, to use the light equation, or time taken by the light to reach us at the varying distances of Jupiter, but great accuracy is hardly obtainable in this way. A second method is by means of the constant of aberration, which gives the ratio of the velocity of the Earth in its orbit to the velocity of light. As aberration produces an annual term of amplitude 20.496″ in the positions of all stars, its amount has been determined in numerous ways. Observations made at Greenwich in the years 1911 to 1936 gave the value 20.489″ ± 0.003″ leading to the value 8.797″ ± 0.013″ for solar parallax. This method is not free from the suspicion of systematic error.

The velocities of stars toward or away from the Earth are determined from spectroscopic observations. By choosing times when the orbital motion of the Earth is carrying it toward or from a star, the velocity of the Earth in its orbit may be obtained. In this way the solar parallax was found from observations at the Cape of Good Hope to be 8.802″ ± 0.004″.

In a similar way, and with higher precision than by the method just mentioned, the solar parallax has been determined from radio wavelength velocity measurements of the 21-cm neutral hydrogen features of the four brightest radio sources in space. In this manner S.H. Knowles, at the U.S. Naval Research Laboratory, obtained in 1967 a solar parallax of 8.7940″ ± 0.0006″.

Radar measures of the distance from the Earth to Venus have provided the best determination of the solar parallax. By measuring the flight time of a radar pulse to Venus, the distance between the two planets can be obtained, allowing the determination of the unit distance between the Earth and the Sun.

Value of the astronomical unit

The present value for the radar astronomical unit is 149,598,000 km ± 200 km, corresponding to a solar parallax of 8.79414″ ± 0.00004″. The principal limitations of the method are its dependence on knowledge of the planetary orbits, the uncertainty in the value of the velocity of light, and the possibility of electromagnetic effects in the Earth-Venus plasma delaying the radar pulse.

Gravitational methods. In lunar theory there is a term of period one month known as the parallactic inequality. The coefficient of the term contains the ratio of the parallaxes of the Sun and Moon as a factor. The large size of this coefficient makes it of value. From the discussion of occultations of stars from 1672 to 1908, Spencer Jones found the value 125.023″ ± 0.033″ for this term giving for the solar parallax 8.796″ ± 0.004″.

The ratio of the combined mass of the Earth and the Moon to that of the Sun may be determined from the disturbing action of the Earth and Moon on the elliptic motion of the planets. The ratio of the Moon's mass to that of the Earth is 1/81.30, and thus the ratio of the Earth's mass to that of the Sun is found. In a manner similar to that described above for the Moon's parallax, the solar parallax is then derived.

From an exhaustive discussion of the perturbations of the minor planet Eros from 1926 to 1945, which included the favourable opposition in 1930–31, Eugene Rabe in 1945 made a determination of the solar parallax as 8.7984″ ± 0.0004″. In 1967, G. Zeck, using the same data as Rabe but including revised values of the coefficients of the reciprocal mass of the Earth–Moon system, obtained a value of 8.7944″ ± 0.0003″ in excellent agreement with the radar results.

At the General Assembly of the International Astronomical Union in 1964 the value 8.79405″ (8.794″) for the solar parallax was adopted, corresponding to an astronomical unit of 149,600,000 km.

STELLAR PARALLAX

The stars are too distant for any difference of position to be perceptible from two places on the Earth's surface; but as the Earth revolves at 149,600,000 km from the sun, stars are seen from widely different viewpoints during the year. The effect on their positions is called annual parallax, defined as the difference in position of a star as seen from the Earth and Sun. Its amount and direction vary with the time of year, and its maximum is a/r, where a is the radius of the Earth's orbit and r the distance of the star (Figure 3). The quantity is very small and never

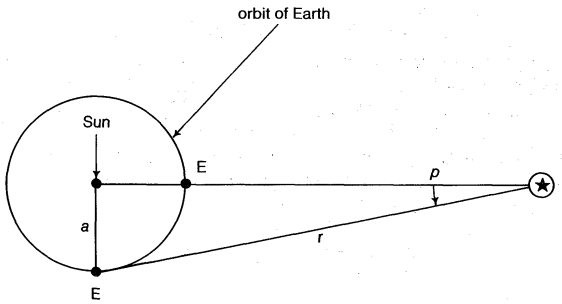


Figure 3: Stellar parallax.

reaches 1/206,265 in radians, or 1″ in sexagesimal measure. Many unsuccessful attempts to measure the parallax of a star were made after the acceptance of the Copernican system, including one by James Bradley, which led to the discovery of aberration (see LIGHT, ABERRATION OF), and one by Sir William Herschel, which led to the discovery of binary stars. The first successful results were announced by F.G.W. Struve in 1837 for Alpha Lyrae, by F.W. Bessel in 1838 for 61 Cygni, and by T. Henderson in 1839 for Alpha Centauri.

First parallax determinations

Direct measurement. Early measurements of stellar parallaxes were made visually with instruments of small size having apertures between 10 to 15 centimetres and focal lengths of the order of 1.5 to 3 metres. The observations were laborious and imprecise, and by the 20th century, parallaxes had been determined for only 72 stars.

The introduction of the photographic method by F. Schlesinger in 1903, using the large Yerkes Observatory refractor, considerably improved the accuracy of stellar parallaxes. In practice a few photographs are taken when the star is on the meridian shortly after sunset at one period (epoch) of the year and shortly before sunrise six months later. Since the star's positions also change because of its motion across the sky (proper motion), a minimum of three such sets of observations is necessary for obtaining the parallax. From approximately 25 photographs taken over five epochs, the parallax of a star usually is determined with an accuracy of about ± 0.010″ (probable error), even though the diameter of the photographic disk of the star is rarely less than 2.0″. The concerted effort of observatories in the Northern and Southern hemispheres resulted in the determination of parallaxes of about 6,000 stars by 1950. These include most of the stars brighter than magnitude 6.0 (roughly the limit of visibility with the naked eye) and many fainter stars whose comparative nearness is suggested by their relatively large annual motion in the plane of the sky (proper motion).

The results have been compiled in a Yale University publication, *General Catalogue of Trigonometric Stellar Parallaxes,* published in 1952.

The unit in which stellar distances are expressed by as-

tronomers, the parsec, is the distance of a star whose parallax is 1″. This is equal to 206,265 times the Earth's distance from the sun, or approximately 30,000,000,000,000 km. When p is measured in seconds of arc and the distance d in parsecs, the simple relation $d = 1/p$ holds. One parsec is equal to 3.26 light years.

The star with the largest known parallax, 0.76″, is Alpha Centauri. Fifty-eight separate stars are known within a distance of five parsecs from the Sun. These stars include the bright stars Alpha Centauri, Sirius, Procyon, and Altair, but the majority are faint telescopic objects.

Indirect measurement. For stars beyond a distance of 30 parsecs (parallactic angle .03″) the trigonometric method is in general not sufficiently accurate, and other methods must be used to determine their distances.

The parallax can be derived from the apparent magnitude of the star if there are any means of knowing the absolute magnitude of the star; *i.e.*, the magnitude the star would have at the standard distance of ten parsecs. For many stars a reasonable estimate can be made from their spectral types (see STAR) or their proper motions. The formula connecting the absolute magnitude, M, and the apparent magnitude, m, with parallax, p, is

$$M = m + 5 + 5 \log p \qquad (2)$$

expressing the condition that the light received from a star varies inversely as the square of the distance.

Some groups of stars, such as the Hyades cluster in Taurus and the Ursa Major cluster, have proper motions converging toward a definite point on the celestial sphere and are called moving clusters. The apparent convergence is due to the effect of perspective on parallel motions. Once the direction toward the convergent point is known, and the proper and radial motion of a member star is known, the parallax can be determined from the geometry.

Mean parallaxes. The solar system is moving through space with a velocity of 19.5 kilometres per second, carrying it four times the Earth's distance from the Sun in one year. This produces a general drift in the angular movement of the stars away from the apex or point in the sky to which the movement is directed. Were the stars at rest, this would give a ready means of determining their individual distances. As the stars are all moving, the method gives the average distance of a group of stars examined, on the assumption that their peculiar motions are eliminated. In this way the average, or mean, parallaxes of stars of successive apparent magnitudes, of different galactic latitudes, and of different spectral types are obtained. Thus the mean parallax of fifth magnitude stars (*i.e.*, of stars just visible to the naked eye) is 0.018″, and of the tenth magnitude stars (*i.e.*, of stars each giving about 1/100 of the light of a star of the fifth magnitude) is 0.0027″.

As spectroscopic observations have given the mean, peculiar velocities of stars of different types, the average proper motions perpendicular to the direction of the solar motion may also be used as a criterion of parallax. There is great diversity in the parallaxes of individual stars included in these mean values, but the means are useful for statistical studies.

Spectroscopic parallaxes. The spectra of nearly all stars can be grouped into a small number of classes, which form a continuous sequence depending on the effective (surface) temperatures of the stars. The Henry Draper (HD) classification, which is of this kind, uses the letters O–B–A–F–G–K–M to denote classes with temperatures descending from about 30,000° K for class O to about 2,500° K for class M. The HD system has been generally adopted, usually in combination with a decimal subdivision for refined work.

Empirical studies show that the spectra of the stars also include important clues to their true luminosities. A. Maury noticed that stars of the same spectral class often had marked differences in line sharpness; E. Hertzsprung found that the sharp-lined stars were intrinsically brighter than the broader-lined objects. In 1914 W.S. Adams and A. Kohlschütter established the spectroscopic differences between giant and dwarf stars of the same spectral type and laid the foundation for the determination of spectroscopic parallaxes. These differences, depending upon the intrinsic brightness of the star, allow an estimate of its absolute magnitude, and the parallax can then be deduced by means of the equation (2) given above. This method has been applied to most of the brighter stars in the Northern Hemisphere, using stars of known parallax as standards.

MK classification system

A two-dimensional classification system of stellar spectra, which has been universally adopted, has greatly improved the accuracy of spectroscopic parallaxes. The system, developed by W.W. Morgan, P.C. Keenan and E. Kellman and called the MK-system, assigns a precise system of Draper classes and five luminosity classes, using the Roman numerals I to V. The system divides the majority of stars into supergiants, bright giants, subgiants, and main sequence (dwarf) stars, depending upon their intrinsic brightness, as determined from the spectral lines most sensitive to this property. The luminosity classes are then calibrated in terms of absolute magnitude.

Photometric parallaxes. The colours of the stars can also be used as indicators of their absolute magnitude, as first shown by E. Hertzsprung in 1905 and 1907. A measure of the colour of a star is the difference in brightness, measured in magnitudes, in two selected wavelength bands of its spectrum. Initially the difference between the visual and the photographic magnitude of a star was defined as the colour of its light, and called its colour index. A comparison between the colour index and the spectral classification of a star has made it possible to develop a quantitative method of measuring a star's absolute magnitude. Several photometric systems have been developed. The most widely used system is the two-dimensional quantitative classification method developed by J.M. Johnson and W.W. Morgan, based upon photoelectric measurements in three wavelength bands in the ultraviolet, blue, and yellow (or visual) regions of the spectrum, hence called the *UBV* system. The system of the two colour indices *U-B* and *B-V* is calibrated in terms of spectral class and luminosity class on the MK-system, based upon a set of standard stars.

The relationship between the two indices in the *UBV* system and the absolute magnitudes for the main-sequence stars is of particular interest.

By means of this relationship, and the inverse square law, it is possible to determine the distances to galactic clusters from photoelectric observations of main-sequence stars in these clusters.

Dynamical parallaxes. If the relative orbit of a visual binary system is known, the following relation connects the combined mass, M, of the two stars expressed in the Sun's mass as unit; the orbital period, P, expressed in years, the semi-major axis of the relative orbit; a, expressed in seconds of arc; and the parallax p: $p = a/\sqrt[3]{MP^2}$. Both a and P are known, but not M; it will be noted that an error in the value of M gives rise to a much smaller error in p. Thus, for instance, increasing M by a factor of 8 only halves the value of p. The value of p obtained by assuming the combined mass to be equal to the mass of the Sun is called the hypothetical parallax.

In many visual pairs the complete orbit has not been observed. If s denotes the apparent distance in seconds of arc and ω the relative motion in seconds of arc per year, a hypothetical parallax can be derived from the formula $p = 0.418\sqrt[3]{s\,\omega^2}$.

By use of the relationship between mass and luminosity of a star, it is possible, knowing the spectral type of the star, to derive a correcting factor that will give a more accurate value of the parallax. Parallaxes so determined are called dynamical parallaxes. Details are given in *The Masses of the Stars* by H.N. Russell and Charlotte E. Moore (1940), which lists the dynamical parallaxes of 166 binaries with determined orbits and those of 2,363 pairs showing relative motion.

PARALLAXES OF DISTANT OBJECTS

Globular clusters. Globular clusters are star systems containing 100,000 to 1,000,000 stars and distinguishable by their spherical appearance. Their estimated number of 125 forms a spherical system centred on the Galaxy. The

globular clusters contain the so-called RR Lyrae-type variable stars that are easily recognizable and are known to have mean absolute magnitudes of 0.5. Also, it has been found from the nearer systems that the 25 brightest stars in a globular cluster have a mean absolute magnitude of − 0.8. These absolute magnitudes, compared with the observed apparent magnitudes, lead to the determination of the cluster distance or the parallax from the formula $M = m + 5 + 5 \log p$ above. Two other less reliable criteria are the assumptions that all globular clusters have the same absolute diameters and the same total integrated absolute magnitude; in both cases the distance scales are determined from the nearer globular clusters. Except for a few very distant ones, the globular clusters all lie within a distance of about 70,000 light years of the galactic centre.

Pulsars. The most recently discovered type of astronomical object is the pulsar, a rapidly pulsing cosmic radio source with a period of the order of a second. The pulsar with the shortest period so far discovered, 0.033 second, has been identified optically with a 17th-magnitude star near the centre of the Crab Nebula.

The pulsars discovered so far all appear to be too distant for direct parallax measurement. The Crab Nebula pulsar is assumed to be at the distance of the nebula itself, about 6,500 light years. Distances to some pulsars have been estimated from observations of the galactic neutral-hydrogen 21-centimetre line in absorption.

Galaxies. Galaxies are star systems similar to our own Galaxy and contain many millions of stars.

The determination of their distances depends upon a stepwise procedure, with errors increasing with the distances. For the nearest resolved galaxies (the Local Group), the distances can be determined from observations of the apparent magnitudes of the Cepheid variables, whose absolute magnitudes are related to their period of variation, as determined from distance calibration in our Galaxy. With the 200-inch Hale telescope it is possible to resolve such distance indicators in galaxies as the bright red and blue supergiant stars, novae, and supernovae up to a distance of 60,000,000 light years. Beyond that limit, the red shifts in the spectrum lines of the galaxies (see UNIVERSE, STRUCTURE AND PROPERTIES OF; UNIVERSE, ORIGIN AND EVOLUTION OF) appear to be related to their distances according to the expansion law of Edwin P. Hubble:

$$c\,\Delta\lambda/\lambda = H\,D \qquad (3)$$

where c is the velocity of light, $\Delta\lambda/\lambda$ the observed fractional wavelength shift, D the present distance, and H the Hubble constant, presently estimated to be of the order of 30 kilometres per second per million light years, but perhaps as low as 15 in the same units. The largest red shift observed for a galaxy is of the order of half that of the velocity of light leading to a distance of 5,000,000,000 light years.

Another important method to determine the distances to distant galaxies is based upon observations of the brightest member galaxy in a cluster of galaxies. These are giant elliptical galaxies with a constant order of brightness and can therefore be used as distance indicators.

Quasi-stellar sources. These objects, also called quasars or QSO's and discovered in 1963 as strong cosmic radio sources, were later identified as starlike objects with large red shifts.

The observed, proportional red shifts, $\Delta\lambda/\lambda$ (see equation 3), range from 0.06 to 2.35. The latter figure indicates a recession of 84 percent of the velocity of light if the exact formula for a Doppler shift, from relativity theory (see RELATIVITY), is applied. Red shifts of this magnitude, if interpreted cosmologically, lead to distances so vast and luminosities for these correspondingly so enormous that other possible causes for the red shift have been investigated, including gravitational or unknown physical effects.

The most recent studies, however, of similarities between Seyfert galaxies and QSO's have made it increasingly probable that the QSO red shifts are cosmological. The QSO's are possibly the youngest objects of an evolutionary sequence of galaxies (see QUASI-STELLAR SOURCES).

BIBLIOGRAPHY. An account of determining parallaxes has been treated historically by A. PANNEKOEK, *A History of Astronomy* (1961); updated results on the lunar parallax have been discussed by I. FISCHER in "The Distance of the Moon," *Bull. Geod.*, 71:37–63 (1964). For an evaluation of the accuracy of stellar parallaxes see K.A. STRAND (ed.), in *Basic Astronomical Data*, vol. 3 of *Stars and Stellar Systems*, ed. by G.P. KUIPER and B.M. MIDDLEHURST (1963). For information on the modern techniques in determining stellar parallaxes see P. VAN DE KAMP, *Principles of Astrometry* (1967); and K.A. STRAND, "The 61-inch Astrometric Reflector, Basic Design and Accuracy," in *Vistas in Astronomy*, vol. 8, ch. 2, pp. 9–15 (1966). Statistical parallaxes have also been described by P. van de Kamp, and by D. MIHALAS and P.M. ROUTLY, *Galactic Astronomy* (1967). Spectroscopic and photometric parallaxes have been covered in several chapters of *Basic Astronomical Data*. The distances of quasi-stellar sources are discussed by M. SCHMIDT in "Quasistellar Objects," *A. Rev. Astro. Astrophys.*, 7:527–552 (1969).

(K.A.S.)

Paraná

Paraná is one of the 22 states of the Federative Republic of Brazil. It is situated in the southern part of the country and is bounded to the east by the Atlantic Ocean, to the south by the state (*estado*) of Santa Catarina, to the southwest by Argentina, on the west by Paraguay, on the northwest by the state of Mato Grosso, and on the north and northeast by São Paulo state. Its area is 77,048 square miles (199,554 square kilometres), and in the 1970 census its population was reported to be 6,998,000. The capital of Paraná, named for the Paraná River (Rio Paraná), which forms its west and northwestern border, is Curitiba. Paraná plays an important role in the Brazilian economy because it is the country's principal coffee-producing state. (For associated physical features, see the articles IGUACU FALLS and PARANA RIVER.)

History. After a century of gradual penetration by bands of Spanish explorers from São Paulo and by Jesuit missionaries, the territory of the present state was occupied, to a large extent, by the forces of a Portuguese emissary, Gabriel de Lara, in the 1640s. Gold was discovered at several locations in the 17th century, and attracted settlers. Eventually recognized as belonging to Portugal's sphere of influence, not to Spain's, the territory was attached at first to the captaincy of São Paulo and subsequently to the province of the same name. Paraná became a separate province of the Brazilian Empire in 1853 and later a state of the Brazilian Republic in 1891.

The natural environment. *Physiography.* Paraná is divided into five zones, each running approximately northeast to southwest. Proceeding westward there is the coastal region, backed by the high mountain ranges of the Serra do Mar to the west; a series of three successive plateaus, each lower than the one before; and the Paraná River borderland. Less than 3 percent of the state's area is under 600 feet above sea level; it forms part of the Serra dos Orgãos. Pico Paraná, 6,305 feet (1,922 metres) high, is the highest point in the state.

The coastal region is fringed with dunes and mangrove swamps. Numerous streams, such as the Nhundiaquara and the Cubatão, flow down from the mountainous hinterland to the ocean. The Serro do Mar ranges consist of Precambrian rocks (from 570,000,000 to 4,000,000,000 years old), which in the east overlook the sea's deeper inroads into the coastline and in the west reach their greatest heights in the peaks of Capivarí Grande (5,499 feet, or 1,676 metres, high), Serra da Graciosa (6,193 feet, or 1,888 metres), Marumbí (4,999 feet, or 1,524 metres), and Prata (4,835 feet or 1,474 metres). The Serro do Mar forms a watershed between the coastal region and the first of the three plateaus.

The plateaus

The first plateau, which lies at a height of between 2,700 and 3,000 feet above sea level, is formed mainly of crystalline rock. To the north of Río Ribeira de Iguape collects its headwaters and forms the state border for a short distance before flowing west into the state of São Paulo. Curitiba is located in the south, at about 3,000 feet above sea level.

On the western side of the first plateau, a cuesta (an

escarpment with a steep slope on one side and a gentle slope on the other), rising to heights of from 3,500 to 3,800 feet, marks the beginning of the second plateau, which gradually falls to an average altitude of 2,200 to 2,500 feet.

A basaltic scarp with a maximum altitude of 3,800 feet rises at the western border of the second plateau, forming the eastern edge of the third plateau, which slopes westward and downward till it reaches the fringes of the Paraná River at altitudes of between 640 and 960 feet. This plateau is characterized by fertile red-coloured earth. Rivers subdivide it into four highland massifs (mountain masses)—the Araiporanga massif in the northeast, the Serra da Apucarana massif extending from the east to the northwest, the Campo Mourão massif in the west, and the Guarapuava massif in the south.

Three major rivers cross the second and third plateaus. Two of them are immediate tributaries of the Paraná. The first of these, the Río Iguaçu, which rises on the first plateau, is the great river of the southern quarter of the state and is about 820 miles long. The second is the Rio Ivaí, which rises in the Serra da Boa Esperança (a scarp of the third plateau) and flows northwestward and westward. The third great river is the Río Tibagi, which rises in the southeast of the second plateau and flows northward for 340 miles to join the Río Paranapanema tributary of the Paraná. The Paranapanema, itself, flows westward along the Paraná–São Paulo boundary for a considerable distance.

The Rio Paraná borderland

The Paraná River borderland, with its low altitudes and steep gullies, is dominated by rain forest; some tracts of wilderness occur to the south of Guaíra. On the Paraná, just below the confluence of the Río Piquiri (a river of the third plateau), a break in the level of the adjacent highlands is marked by the Guaíra Falls, or Salto das Sete Quedas, 375 feet or 114 metres high. The falls have a hydroelectric potential of 13,500,000 horsepower at the average rate of flow. The cataracts of the Río Iguaçu, which occur just before its junction with the Paraná, include the Santa Mariá Falls, which is 264 feet (or 80 metres) high and has a potential of 1,485,000 horsepower at an average rate of flow.

Climate. Paraná state, the northern region of which crosses the Tropic of Capricorn, has a climate influenced by the equatorial continental air mass in the summer and by the tropical Atlantic and Polar Antarctic air masses in the winter. Moderate heat prevails. Winters are dry in the northwest, while other parts receive adequate precipitation throughout the year; summers are hot in the lower altitudes and cooler—below 72° F (22° C)—at the higher levels. On the coast the annual mean temperature is 70° F (21° C) at Paranaguá. The rainfall reaches 81 inches annually, with the largest amounts falling in January and February. In the northwest, Londrina has a mean annual temperature of 71° F (22° C), and Umuarama one of 69° F (21° C). Heat and heavy rains occur there in the summer. On the eastern edge of the second plateau, however, mean annual temperatures are lower than 69° F (21° C), with an annual rainfall of 55 inches. More temperate conditions, with greater differentiation between the seasons, are found at Curitiba, where the average temperature is 61° F (16° C), as well as in the south, where at Guarapuava the temperature averages 64° F (18° C), with a rainfall of about 67 inches a year. The climate of the south has encouraged European immigration.

Vegetation. Dense tropical rain forest extends along certain tracts of the Atlantic coast and over the uplands into the Paraná River borderland. Notable species are *Ilex paraguariensis*, or Paraguay tea, the leaves of which are used for maté, a popular South American drink; the imbuia tree, the wood of which is highly esteemed for furniture making; and the Paraná pine, which covers much of the high ground. Some areas consist of treeless savanna—those around Curitiba, Campos Gerais, and Guarapuava, for example. There are also localized bushlands, which are found around Campo Mourão along the Río das Cinzas tributary of the Río Paranapanema in the northeast, and elsewhere.

Population. The population of the state, large areas of which were practically uninhabited as late as 1850, is largely of Portuguese-Brazilian descent. After the Portuguese, waves of other immigrants began to arrive as labourers and businessmen; these included Poles, Ukrainians, Italians, Germans, and Arabs. Some towns, such as Mallet, Prudentópolis, and Vera Guarani, represent comparatively early settlements; Carambeí (Castro), which was developed by Dutch dairy farmers, and Assaí, where Japanese labourers established themselves, were settled later.

The common language is Portuguese, and the principal religion is Roman Catholicism.

The principal towns

Of the state's 288 municipalities, about 20 have more than 50,000 inhabitants. Those with more than 100,000 include the capital, Curitiba (608,000 in 1970); Londrina and Maringá in the coffee-growing north; Ponta Grossa, the important railroad junction; Guarapuava; and Umuarama.

Administration and social conditions. The governor of the state is elected for a five-year term. The Legislative Assembly, elected for four years, seats 37 deputies. The state is represented by three senators and 25 deputies in the federal government. Public health is the responsibility of the state, which maintains a health centre in each town. The state has a separate welfare department, which deals with social assistance and is also responsible for slum clearance.

The state is responsible for primary and secondary education. Among the institutions of higher learning are the Universidade Federal do Paraná (Federal University of Paraná) and the Catholic university, both located in Curitiba.

Economy. *Agriculture and mining.* Paraná is one of the richer states of Brazil. Intensively developed plantations made Paraná Brazil's chief producer of coffee (1,645,000 tons in 1969), the most important centres for this crop being Umuarama, Rondon, and Londrina. Woodcutting for various industries produced 543,400,000 cubic feet of timber in 1967. Other important crops are maté (34,000 tons in 1967); cotton (471,000 tons), with Assaí as the major centre of production; and peanuts, with Umuarama as the major centre. There is also a notable output of ramie (a strong lustrous fibre capable of being spun or woven).

Agricultural products

Maize (almost 3,000,000 tons in 1969) is grown for the most part around Ivaiporã, Rondon, and Toledo; rice, manioc, potatoes, beans, oats, rye, barley, and wheat are cultivated quite widely, as are garlic, onions, and tomatoes; and there is a continually increasing output of soybeans. Sugarcane (2,220,000 tons in 1969) is grown mainly around Porecatu. Fruit production includes oranges, bananas, grapes, and pineapples.

Paraná's livestock herds have been much enlarged and improved with help from official agencies: in 1969 cattle numbered more than 4,000,000, pigs more than 8,000,000, sheep and horses less than 1,000,000. Dutch and East Indian (zebu) breeds of cattle have become acclimatized, notably around Guarapuava and Rondon. Yields in 1968 amounted to 129,400,000 gallons of milk, 2,200,000 tons of butter, and 550 tons of wool, mainly from the north and from the east.

In addition to a regular output of dolomite (a type of limestone or marble), lead, iron, talc, and lime, Paraná in 1969 produced 457,000 tons of coal from Venceslau Brás and 341,000 tons of cement from Rio Branco do Sul.

Transport. The main railroad in Brazil from São Paulo to the south traverses the eastern half of Paraná state; an important branch line from Ponta Grossa serves Curitiba and the Atlantic ports; other branch lines serve some of the major centres of the north and south. Highways link the state with São Paulo to the north and Santa Catarina to the south, and also link Curitiba with the hinterland. Since the rivers are generally navigable only for limited distances, the communications of the hinterland depend mainly on motor highways. Paranagua and Antonina are the chief seaports, and Curitiba and Londrina the chief airports.

Cultural life. Several towns have public libraries, that at Curitiba being the most important. The principal towns also have theatres and cinemas. The theatre at Guaíra, the best known in the state, is widely known in South America; its architecture is in the modern style.

BIBLIOGRAPHY. THOMAS PLANTAGENET BIGG-WITHER, *Pioneering in South Brazil: Three Years of Forest and Prairie Life in the Province of Paraná*, 2 vol. (1878, reprinted 1968); GRAFICA EDITORA PARANA CULTURAL, *História do Paraná*, 4 vol. (1969); REINHARD MAACK, *Geografia física do estado do Paraná* (1968).

(J.C.de F.)

Paraná River

The Paraná River, or Río Paraná (Paraná means "Father of Waters" in the Guaraní language) together with its tributaries forms the larger of the two river systems that drain into the Río de la Plata Estuary on the east coast of South America. Its basin of 1,081,000 square miles (2,800,000 square kilometres) includes the greater part of southeastern Brazil, Paraguay, eastern Bolivia, and northern Argentina. The third greatest river of the New World (after the Amazon and the Mississippi), and an important commercial waterway, the Paraná is 2,485 miles (3,998 kilometres) long. It extends from the confluence of the Grande and Paranaíba rivers in southern Brazil, and runs generally southwestward for most of its course, before turning southeastward to drain into the Río de la Plata. (For associated physical features, see IGUACU FALLS; PARAGUAY RIVER; and PLATA, RIO DE LA.)

The natural environment. *Physiography.* The Paraná begins in Brazil, at the confluence of the Grande and Paranaíba rivers, in latitude 20° S. The Grande rises in the Serra da Mantiqueira, part of the mountainous hinterland of Rio de Janeiro, and flows westward for approximately 800 miles; but its numerous waterfalls (such as the Cachoeira do Marimbondo [Marimbondo Falls], with a height of 72 feet) makes it of little use for navigation. The Paranaíba, which also has numerous waterfalls, is formed by many affluents, the northernmost headstream being the São Bartolomeu, which rises in the Serra dos Pirineus near Brasília.

The Alto Paraná and its tributaries

From its origin in the Grande–Paranaíba confluence to its junction, 1,740 miles downstream, with the Paraguay, the river is known as the Alto (Upper) Paraná. This upper course of the river receives many tributaries, some from the right, some from the left. The three most important of these tributaries, namely the Tietê, the Paranapanema, and the Iguaçu, all belong to the left bank, having their sources near the Atlantic coast of Brazil (the first two in São Paulo state, the last-named in Paraná state).

The Alto Paraná flows at first in a southwesterly direction down a deep cleavage in the ancient Brazilian massif (mountainous mass), the configuration of which determine its course. Just before it begins to run along the frontier between Brazil to the east and Paraguay to the west, the river has to cut through the Serra de Maracaju (or Cordillera de Mbaracayú), which has the effect of a dam; as the river approaches this natural barrier, it expands its bed into a lake two and a half miles wide and four and a half miles long. Pôrto Guaíra in Brazil stands on the southern shore of this lake. The river's passage through the mountains is marked by the Salto das Sete Quedas (Guaíra Falls), which have eight times the water volume of the Niagara River of North America. Below these falls the river flows through a narrow canyon, 196 feet wide and 328 feet deep, for a distance of two miles; farther south, however, its bed becomes wider again and the bordering heights lower, so that at Pôrto Mendes, on the left (Brazilian) bank 37 miles downstream from the falls, the canyon is 492 feet wide and only 295 feet deep.

The Río Iguaçu ("Great Water" in the Guaraní language) joins the Alto Paraná at the point where Brazil, Paraguay, and Argentina converge, 82 miles downstream from Pôrto Mendes. Rising in the Serra do Mar near the Brazilian city of Curitiba (for which reason it is sometimes called the Rio Grande de Curitiba), the Iguaçu has a course of 820 miles from east to west, during which some 70 waterfalls reduce its bed's altitude above sea level by a total of about 2,650 feet. While the Ñacunday Falls are 131 feet high, the spectacular Iguaçu Falls, on the frontier between Brazil and Argentina, 14 miles upstream from the Iguaçu–Alto Paraná confluence, have a height of 269 feet. One of the currents into which the river is divided descends from this height at a single drop, but most of the others make two successive drops of 98 feet each. The cataracts have a mean water volume of 59,400 cubic feet per second. The crescent-shaped edge over which the water pours is more than one and a half miles wide; thereafter, the river passes for several miles through the Garganta del Diablo, or Devil's Gorge—a canyon only 164 feet wide between heights varying from 65 to 328 feet. At its junction with the Alto Paraná, however, the Iguaçu is 820 feet wide and 40 feet deep.

From the Iguaçu confluence to its junction with the Paraguay River, the Alto Paraná continues as the frontier between Paraguay and Argentina. So long as it is flanked on the left (Argentinian) bank by the steep edge of the Sierra de Misiones, the river proceeds in a generally southwesterly direction, but twists repeatedly to and fro over a rocky bed studded with outcrops of melaphyre (a dark-coloured rock of porphyritic texture). At Posadas in Argentina, however, where it is about one and a half miles wide, it turns abruptly westward and begins a more amply meandering course, embracing islands of quite considerable size and punctuated so frequently by rapids and by outcrops of basalt that navigation is difficult. At the Apipé Rapids the river is only about four to six feet deep.

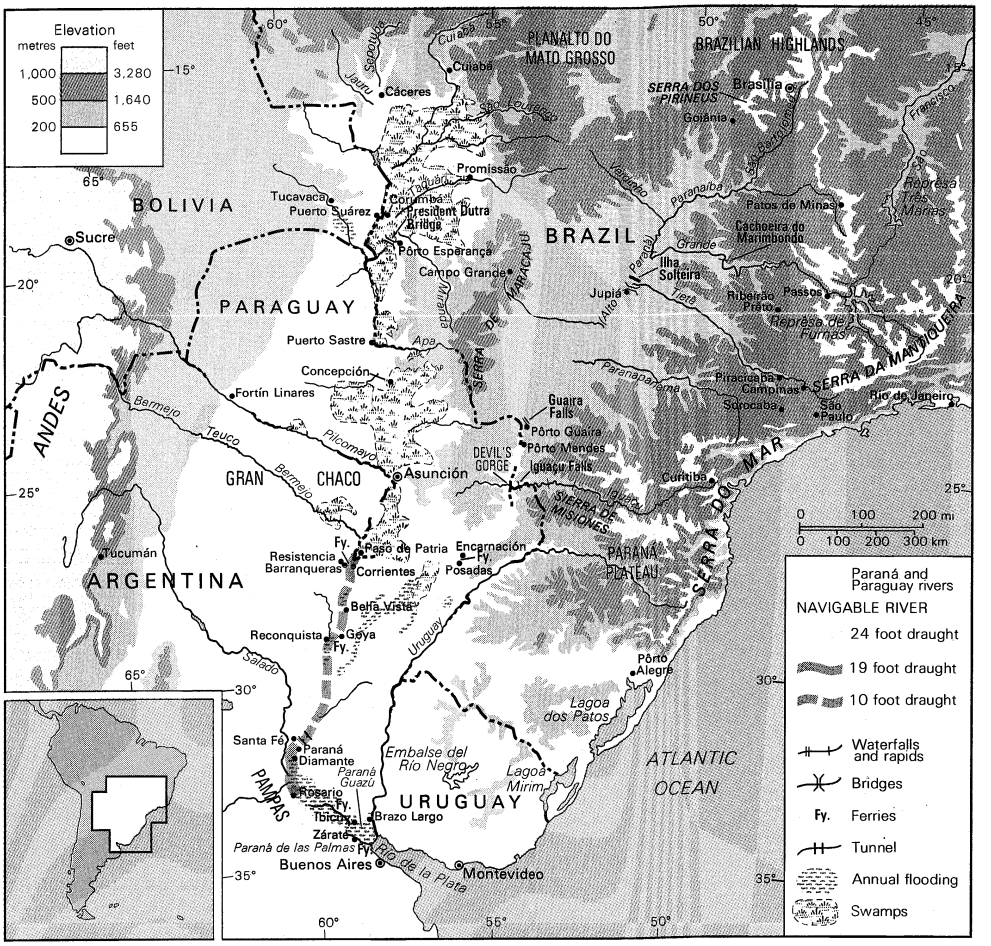
At Paso de Patria, on the right (Paraguayan) bank, at a place appropriately named Confluencia, the Paraná receives its greatest tributary, the Paraguay River, that joins it on the right bank. Thenceforth it flows on through only Argentinian territory.

The lower Paraná

The combined stream of the Paraná turns southward as it passes Corrientes. It now becomes a typical "plains-type" river, banked by its own alluvial deposits and having an extensive floodplain on its righthand side for the next stretch of its course, with tracts up to 24 miles wide liable to inundation. Its permanent bed, about two and a half miles wide at Corrientes, narrows to about 8,000 feet at Bella Vista, to about 7,000 feet at Santa Fé, and to about 6,000 feet at Rosario and is strewn throughout with chains of islands. Santa Fé, on the right bank opposite the port of Paraná, stands where the Paraná River receives its last considerable tributary, the Salado. Between Santa Fé and Rosario, however, the right bank begins to rise as the river skirts the edge of the undulating pampa (grassy plain), which flanks it all the way down to the delta, and reaches altitudes ranging from about 30 to 65 feet. The left bank, meanwhile, is always higher than the right but has to sustain the erosive action of the water, which becomes more and more turbid as great masses of earth are constantly falling into it; in the delta the main branch of the river runs along a break in the terrain, with its left bank consisting of a cliff about 75 feet high.

The delta

The delta of the Paraná has its apex as far north as Diamante, upstream from Rosario, where branches of the river begin to turn southeastward. About 11 miles wide at its upper end, the delta is 40 miles wide at its lower, where the separated branches flow into the Río de la Plata, about 200 miles from Diamante. With an area of 5,500 square miles in 1970, the delta—from calculations based on the maps made in 1818, 1912, and 1945—appears to be advancing at the rate of 230 feet per annum, the yearly volume of alluvial deposit being estimated at 165,000,000 tons. Within the delta the river divides again and again into distributary branches, the most important being the two last great channels, the Paraná Guazú and the Paraná de las Palmas. The islands of the delta, alluvial in origin, are low-lying and of varying size. Their shores and the outer fringes of the river have protective embankments covered with trees, but may nevertheless be submerged in times of flooding, when they present the appearance of flooded forests.



The Paraná and Paraguay river basins.

Hydrology. The velocity of the current of the Paraná changes frequently during the River's long course. For the Alto Paraná, the rate becomes slower wherever the bed widens (especially when a real lake is formed, as before the Guaíra Falls) and very much faster wherever the bed narrows (as in the canyon downstream from Guaíra). Farther downstream it slackens on its way to Posadas, but accelerates thereafter over a series of rapids and races. Downstream from Corrientes it becomes slower again, stabilizing its flow at a mean rate of two and a half miles per hour on the way to the Río de la Plata.

The volume of flow

The volume of the Paraná is, for practical purposes, correlated to the amount it receives from the Paraguay, which supplies about 25 percent of the total. Where it flows into the Río de la Plata, the river's overall mean volume is 610,740 cubic feet per second. While the highest recorded volume is 2,295,000 cubic feet per second (1905) and the lowest 86,400 (1945), averages are 1,587,600 for the seasonal high-water period and 218,700 for the low-water period.

High periods occur normally in the summer (November–February); low periods in August and September. An important factor is that the Alto Paraná and the Paraguay reach their maximum flow at different times. Whereas the mountainous basin of the Alto Paraná is drained so rapidly that water begins to rise at Corrientes in November, reaching its maximum height there in February, the swamps of the upper basin of the Paraguay retain precipitation so much longer that the Paraguay's high water does not reach Corrientes till May, reaching its maximum in June. Thus, levels on the lower Paraná begin to sink in March, then rise again from May, and sink again from July to September. Whenever both the Alto Paraná and the Paraguay reach their highest levels at the same time, the lower Paraná has to carry an exceptionally heavy volume of water—as it did in 1905, when the whole delta and much of the valley experienced heavy flooding.

Climate. The basin of the Upper Paraná has a hot and humid climate all the year round. The winters are dry, and the summers (lasting from November to April) are rainy, with precipitation averaging 60 inches a year but decreasing toward the west. Rainfall takes the form of drenching downpours often accompanied by hailstorms; the rainwater is drained away rapidly into the mainstream. The middle and lower basins have a subtropical climate of temperate humidity, with less plentiful rainfall (except in the Misiones Region) but without seasonal differences, the mean annual precipitation being 39 inches in the delta as opposed to 66 in Misiones. The climate along the Lower Paraná is very humid indeed and sometimes quite stifling in summer; the moist vapours become still thicker in the delta when the river brings down the torrential waters of the tropical basin.

Vegetation. The Brazilian section of the Alto Paraná forms the boundary between two natural zones: that of the forest to the east, that of the savanna (grassy parkland) to the west. Wood and maté tea leaves from the forest are shipped on the river, and the treeless savannah, with grasses and bushes, is good for cattle raising. On Argentina's section of the river, forest extends from the Misiones region to the vicinity of Corrientes; some forest trees, outside the zone of the forest proper, still occur in areas of woodland all the way downstream to the delta.

Animal life. The Paraná has a rich and varied animal life. Among its many edible fish, the pejerrey (a marine fish, silver in colour, with two darker bands on each side), the dorado (a brilliantly coloured river fish with a golden appearance), the surubi (a fish with a long rounded body, flattened at the nose), the pati (a large scaleless fish, which frequents deep and muddy waters), and the pacu (a large river fish with a flat body, almost as high as it is long) are worth mentioning; every year sportsmen visit the stretch of the river upstream from Corrientes to angle for dorado. Reptiles include the iguana lizard, two species of cayman (a crocodilian),

Species of fish

the water boa, the rattlesnake, the cross viper, and the yararáa (the most prevalent South American representative of the viper family). Frogs and toads are plentiful, as also are freshwater crabs. There are innumerable species of insects and spiders, and the islands are plagued by mosquitos. Herons, cormorants, storks, and game birds are also seen.

The human imprint. The peoples who live up and down the Paraná, be they fishermen, farmers, herdsmen, boatmen, pilots, shipwrights, or raftmakers, all share certain customs and interests in common arising from their environment, from their racial identity, and from the cultural underdevelopment of the region. Originally, almost the whole length of the river region was populated by Guaraní Indians. The Spaniards, for whom the river was the only practicable way of penetrating the interior of the country, mingled freely with the Guaraní. The cultural, physical, and linguistic characteristics of the present population reflect this mingling.

On the Alto Paraná most of the people are poor fishermen living in huts on the fringe of the forest, and depending on the river both for their food and for their transport: apart from their fishing, they profit from services rendered to passing river traffic.

On the lower Paraná the economy is more diversified, but still primitive. Apart from fishing and providing services, the islanders raise livestock; some crops, such as maize, manioc, tobacco, and fruit, are grown. Settlement of the adjacent country was begun from the banks of the river and the economy remains dependent on the river in the absence of a system of roads.

The most distinctive features of human life and activity on the Paraná can best be observed in the delta, where water transport is the sole means of movement for people and for goods alike. Such local fruit growing as there is contributes little to ecological development, as the fruit trees are not suited to the climate and are frequently damaged by floods; it would seem more rational to proceed with the afforestation of the islands liable to flooding with pine and other softwood species.

History of the mapping of the river. The Paraná, as has been said, was the principal thoroughfare used by the Spaniards for their conquest of the interior of this part of South America; Santa Fé and Corrientes in Argentina and Asunción in Paraguay (up the Paraguay River) were among the earlier settlements founded. The mapping of the river, however, preceded the European exploration of it: thanks to information obtained from Indians, Sebastián del Cano was able to include relatively accurate markings of the Paraná, Paraguay, and Uruguay rivers in the map of the estuary of the Río de la Plata that he drew in 1523—three years before Sebastian Cabot's expedition began sailing up the Paraná in 1526. Further cartographic work by agents of the Spanish crown was considerably supplemented by that of Jesuit missionaries, who first covered the whole basin of the Paraná in a series of 98 maps. In the second half of the 18th century, commissioners demarcating the frontiers between Spanish and Portuguese possessions produced a fresh quantity of maps. Of later cartographers, Félix de Azara (a Spanish naturalist and geographer sent to study the frontier question) and Martin de Moussy (a French doctor and naturalist, commissioned to undertake geographical work for the Argentinian government) are the most important.

Navigation. The Paraná is Argentina's and Paraguay's most important waterway. If one allows for changes in the water level, it is navigable from the Río de la Plata to Pôrto Mendes and, then, after a short interruption, from Pôrto Guaíra to Jupiá, just below the Salto do Urubupungá (Urubupungá Falls) in Brazil—that is to say, for a distance of 1,677 miles all the year round.

Thanks to regular dredging, seagoing vessels can go upstream for 250 miles from the Río de la Plata in any season of the year. Vessels with a draught of 24 feet can use the port of Rosario, those with a draught of 19 feet can reach the ports of Santa Fé and of Paraná, 100 miles upstream. Draughts, however, must not exceed 10 feet between Paraná and Corrientes or 4 feet between Corrientes and Pôrto Mendes, and navigation on the latter stretch is further hindered by rapids such as those of Apipé. Corrientes has consequently become a major port for transshipment. Between Pôrto Mendes and the Salto das Sete Quedas, navigation is interrupted, but transport over the 37 miles from Pôrto Mendes to Pôrto Guaíra is provided by a Brazilian railroad. From Pôrto Guaíra as has been said, the river is navigable again up to Jupiá.

River traffic

There are also regular boat services on the Lower Paraná and the Paraguay linking Buenos Aires on the Río de la Plata with Asunción in Paraguay. The Paraná carries 90 percent of Paraguayan exports and 30 percent of Argentinian exports.

Principal crossings. As the Paraná is a wide river, for much of its course, with a network of channels round its islands, it is difficult to establish permanent crossing points. The international railroad from Buenos Aires to Asunción crosses the river twice by means of ferries—between Zárate and Ibicuy in the delta and between Posadas and Encarnación on the southern frontier of Paraguay. Work is in progress for a Zárate–to–Brazo Largo road and rail link to replace the delta ferry by bridging the Paraná de las Palmas and the Paraná Guazú, and a tunnel one-and-a-half miles long under the river between Paraná and Santa Fé was built in 1969. An old-fashioned crossing by means of balsa rafts between Goya and Reconquista, upstream from the tunnel, is not of much importance; the busier crossing by the same means between Corrientes and Barranqueras is being replaced by a long bank-to-bank bridge 5,465 feet long. In Brazil the São Paulo–Corumbá railroad crosses the Alto Paraná at Jupiá, where the river is narrow.

Economic resources and prospects. The Paraná Basin is economically underexploited.

Whereas in Brazil the Alto Paraná and its affluents account for 50 percent of the river's hydroelectric potential, only those tributary rivers that rise near the comparatively populous zones of the Atlantic littoral are at present harnessed for power.

Hydroelectrical schemes

An agreement between Brazil and Paraguay for the building of one of the world's greatest dams at the Salto das Sete Quedas, due to start in 1973, was to bring some progress to a thinly populated and barely developed region. There is also a dam at Jupiá (with a capacity of 1,200,000 kilowatts [kW]), and another under construction at Ilha Solteira is to have a capacity of 3,200,000 kW. Meanwhile, Brazil makes use of the Paraná as a waterway for the transport of agricultural and forest products and for trade with neighbouring countries.

For Paraguay, pending exploitation of the river's hydroelectric potential, the Paraná is of major importance as a channel of access to the Atlantic. Paraguay's freedom of navigation on the thoroughfare is assured by its treaties with Argentina.

For Argentina, the Paraná supplies water both to the rice fields of the lands between the Paraná and the Uruguay River and to a number of cities and industrial towns; the river also provides a route for the transport not only of the river basin's produce—cereals, fruit, wood, tobacco, cotton, vegetable oils, and meat—but also of its imports—petroleum, coal, iron, and manufactured goods. The downstream traffic, however, is noticeably greater than the upstream. Argentina might also derive hydroelectric power from the Apipé Rapids and from the cataracts of the Iguaçu.

If the water power of the Paraná is to be exploited without any one country's profiting from it to the disadvantage of any other, multilateral agreements between the interested states must be concluded. The five states with river systems draining into the Río de la Plata—Argentina, Bolivia, Brazil, Paraguay, and Uruguay—in 1967 declared their readiness "to co-operate in the study and integrated development" of practical means, both by national and by international undertakings, of bringing progress to the region of which the Paraná River constitutes the major axis.

BIBLIOGRAPHY. F.A. SOLDANO, *Régimen y aprovechamiento de la red fluvial argentina*, vol. 1, *El Río Paraná y sus tributarios* (1947), a thorough study of the Paraná River.

(D.O.)

Parapsychological Phenomena, Theories of

Parapsychological phenomena of two types have been described. They may be cognitive, as in the case of clairvoyance, telepathy, or precognition and prophecy; here one person is believed to have acquired knowledge of facts, of other people's thoughts, or of future events, without the use of the ordinary sensory channels—hence the term extrasensory perception (ESP), often used to designate these phenomena. Alternatively, parapsychological phenomena may be physical in character: the fall of dice or the dealing of cards is thought to be influenced by a person's "willing" them to fall in a certain way; or objects are moved, often in a violent fashion, by "poltergeists"; the term psychokinesis (PK effect) is often used in this connection. The general term psi has become established to denote all kinds of parapsychological phenomena. Scientific interest in the subject is of relatively recent origin, but belief in the reality of such phenomena has been widespread since the earliest recorded times. Before the rise of modern science the causation of all complex physical phenomena was very poorly understood, and hence appeals to nonmaterial agencies (ghosts, devils, angels, sorcerers and witches, warlocks, demons, mythological beings) took the place of a causal, scientific explanation. Even so, there were widespread debates about the reality of phenomena that obviously transcended the bounds of everyday happenings, such as veridical prophecies, as by the oracle of Delphi, or the revival of the dead. The field was an obvious one for charlatans to enter, and even outstanding scientists believed in the possibility of astrological prediction—Johannes Kepler, for instance, earned his living by casting horoscopes.

Even now, the very existence of parapsychological phenomena is still very much in dispute, although societies for the study of psychic phenomena, made up of eminent scientists and laymen, have been in existence for almost a century and although the subject has been studied by many academic departments in British, continental, and American universities, often by scientists of outstanding ability. The discussion has sometimes assumed emotional overtones, unsuitable to scientific discipline, and outspoken but contradictory opinions are still frequently voiced. Believers and nonbelievers in psi may base their belief or disbelief on what they consider to be the scientific evidence, on their personal experiences, or on some larger system of attitudes and values into which ESP does or does not fit. When such extreme and contradictory views are widely held, it is almost certain that the evidence is not conclusive either way and that confident conclusions are unlikely to be supported by a survey of all the known facts.

Types of ESP

Extrasensory perception. While it is possible to differentiate various types of ESP, this differentiation is difficult to maintain experimentally because in any actual investigation the types tend to blend into each other. The terms used in this connection are as follows: (1) clairvoyance, defined as the perception of objects and events by paranormal means; (2) telepathy, defined as the perception of the thoughts and mental states (*e.g.*, emotions) of another person by paranormal means; (3) precognition, defined as the perception of some future event, which may be an act, a thought, or an emotion. Evidence regarding any of these three alleged types of phenomena may be by reference either to naturally occurring phenomena or to especially arranged experiments. The earlier investigations were almost exclusively of naturally occurring phenomena; it is only in the last half century that methods of experimental investigation have been elaborated.

Naturally occurring phenomena are quite diverse; they include prophetic dreams, other types of prophecy, thought reading, and allied events, which are frequently reported by normal and relatively disinterested people who are themselves surprised by the events that they recount. An unusual example is the case shown on British television of a middle-aged woman, Mrs. Rosemary Brown, who had had, as a girl, some slight musical tuition but had since then not practiced on the piano or shown any interest in music; upon the death of her husband she suddenly claimed that famous musicians like Beethoven, Brahms, and Schubert had appeared to her and were dictating musical scores to her, which she took down. When these were shown to outstanding experts they agreed that had these scores been found in some derelict attic they would have been regarded as genuine; each was not simply reasonable pastiche but genuinely expressive of what was known of the composer's emotion and personality. Apparently even highly trained musical experts could not easily (if at all) have produced work of this calibre; how a simple working-class woman with very little musical training could have done so is baffling, particularly as she was never taught composition. On the other hand, the idea of the ghosts of these Germanic musicians queuing up to dictate their recent compositions to this woman in English is not appealing. The facts are undisputed; no obvious explanation is forthcoming. As such, the story is typical of many others.

Much work has been devoted to professional mediums, persons (usually female) who make a profession of parapsychological experiences; Mrs. Leonore Piper, an American, Mrs. G.O. Leonard, British, and many others provided material for numerous books and articles. Most mediums claim to be controlled by someone in the spirit realm who speaks and sometimes acts through them; these controls enable them to perform acts of clairvoyance, telepathy, and precognition and to put people in touch with departed relatives and friends. Many mediums have been exposed as frauds; whether this makes it likely that those not so exposed were simply lucky or whether they were genuine as opposed to the fakes is still the topic of controversy. It seems unlikely that the study of naturally occurring phenomena, whether involving normal people or mediums, can throw much light on the scientific problem itself. Odd events happen all the time and every day; unless the probability of a particular event's happening can be calculated with some accuracy, it is not possible to adduce it as meaningful evidence in favour of the occurrence of ESP. Many people have reported dreaming that a particular horse won the Derby, only to find that in fact this horse did win. This does not prove that these dreams were veridical; it is equally important to know how many people dreamed about other horses winning the Derby that did not in fact win. Suppose that there are 20 runners; if 100 people report dreaming the correct winner but 1,900 dream the incorrect winner, then clearly there is nothing to be explained. Unfortunately the people who dream about the wrong horse seldom come forward so that this control figure is not known. Without such knowledge nothing can be said about the reported accurate dreams; they may be precognitive, but the evidence is completely inconclusive. The same argument may be applied to most of the claims made for ESP under naturally occurring conditions. The verdict must be that it is not proved.

Formal testing of ESP

The experimental arrangement of conditions such that proof for or against ESP can be obtained is difficult; best known in such investigations is the work of the American Joseph Banks Rhine and his many collaborators and students. Much of this was done with the so-called Zener cards; these bear one of five symbols (cross, star, circle, wave, rectangle), and 25 of them make up a pack. The experimenter may shuffle a pack and put it on the table, hidden from the subject, who then has to try to guess the order of the cards (clairvoyance), or the experimenter may look at each card in turn and ask the subject to guess at which card he is looking (telepathy, or possibly a mixture of telepathy and clairvoyance). There are many ways of varying this procedure, and millions upon millions of guesses or calls have been made in many different countries and by many different people. The advantage of this procedure is twofold: (1) a perfectly foolproof experimental design can be formulated that excludes completely the possibility of sensory knowledge of the cards and which is repeatable by other experimenters; and (2) the probability of any particular score (number of correct matchings of cards and calls) can be calculated and evaluated according to standard statistical formulae. When this is done

there emerge a number of studies where the probability of achieving the reported success in guessing correctly by chance is so small that it may be ruled out for all practical purposes. Such studies are typically of two kinds: (1) it is possible to find isolated individuals who score so highly that coincidence is ruled out; and (2) when groups of individuals are tested no one does *much* better than chance, but there is on balance a slight excess over chance that is statistically unlikely to happen. The most likely value in calling through a pack is five; this represents complete chance. If a talented person scores consistently around 15, he only has to run through three or four packs to make chance an unlikely explanation of his performance. If 100 people are tested and each person averages 5.5 correct calls, this means little for each person, but the excess adds up to a most unlikely total. Some of the most exceptionally high-scoring individuals have been discovered by the British mathematician S.G. Soal, but these are few and far between, and most work is done on large groups of subjects. Occasionally subjects persistently "recognize" not the card presented but that next to be presented; this may be evidence of precognition. Sometimes believers in ESP score positively, disbelievers negatively; when the scores are averaged they do not depart from chance, but, when they are separated according to the prior belief or disbelief of the subject in the existence of ESP, then scores significantly higher (for believers) and lower (for disbelievers) than chance are found. Many other detailed findings are reported in the experimental literature.

Criticism of ESP tests

Critics have argued about the experimental designs and about the statistical treatment. It seems clear that early experiments suffered from poor controls experimentally and that statistical analyses were subject to valid criticisms. This is not true of more recent work, and it would be difficult to fault either experimental design or statistical treatment of studies done in the last 20 years; the former has been investigated by the American Psychological Society, the latter by the American Statistical Society, and neither society made any adverse criticisms. Hence critics now tend either to throw doubt on the applicability of statistics to these phenomena or to claim that there was cheating on the part of the subject or the experimenter or both. No coherent and acceptable argument, in fact, has been put forward to show that statistical methods are not applicable to ESP work, although the possibility that negative conclusions will not be published and that positive ones will must make one doubtful about the exact values reported. But the reported values add up to such a large probability that every inhabitant of the earth would have to guess cards continually and unsuccessfully for many centuries in order to render the reported results insignificant. The question of cheating is not capable of rational discussion; no case of an academic ESP research worker actually cheating has in fact been reported, and the probability that this will happen on a large scale seems slight. Nevertheless, the possibility cannot be completely ruled out that there is a massive conspiracy on the part of a hundred or more respected psychologists and other research workers to defraud the public into believing in the reality of ESP phenomena that in reality do not exist, a conspiracy which in many cases would also have to involve their subjects. Why these people should expose themselves to the danger of being found out when their only reward for their work is that of being vilified is not made clear. If the hypothesis of widespread fraud is rejected, then the evidence for ESP is stronger than that for many tenuously supported psychological phenomena.

Psychokinesis. Here, too, there is need to distinguish between naturally occurring phenomena and experimentally arranged sessions. One of the best known of the former is the story of how D.D. Home, surrounded by witnesses, floated out of a third-story window and in at another—over 100 years ago. The event took place in Ashley House, London, in 1868, and was attested by Lord Adare, Lord Lindsay, and Captain Charles Wynne; possibly because of the high social standing of the witnesses this became the most famous of many such "proofs" of the existence of parapsychological faculties. Certainly many scientists and public men accepted it as such; yet when the original accounts are consulted little remains of the mystery. Home went to room A, leaving his companions in room B; he was then seen outside the window of room B and finally arrived in room C, which was on the other side of room B to room A. It would have been easy for him to have made arrangements beforehand to enable him to appear outside the window of the room containing his friends, either by way of a plank or by swinging from a rope attached to the roof and hanging over the parapet. Every stage magician nowadays does more difficult tricks before larger audiences and under conditions of much brighter illumination. Reports of this kind are inherently unreliable, and the uncontrolled conditions under which they are gathered make them impossible to evaluate properly. Most psychologists realize the weakness of their unaided senses in giving veridical reports of happenings, particularly when these occur quickly, unexpectedly, under conditions of poor illumination, or when powerful emotions are involved; all these factors are at work in the majority of reported psychokinetic phenomena and make reports on them, even by trained scientists, unreliable and useless. Investigators who cannot explain every trick performed by stage magicians should consider themselves barred from investigating alleged psi phenomena.

Poltergeists and children

Poltergeists constitute a particular class of agencies supposed to produce PK phenomena. What is reported to be happening is that furniture is moved, things are thrown, windows are broken and other, sometimes massive, rearrangements of physical objects take place. Loud noises may be made, but there is seldom any physical injury done to people. It is usually found that when poltergeists plague a house a young child is living in that house; whether poltergeists are attracted by children or whether children produce the phenomena is an open question. Efforts have been made to photograph such happenings and to record them by other means, but such efforts have not led to any notable results. It is still necessary to rely on personal reports, and for reasons already stated such reports are not on the whole acceptable as scientific evidence. Some serious students of the subject report themselves convinced that no natural explanation is possible of the phenomena they have observed, but this belief is not widely shared. What is curious is the similarity of the events in places separated by thousands of miles and by hundreds of years; the question remains of course whether this similarity is between different groups of poltergeists or between different children. No other PK effects are so extensive, involving such large expenditure of physical energy, as are the alleged efforts of the poltergeists, and it often seems quite beyond the capacity of children to produce such effects; some critics have suspected the help of more adult members of the family or even efforts made by outsiders having their own reasons for frightening the inhabitants.

Intermediate between accounts of poltergeist phenomena and serious experimental work on PK come studies like those done recently on Ted Serios, a poorly educated, hard drinking, ex-bellhop from Chicago who claimed to be able to produce "thoughtographs"—*i.e.*, images on film under conditions when the physical objects whose pictures appeared on the film were not present and where all physical interference with the film by Serios had been prevented. Serios was brought to scientific notice by a psychoanalyst who later accompanied him around the country and argued forcibly that this phenomenon (fitful though it was—successful pictures were only rarely produced, even in lengthy sessions) provided the breakthrough so often demanded of parapsychologists. Scientists were asked to investigate Serios; they tended to complain that they were not allowed to arrange conditions in such a way that trickery was excluded. Two expert photographers, who also had experience of amateur conjuring, investigated Serios and showed that his performance could be imitated under conditions similar to those under which he worked, without the help of occult forces; when they tried to search Serios in order to find evidence that he was using certain physical agents, they were prevented from doing so by the psychoanalyst, whom they also found ignorant of simple

laws of photography and "naïve" in many ways. It is clear from their account that such studies of Serios as have been reported are far from being scientifically impregnable and leave ample leeway for natural ways of producing the phenomena; as is often the case with mediums, here too the excuse for not allowing certain precautions to be taken is that the medium (or Serios in this case) would not like it and would be upset and thus would not be able to function properly. Even stage magicians do not have to invoke such excuses. When conditions were indeed more rigorous, as when the two photographers investigated Serios, no thoughtographs were produced.

Experimental investigation of PK effects

PK effects have been investigated experimentally by means of dice thrown at random, either by people or by mechanical means; the subject of the experiment "wills" the dice to come up high or low or with a particular face upwards, and as the probability of that happening is known, it is possible to calculate the departure from chance, very much as in the card guessing experiments. The results of much work in this field have been similar to those for ESP, but perhaps somewhat less impressive; positive results have been reported sufficiently often to make a strong case for the existence of PK, but the case is still not strong enough to convince all those who consider the evidence in favour of ESP sufficient. There is nothing wrong with the method, and, indeed, several ingenious variants on it have been used; it is simply that far fewer experiments have been done with it and that no really high-scoring subjects have yet appeared. Reliance therefore has to be placed on large numbers of marginally above-chance scoring subjects and the detailed statistical evaluation of their answers. Many gamblers, of course, are convinced of the reality of PK, as is shown in their obvious attempts to influence the fall of the dice by "willing"; this, of course, does not constitute evidence.

Theories of psi. Two difficulties with ESP and PK effects are that they (1) contradict orthodox scientific theories, notably those relating to space and time barriers, and (2) have not given rise to meaningful theories of their own. There are, indeed, so-called theories of psi, but these are mostly ad hoc and in any case do not serve the main function of a scientific theory, which is to guide research in a rigorous fashion. At best, these theories are philosophical attempts to integrate putative phenomena with some highly speculative version of "reality"; usually these theories are idealistic and opposed to materialistic views, and frequently theorists have simply used psi phenomena as a stick with which to beat materialists. Such theories as exist may be grouped as follows: (1) Physical theories. These suggest the existence of some as yet undiscovered form of energy that has produced the phenomena in question; such a form of energy would have to differ from all others by not obeying the inverse-square law, as simple physical distance does not seem to affect psi much. (According to the inverse-square law, one quantity varies inversely as the square of another; for example, light decreases by the square of its distance from its source.) (2) Field theories. These suggestions, not in essence very different from the physical theories, deal with fields of force in the language of Maxwell; essentially they fall into the same class and are subject to the same difficulties. (3) The collective unconscious. According to this theory, all minds participate in some mysterious, common, unconscious source of knowledge; and the unconscious portions of one mind may interact with those of another. Such an explanation raises more problems than it answers. The so-called theory of a subliminal self is perhaps little more than another variant of the collective unconscious type of theory; it encounters the same difficulties. (4) Projection hypothesis. This theory endows the mind with some powers that can act independently of the physical world but which can interact with it. Such a hypothesis simply asserts the facts and does little to explain them. Many other theories, similar in kind to those mentioned, could be cited, but none of them has gained wide acceptance, and none of them has been found useful by research workers. In short, the how and why of psi, assuming even that psi does in fact exist, are unknown.

One of the reasons for this state of affairs may be that research workers have been obsessed by the need to prove the existence of psi and have neglected the need to investigate the effects of altering conditions. How much and precisely in what way is psi influenced by distance, by changes in the physical properties of the cards or dice used, by electric screening, by drugs administered to subjects, by learning, by state of mind, by rewards and punishments for correct or incorrect guessing? Extroverts seem to do better than introverts. Are there other personality traits that predispose subjects to do well or poorly? Until a certain amount of knowledge is available on these points (and some efforts are now being made to study these factors) theorizing is perhaps premature. Simple repetition of experiments is unlikely to convert unbelievers and has thus a limited usefulness; what is needed are parametric experiments of the kind suggested above.

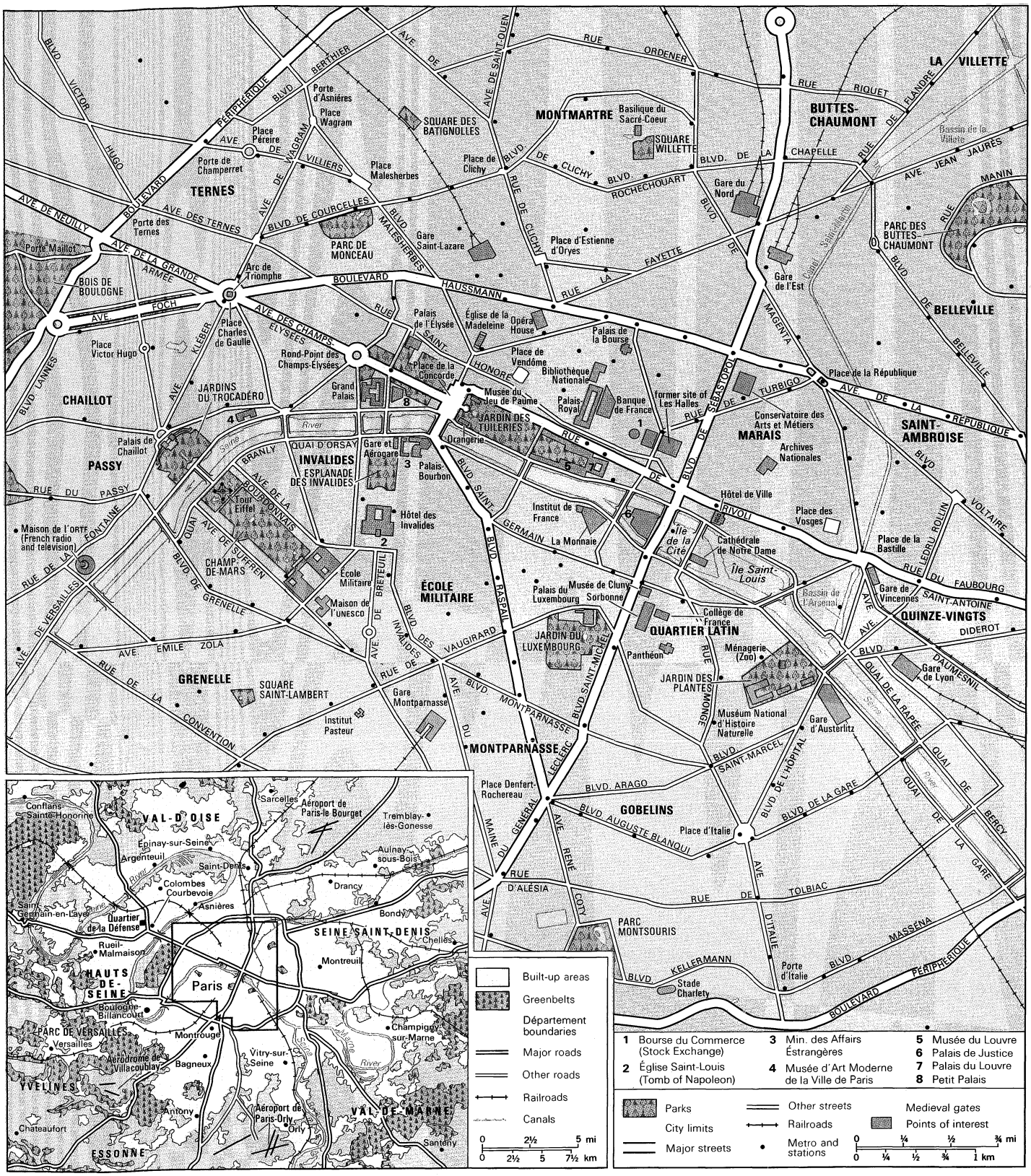
Can it be said that the existence of psi has been proved? No simple answer can be given because different people require different standards of proof. In the sense that there is much experimental evidence, collected under sound and well-controlled conditions and properly analyzed statistically, which supports such a view, the answer must be that ESP certainly, and PK probably, does exist. If to prove the existence of psi it is necessary to have replicable phenomena that can be demonstrated with certainty and that behave according to certain well-known laws, the answer must be in the negative. This position is not unusual in science, particularly in the initial phases of research; stellar parallax was searched for but not definitively discovered for over 200 years, although it formed an essential part of Copernicus' heliocentric theory. Neutrinos, having no charge and hardly any mass, were difficult to find and pin down; antimatter has a similar shadowy existence. The best summing-up might be that very intriguing demonstrations have been given that suggest the existence of something outside the purview of orthodox physics and psychology but that no one has yet succeeded in bringing this something under adequate experimental control. Until this is done and until the ways in which this something responds to changed experimental conditions are known, and until it is possible to formulate quantitative theories that are not simply question begging, it would be unwise to claim any more. Certainly it is premature to look for proof of immortality or survival of the soul in experiments with dice and playing cards; neither is it reasonable to claim that these findings undermine modern science. The relevance of these phenomena to either cannot reasonably be discussed until a better understanding of their nature has been achieved—assuming always that they exist at all and are not the product of imagination, cheating, and chance. The probability that these factors are responsible for all that has been discovered is small, but it cannot be completely excluded.

BIBLIOGRAPHY. A popular account and discussion of the field is given in H.J. EYSENCK, *Sense and Nonsense in Psychology* (1957). A thorough review of the experimental literature, written from a point of view favourable to parapsychology, is K.R. RAO, *Experimental Parapsychology* (1966). It is complemented by C.E.M. HANSEL, *ESP: A Scientific Evaluation* (1966), a highly critical review. J. EISENBUD writes on *The World of Ted Serios: Thoughtographic Studies of an Extraordinary Mind* (1967); D.B. EISENDRATH, JR., and C. REYNOLDS give a critical account of their study, "Spend an Amazing Weekend with the Amazing Ted Serios," *Popular Photography* (Oct. 1967). A theory of parapsychology and descriptions of some famous mediums are presented in C. MCGREERY, *Science, Philosophy and ESP* (1967). Two works by parapsychologists discussing their own experiments are J.B. RHINE and J.G. PRATT, *Parapsychology: Frontier Science of the Mind* (1957); and S.G. SOAL and F. BATEMAN, *Modern Experiments in Telepathy* (1954). A philosophical approach is taken by C.D. BROAD in *Lectures on Psychical Research* (1962).

(H.J.E.)

Paris

Founded on the island where a natural north–south highway crosses the Seine River, some 233 miles (375 kilometres) upstream from the river mouth on the English Channel, Paris, the largest city proper of continental Europe and the capital of France, is over 2,000 years old.



The city of Paris and (inset) its metropolitan area.

THE CHARACTER OF THE CITY

For hundreds and hundreds of years, by a process never successfully explained, Paris has radiated an enchantment irresistible to millions around the world, including hosts of people who would live and die without ever seeing the place. In the early 1970s, however, this magical Paris exists almost exclusively in the hearts of the believers.

"Gay Paree"—that honeycomb of breathtaking women and women-taking men, capital of the arts, of savoir-faire (literally, "knowing-how-to-do") and savoir-vivre ("knowing-how-to-live")—survives only in scattered, scanty fragments. It has shrunk to the dimensions of the same small international set who made it gay in the 16th century, together with an excursion version for the tourist. The naughtiness of French bedroom farces, French cabaret nudity, and French yellow-backed novels has long

The Parisian image

since been rendered innocuous by the "permissive society" of other rich Western nations. The celebrated *cocottes* have become anonymous call girls, the famous *maisons de tolérance* were closed down in 1948, and there remains nothing unique about commercial vice in Paris. Even the practice of amateur dalliance has altered: the traditional *cinq-à-sept* (literally "five-to-seven"; *i.e.*, those hours, intervening between office and home, used for mistress-visiting) has, it appears, been abandoned.

Once capital of the arts, Paris has become provincial. During the 19th century and halfway through the 20th, Paris was the great garden of creative talents. Painters and sculptors, writers, musicians, and dancers came from around the world, hoping to blossom in the incredibly propitious intellectual climate. By the 1970s, however, nothing was left but nostalgia.

Savoir-faire has become largely managerial-technological in orientation and is the specialty of other cities in other countries. Savoir-vivre, too, in Paris, as in every other major city, has been outstripped by the savoir-faire.

The city has always been quick to anger. In spite of the French *mission civilisatrice*, no year in Paris has passed without some manifestation of mass violence. In this, contemporary Paris remains unchanged.

As Parisians themselves took note in the late 1960s, when they launched a campaign in favour of courtesy, their traditional noninterference in the lives of others had degenerated into indifference. By the 1970s the satisfaction of a legitimate demand in Paris often required either bluster or wheedling, so that a perfectly banal transaction seemed to become a conspiracy or a contest.

The delivery boy seems to have ceased his whistling, and the house painter to have stopped his habitual singing. In the past, even though life was often brutal for the masses, the small people of Paris seemed as much under its spell as anyone else. The city seemed to bewitch even those whom it devoured, and songs celebrating the joys of Paris bubbled up from the most wretched of corners.

Despite this devotion, Parisians began leaving Paris after 1921, the year that the population attained the total of 2,906,500. By 1971 the number of inhabitants (2,550,000) was lower than it had been in 1895. As more dwellings disappeared to make room for offices, the remaining population was squeezed into smaller areas of the city. The very poorest—mostly immigrant labourers—could not cram themselves in at all: between 25,000 and 35,000 existed in miserable circumstances in the more than 150 shanty towns (*bidonvilles*) huddled outside the city limits. With more people leaving Paris, the notably morose fringe suburbs began to spread like a fungus onto the rolling richness of the Île de France.

A crisis of expansion

To save the surrounding spaces from speculator-induced blight and to save Paris itself from the epidemic ills of great cities in mid-20th century, the government went into action as early as 1958. Parisians and suburbanites were cynical. Previous government efforts in the suburbs had been catastrophic, resulting in barren collections of housing units intended to receive a maximum number of population units. In the city itself the last improvement of first magnitude had been made in 1898, when—after a nasty 26-year fight—tunnelling began for the métro (*chemin de fer métropolitain*, designated the *Nécropolitain* by its opponents).

Table 1: Paris, General Data

Geography					
Altitude			85–128 feet (26–128 m)		
Area					
City proper (Ville de Paris)			41 sq mi (105 sq km)		
Metropolitan area (Région Parisienne)			4,636 sq mi (12,008 sq km)		
Population Growth					
year	population (000)	year	population (000)	year	population (000)
1801	548	1891	2,447	1946	2,691
1831	786	1911	2,888	1954	2,821
1851	1,053	1921	2,906	1962	2,780
1872	1,851	1936	2,830	1968	2,591
			(Région Parisienne)	1968	9,251

THE REGION

This time, given the personal intervention of Pres. Charles de Gaulle and his successor, the action was carried through. The area in question was huge—2 percent of the entire national territory—comprising 4,600 square miles (12,000 square kilometres) and including 1,305 communes with a total (1971) population of 9,690,000. It was decided to develop the area as a region, rather than to extend the city limits to include the suburbs, thus forming overnight a grotesque city a dozen times its present size. The next move was to give the new region coherence, authority, and autonomy, enabling it to operate as an entity and to insure its survival against the hostility of often powerful interests. To accomplish this, new political units were devised, equipped with new administrative machinery, and provided with ample legal and financial powers.

By 1968 two of the three *départements* comprising the region had been divided into seven new units: Seine became Ville-de-Paris, Hauts-de-Seine, Seine-Saint-Denis, and Val-de-Marne; Seine-et-Oise was carved into Essone, Yvelines, and Val-d'Oise. Seine-et-Marne was left intact. There were now eight prefects, instead of a mere three, to administer the territory, and as a common financial tool they had the district of the Paris region, which was created as an autonomous public body in 1961. Deriving its considerable treasury from a special regional tax (the only such in France), it invests directly in local development projects through its administrative council of mayors and departmental councilmen.

Studies were undertaken to establish what the city and region contained, how it was being used or abused, and what solutions were possible to problems thus identified and defined. An early result was the discovery that previous estimates of future population were grievously short and that the region would probably have 14,000,000 to 16,000,000 inhabitants by the year 2000.

A master plan for the region was made public in 1965 and that for the city of Paris in 1968. The regional plan bore an epigraph from Seneca, "It is not because things are difficult that we do not dare: it is because we do not dare that they are difficult." The most daring proposal was for the establishment of five new cities on two axes of urbanization extending both north and south of Paris.

These are: to the southeast, Évry (with 300,000 inhabitants) and Melun-Sénart (400,000); to the east, Vallée de la Marne (500,000 inhabitants in four settlements separated by parkland); to the west-northwest, Cergy-Pointoise (100,000 inhabitants); and, finally, to the south-southwest, Trappes (with six major settlements and a total of 350,000 inhabitants). The architecture is bold, the siting combats monotony, and the planners start with the essentials of good small-town life, attempting to find the modern equivalents of the parish pump, the village green, and the old swimming hole. The regional plan makes sweeping provisions for green spaces, and the development of recreational, educational, and hospital facilities, while highway construction and public transport are also an integral part of the scheme.

The power of the *préfet*

In 1966, the Cabinet named a *préfet de la région parisienne*, who commanded powers greater than the 100 other prefects of France. To give him financial leverage, he also was named delegate general of the district. The eight departmental prefects take orders from him in matters relating to planning and development, and he has authority over the public transit system, the Port of Paris, the Public Housing Office, and other public bodies and teams of experts dealing with plans for expansion. Since ministers can initiate or quash projects, or withhold funds, he also was given a seat on the interministerial committees. By 1968 the new Paris began to emerge.

THE CITY PROPER

By the early 1970s the heart-shaped city, comprising the 20 *arrondissements* (wards) threaded by the Seine, had already lost its cherished skyline. Clutches of 30-story towers (and one of 56 stories) look down upon scores of 16-story apartment houses. The clover-leaf, the elevated highway, the underpass have arrived on the outskirts.

Table 2: Paris, Transport Data

	air		rail		road	
	(000)	(percentage)	(000)	(percentage)	(000)	(percentage)
Passengers						
Inbound	5,237	48	...	...	...	...
Outbound	5,304	49	...	...	...	...
Transit	341	3	...	...	...	...
Total	10,882	100	380,500	...	8,720,000	...

	air		rail		water	
	metric tons (000)	(percentage)	metric tons (000)	(percentage)	metric tons (000)	(percentage)
Freight*						
Inbound	99	45	...	...	24,577	90
Outbound	119	55	...	...	2,586	10
Total	218	100	...	...	27,163	100

*Including mail.
Source: *Annuaire statistique de la France 1970–71; Airport Traffic, 1969*, International Civil Aviation Organization.

A progress report from one of the planning bodies expresses the hope that the inhabitant may "recognize the Paris he does not know through the Paris that he loves." The Paris that he loves was in large measure created from 1835 to 1870 by Baron Georges Haussmann, the iron-fisted *préfet de la Seine* under Napoleon III. There were cries of alarm from the more traditionally minded when he slashed the boulevards through the tangles of slums (at the same time beginning the modern sewer and water systems), gutted the Île de la Cité, the heart of Paris, rebuilt the ancient market of Les Halles and constructed the Opera House (Opéra), as well as adding four new Seine bridges and rebuilding three old ones. The modern architecture that bordered his new streets was for the most part pretentious and lugubrious, but Paris remained lovable for all that. Tomorrow's Paris—the master plan carries to the year 2000—will have harder work being seductive. The planners' announced aim is to make a Paris "more human" than the one given into their hands.

According to them, the old Paris is dying of "asphyxiation." They envisage not only a cure for the malady but also a restoration of some ease and elegance to Paris life.

The principal visible agent of suffocation is traffic, human and vehicular. Under its pressures rush hours last longer and start earlier in the slowing flow of 2,000,000 commuters. Of the travellers, 900,000 come in from the suburbs—334,000 by railway and about 271,000 by automobile—and the rest come by subway and bus from the immediately adjacent communities. In 1970 a report by the prefect of the region said that street traffic was approaching the "limit of the absurd."

The traffic crisis

The daily migration is further complicated by the fact that most of the city's employment opportunities lie west of the Cathedral of Notre-Dame, while the greater part of the population lives east of that central point. Four of the nine stations of the suburban railways are also west of the centre. Humanity thus sloshes back and forth across the city at peak hours like slurry in a basin. A frequent plaint scrawled on the métro walls is: "Métro-Boulot-Dodo" (*boulot* is slang for "work"; *dodo* is baby talk, like "beddy bye"); a free translation might chronicle the unhappy commuter's day as "Subway line, salt mine, beddy mine."

To reduce this daily abrasion, the authorities are promoting the establishment of new "poles of attraction" in the west of Paris, in the shape of skyscraper commercial complexes sited around the railway stations. New westward office construction is slowed through restriction of building permits and increased taxation. Outlying parking garages (*parkings de dissuasion*) have been constructed in the hope that drivers will leave their cars at the city gates. The first (east–west) leg of the regional subway system called the Réseau Express Régional (RER), a high-speed subway in its own deep-level tunnel, was put into service from the suburbs in 1970. More métro lines —and they are all being modernized—have been extended into abutting municipalities.

The last of the much-romanticized open-platform buses made their final run in 1970. While London was experimenting with small single-deck red buses, manned only by the driver, Paris was experimenting with big green double-deckers, which require the services of conductors as well. The number of miles of traffic lanes for the exclusive use of buses and taxis was tripled in 1971 to get the average speed above seven miles (ten kilometres) per hour.

All projects to run express highways through the city have been vetoed by the Paris municipal council. An expressway (built 1961–72) encircles the city, following the line of the old ramparts. Other highways will mesh the Paris region in a spider-web design centred on the capital.

For in-town parking, many leafy Paris squares have been dug up, scooped out, and re-covered, often with disastrous defoliatory results, although public outcry at the start of the 1970s forced the city to abandon plans to eviscerate some favourite neighbourhood squares. The giant new buildings, however, are equipped for underground parking.

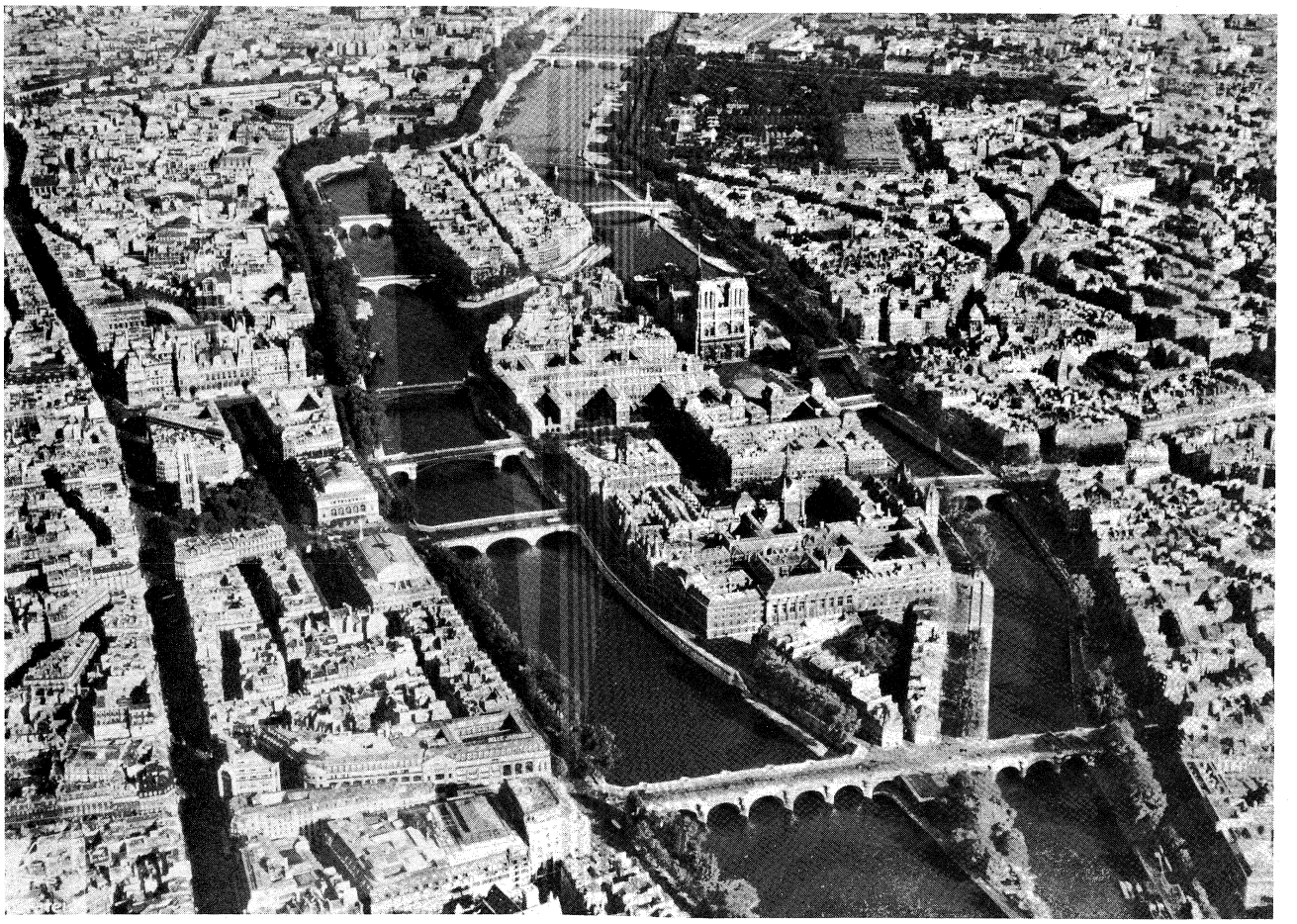
The plan for the year 2000 provides for more of everything the city needs: hospitals, schools, housing, day nurseries, additional centres of sport and leisure, and more green space. (For related information see FRANCE; FRANCE, HISTORY OF; and SEINE RIVER.)

THE HEART OF THE CITY

Paris and its river. Paris is small; no corner is farther than six miles from the square in front of Notre-Dame Cathedral. The city has a total area of 41 square miles (105 square kilometres), if the two big parks at either extremity are included, and 34 square miles without them. The city occupies a bowl hollowed out by the Seine in its prehistoric vigour, and the surrounding heights have been respected as the limits of the city. The river arches through the centre of town, visiting 10 of the 20 *arrondissements*. Entering the city at the southeast corner, it arcs northward and bends out of Paris at the southwest corner. As a result, what starts out as the stream's east bank becomes its north bank and ends as the west bank, and the Parisians therefore adopted the simple, unchanging designation of Right and Left Bank (when facing downstream). These terms are not much used in conversation, as specific places are usually indicated by *arrondissement* (*e.g.*, *quinzième*) or by *quartier* (*e.g.*, *Observatoire*).

The river bank

At water level, some 30 feet below street level, the river is bordered—at least on those portions not transformed into expressways—by cobbled quays graced with trees and shrubs. From street level another line of trees leans toward the water. Between the two levels, the retaining walls, usually made of massive stone blocks, are decorated with the great iron rings of a past age's commerce and sometimes pierced by mysterious openings (water gates for old palaces or inspection ports for subways, sewers,



Île de la Cité and Île Saint-Louis, on the Seine River at the heart of Paris.
By courtesy of the French Government Tourist Office

and underpasses). Here and there the wall is shawled in ivy.

The old buildings, the riverboats, the changes of colour reflected by the water, the gardens, and the 32 bridges (many of them handsome) compose one of the world's grandest, yet most endearing cityscapes. Along the river are two of the great set pieces of urban spectacle in the contemporary world. The first sweeps down from the Palais de Chaillot on the Right Bank, crosses the river to the Eiffel Tower, and continues through the gardens of the Champ-de-Mars to the 18th-century École Militaire; the other begins at the Seine and marches up a broad esplanade to the golden dome of the Invalides.

The Palais de Chaillot dates from the International Exposition of 1937 and is a period piece of between wars, timid-modern style. It replaces a structure of tepid Moorish sympathies left over from the 1878 International Exposition. Earlier in the 19th century, after demolition of the Convent of the Visitation, the top of this 230-foot (65-metre) hill had been levelled for the construction of a palace (never built) for the King of Rome, son of the emperor Napoleon.

The Palais is made of two separate pavilions, each of which sprouts a curved wing. The Musée de l'Homme (Museum of Man), the Musée de la Marine, and the Musée des Monuments Français (Museum of French Monuments) are located there. Under the terrace which separates the two sections are two theatres, the variable-formation (1,500 to 3,000 seats) National Popular Theatre (TNP) and a small hall that serves as one of the two cinemas of the National Film Library (Cinémathèque Française).

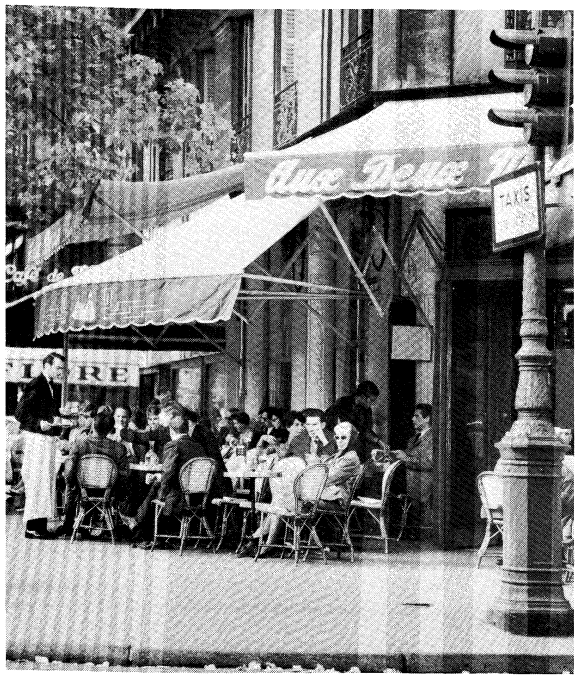
The statue-guarded terrace gives a splendid view across Paris and makes an enduring travel-poster setting for photographs of visitors and fashion models. The hill descending to the river has been made into a terraced park, the centre of which is adance with mighty fountains, cascades, and pools. The Paris aquarium is in a grotto to the left.

One of the enchantments of the view—and some others in Paris—is that it has all the qualities of a *trompe-l'oeil* (literally, "deceive the eye") painting into which, extraordinarily, one can walk. From the bottom of the hill the five-arched Pont d'Iéna springs across the river. It was built for Napoleon I in 1814, although the imperial "N's" with which it is decorated were in fact put there by Napoleon III. After the bridge comes the unclad metal truss tower of Gustave Eiffel. It was built for the International Exposition of 1889, against the strident opposition of national figures who believed it to be unsafe or ugly, or both. When the exposition concession expired in 1909, demolition of the 984-foot (300-metre) tower was averted by demonstration of its value as an antenna for the newly developed radio. Additions made for television transmission have added 56 feet (20.75 metres) to the height. From the topmost of the three platforms the view extends for 50 miles—when air pollution is low and the sun is near the horizon.

The Eiffel Tower

From the two-acre base of the tower the Champ-de-Mars stretches inland to the École Militaire (built 1769–72 and still used by the École Supérieure de Guerre [War College]), where the 15-year-old cadet Napoleon Bonaparte was enrolled in 1784. Originally the school's parade ground, the field was the scene of two vast revolutionary rallies, that of the Federation (1790) and that of the Supreme Being (1794). From 1798 there were annual national expositions of crafts and manufactures, followed by world's fairs between 1855 and 1900. The International Exposition of 1937 spread around it, for between 1908 and 1928 the field had been made into a formal park.

Behind the École Militaire, which was designed by Gabriel, architect of the Place de la Concorde, stands the Y-shaped headquarters of the United Nations Educa-



(Left) Aux Deux Magots, a famous Paris sidewalk cafe. (Right) The fountain and obelisk in the Place de la Concorde.
(Left) Roger-Viollet, (right) Josef Muench

tional, Scientific, and Cultural Organization. The building, erected in 1958, was designed by an international trio of architects and decorated by artists of member nations.

The Invalides. Just one street away to the northeast is the Hôtel des Invalides, founded by Louis XIV to shelter 7,000 aged or crippled former soldiers. The enormous range of buildings was completed in five years, (1671–76). The gold-plated dome (six kilograms of gold leaf were involved) that rises above the hospital buildings belongs to the Church of Saint-Louis (1675–1706), designed by Jules Hardouin-Mansart. The architect employed a style known in France as "Jesuit," since it derives from the Jesuits' first church in Rome, built in 1568. The churches of the Académie Française (French Academy), the Val-de-Grâce Hospital, the Sorbonne, as well as three others in Paris, all of the 17th century, followed this style. By a freer use of the classical elements, the French made it something recognizably Parisian.

Napoleon's tomb

In the chapels of Saint-Louis are the tombs of Napoleon's brothers Joseph and Jérôme, of his son (whose body was returned from Vienna in 1940 by Adolf Hitler), and of the marshals of France. Immediately beneath the cupola is a red porphyry sarcophagus that covers the six coffins enclosing the body of Napoleon I, which was returned from Saint-Helena in 1840 through the efforts of King Louis-Philippe. Napoleon's uniforms, personal arms, and death bed are displayed in the rich Musée de l'Armée (Army Museum) at the front of the Invalides. Fewer than 100 pensioners now live at the hospital, which is used as a paraplegic centre.

The grassy, tree-lined Esplanade des Invalides (810 feet wide) slopes gently for 1,470 feet to the Quai d'Orsay and the Pont Alexandre III. The first stone for the bridge was laid in 1897 by Alexander's son, Tsar Nicholas II. A steel span with upper works of stone, it embodies the "Gay Nineties" (*la Belle Époque*), solid, sumptuous, and luxuriant, with its pomposity mocked by its own gaiety. Finished in time for the International Exposition of 1900, it leads to two faded souvenirs of that year's fair, the Grand Palais and the Petit Palais. Both are still used for seasonal painting salons and major visiting art exhibits, and the Grand Palais also shelters overflow classes from the Sorbonne and a science museum.

For millions around the world the name "Paris" connotes an image of a mile-long stretch of the Seine between the Pont du Carrousel and the Pont Sully. The city's most celebrated bridge, its most famous museum, its most admired Gothic churches are all found here in the ancient heart of the capital. Here, too, are the quayside bookstalls, the bird and flower markets, and the sempiternal anglers of the Seine.

The Louvre. The Vikings camped on this spot on the Right Bank in their unsuccessful siege of Paris in 885, and in 1220 King Philippe-Auguste selected it to plant a square crusader's castle before the new city wall, as key to the western defenses. Through subsequent centuries and the attentions of 11 monarchs, the château-fort was made into one of the world's biggest palaces, completed in 1852. The last of the Gothic portions disappeared only in 1673. From the original square, two galleries extend westward for 1,640 feet, one along the river, the other along the rue de Rivoli. Only 19 years after the huge oblong was finally completed, its western face, the Tuileries Palace (begun 1563), was destroyed by the insurrectionists of the Commune, in 1871.

The Cour Carée

Two of the facades of the original palace square, the Cour Carée, have considerable architectural importance and were strong influences on the development of French styles. Pierre Lescot began his inner courtyard facade in 1546, adapting the Renaissance rhythms and orders he had observed in Italy, and adding purely French decoration to the classical motifs. Claude Perrault, also distinguished as an anatomist and court physician, created a masterpiece for the outer east face of the palace in 1673. It, too, employs classic elements—coupled columns and a pediment—but they are handled with a grace and originality that makes it perfect for the late reign of the Sun King.

For so many centuries the seat of French power, the Louvre of the early 1970s still contains some national administrative offices in the rue de Rivoli gallery, where a separate museum—the Musée des Arts Décoratifs—is also housed. The gallery is an awkward 19th-century counterfeit of the riverside arm (Galerie du Bord de l'Eau) built in the last half of the 16th century. The Louvre, known formally as the Musée du Louvre, occupies the older gallery, the palace around the Cour Carée, and, at the far end of the Tuileries Garden, the galleries of the Orangerie and the Musée du Jeu de Paume. Among the treasures of the museum—one of the world's greatest—are the Victory of Samothrace, the Venus de Milo, and the Mona Lisa. The enormous collections contain works from 7 BC to the mid-19th century, with a

huge cultural and geographic spread. The École du Louvre, in addition to its regular university-level curricula, offers free public evening classes in art history.

Crossing from the Louvre to the Institut, the Pont des Arts is one of the most charming of all the Parisian bridges. It was the first (1803) to be made of iron, and has always been reserved for pedestrians: it provides an intimate view of riverside Paris and of the Seine itself.

Institut de France. Since 1806, these Left Bank buildings have been the home of the five French academies. Originally, a defense work for the Left Bank terminus of the 1220 city wall stood here. Named the Tour de Nesle, it was for centuries celebrated in romantico-erotic song and story. Louis Le Vau, architect to the king, designed the new buildings in 1663 to house the Collège des Quatre-Nations, paid for by a legacy from Cardinal Mazarin who had brought the four entities in question—Piedmont, Alsace, Artois, and Roussillon—under the French crown. Again, Italian notions were subtly gallicized: from a domed church, twin wings arch out to twin pavilions, more suavely and less heavily than in the Italian models. The five contemporary academies are: the Académie Française, founded by Richelieu in 1635, which edits the official French dictionary (the two-volume 8th revised edition was published in 1932 and 1935), awards literary prizes, and has a membership of "40 Immortals"; the Académie des Inscriptions et Belles-Lettres, founded 1663 by Colbert; the Académie des Sciences, founded 1666, also by Colbert; the Académie des Beaux Arts, two sections formed at different times by Mazarin and Colbert and joined in 1795; and, finally, the Académie des Sciences Morales et Politiques, created by the Convention in 1795 to ponder questions of philosophy, economics, politics, law, and history. After he became one of the Immortals, François Mauriac was asked what they did at their meetings, and he said, "We watch ourselves grow old."

The Académie Française

Almost next door is La Monnaie (the Mint), a sober late-18th-century building where visitors may watch coins being struck and visit a museum of coins and medals.

Pont Neuf ("New Bridge"). This is the oldest (1578–1604) of the Paris bridges, the sturdiness of which has become literally axiomatic: Parisians still say, "Solid as the Pont Neuf." It crosses the broadest part of the river and is supported in the middle by the tip of the Île de la Cité, to which it extends five arches from the Left Bank and seven from the Right. The parapet corbels are decorated with more than 250 different grotesque masks, and the parapet curves out toward the water at each bridge pier, forming half-moon bays along what was the first sidewalk in Paris; it was in these that the street vendors and medicine men set up shop. For 200 years this bridge was the main street and the perpetual fair of Paris. Although the structure undergoes regular repair, in the main today's Pont Neuf is the original bridge.

At the island crossing the bridge touches the spot where Jacques de Molay, Grand Master of the Knights Templars, was executed March 11, 1314, after Philip the Fair (Philippe le Bel), King of France, and Pope Clement V agreed to suppress the financially and politically powerful order. Downstream and just below the bridge is the prow of the island, fashioned into a triangular gravel-pathed park bordered by flowering bushes, with rustic benches under the ancient trees. It bears the nickname of Henry IV, being titled Square du Vert-Galant (Square of the Green Gallant). It is surrounded by a wide cobbled quay much used for sunbathing and lovemaking.

Where the steps come onto the bridge from the park there is a bronze equestrian "statue of Henry IV," who insisted on completion of the Pont Neuf. It is an 1818 reproduction of the 1614 original, the first statue to stand on a public way in Paris. Opposite the statue is the narrow entrance to the Place Dauphine (1607), named for Henry's heir. The *place* was formerly a triangle of uniform red-brick houses pointed in white stone, but the row of houses along its base was ripped out in 1871 to facilitate the exposure of the new rear elevation of the Palais de Justice.

Île de la Cité. This ship-shaped island is ten streets long and five wide, with eight bridges to the riverbanks and a ninth to the scow-shaped Île Saint-Louis, which lies abaft to starboard. The first recorded name (*c.* 3rd century BC) for the settlement of rivermen on the island was Lutèce, which meant midwater-dwelling. When the Romans arrived, the Parissi tribesmen were already sufficiently organized and wealthy to have their own superior gold coinage. Caesar's *Commentaries* record that in 52 BC the inhabitants burned their town rather than surrender it to the Romans.

There is some evidence that the Roman governors' palace was on the site of the present Palais de Justice, although their new town spread up the hill on the Left Bank, the present Latin Quarter (Quartier Latin). By the end of the 5th century the Salian Franks, under Clovis, had conquered Paris, which they made the capital of Gaul. It remained the capital until the end of Chilperic's reign in 584, but succeeding Merovingians carried the crown elsewhere and the town's political importance diminished. Charlemagne's line, the Carolingians, left the city in the charge of the counts of Paris who eventually took the town and the kingdom for their own. Their power was intermittent until the election of Hugues I, Capet to the throne in 987, when the palace in Paris again became the seat of power.

Foundation of the city

The palace was rebuilt on the same site by St. Louis (Louis IX) in the 13th century, and enlarged 100 years later by Philip the Fair, who added the grim, gray-turreted Conciergerie, with its impressive Gothic chambers. The Great Hall (Grand Salle) was known throughout Europe for its Gothic beauty, but today, after the fires of 1618 and 1871 (most of the rest of the palace was devastated by flames in 1776), there is almost nothing left of the original room, which, under the kings, was the meeting place of the Parlement (the high court of justice). It serves now as a waiting room for the courts, in one of which, the adjoining first Civil Chamber, the Revolutionary Tribunal sat from 1793, condemning 2,600 persons to the guillotine. After sentencing, the victims were taken back down the stone stairs to the dungeons of the Conciergerie to await the tumbrils. The Conciergerie still stands and is open to visitors; the Gothic Salle des Gens d'Armes, restored in 1970 along with the royal kitchens, can be hired for "receptions of quality."

After the counsellors of the Dauphin Charles (later Charles V) were slain before his throne in 1358, the 20-year old regent turned the palace over to the Parlement and went to live in his new Right Bank residence, the Hôtel Saint-Paul. No monarch ever again lived in the palace on the Île de la Cité.

Tucked away in the palace courtyards is one of the great monuments of France, the 13th-century Sainte-Chapelle. Built between 1243 and 1248, it is a masterpiece of Classic (known in French as *rayonnant*, radiant) Gothic. With great daring, the architect (possibly Pierre de Montreuil) poised his vaulted ceilings on a trellis of slender columns, the walls between being made of stained glass. This jewel box was designed to hold the Crown of Thorns, bought from the Venetians, who had it in pawn from Baldwin, the Latin king of Byzantium. Other holy relics, such as nails and pieces of wood from the True Cross, were added to the chapel's miraculous collection, the remnants of which are now in the treasury of Notre-Dame.

Under King Louis-Philippe (1830–48), the "sanitization" of the island was begun and continued for his successor, Napoleon III, by Baron Haussmann. It was a mass clearing of antique structures, widening of streets and squares and the erection of ponderous new government offices, including parts of the Palais de Justice. That portion which borders the Quai des Orfèvres is headquarters of the Paris municipal detective force, the Police Judiciaire, which keeps a small museum on the fourth floor.

Across the Boulevard du Palais is the Préfecture de Police, another 19th-century construction. Paris is the only French city to have its own police prefect, and his authority in the capital is broad, covering a multiplicity

Police agencies

of duties that in other cities usually devolve on other officials. In addition to dealing with crime, traffic, and public order, the Paris police register all vehicles; license drivers; issue passports, identity cards, and aliens' residence permits; and conduct political surveillance. Civil defense and many aspects of public health are part of police duties, together with some responsibility for fire fighting, although Paris firemen are, in fact, an army unit. The Garde Républicaine are detached for service to the prefect, but this mounted squadron of spurred, helmeted, plumed, breastplated sabre-flourishers are really *gendarmes,* members of the national rural police run by the Ministry of Defense.

There are several other police agencies in Paris, including several shadowy investigatory and counter-espionage agencies as well as the famed national detective force, the Sûreté Nationale, and the Compagnies Républicaines de Sécurité (CRS), whose duties range from teaching youngsters to swim to dispersing demonstrations. The Sûreté and CRS are responsible to the Minister of Interior, who also appoints the *préfet de police,* the *préfet de la Ville-de-Paris,* and the 20 mayors in the town halls of the *arrondissements.* Some ministers take a strong personal role in the policing of Paris, but at times the prefect of police himself has been an important figure in the city's ever-turbulent life.

It was the police who hoisted the tricolour above the Préfecture in August 1944 to begin the insurrection that led to the liberation of Paris. They held out against the Germans for three days until the arrival of French and American forces.

On the far side of the Préfecture is the Place du Parvis-Notre-Dame, an open space enlarged six times by Haussmann, who also moved the Hôtel-Dieu, the first hospital in Paris, from the riverside to the inland side of the square. Its present buildings date from 1868.

Notre-Dame de Paris. At the east end of the square is the Cathedral of Notre-Dame, sited on a spot Parisians have always reserved to the practice of religious rites. The boatmen of the *cité* erected their altar to Jupiter there (it is now in Cluny Museum), and, once Christianity was safely established, a church was built on the temple site. The first bishop of Paris, St. Denis (Dionysius), became its patron saint. The red in the colours of Paris represents, in popular legend, the blood of this martyr who, after decapitation, picked up his head and walked.

When Maurice de Sully became bishop in 1159 he decided to replace the decrepit Cathedral of Saint-Étienne (St. Stephen's Cathedral) and the 6th-century Notre-Dame with a new church in the newest style, Gothic. The style was conceived in France, and a new structural development, the flying buttress, which added to the external beauty and permitted interior columns to soar to new heights, was introduced in the building of Notre-Dame. Construction began in 1163 and terminated in 1345.

After being damaged by the celebrated crowds of the French Revolution, the church was sold at auction to a building-materials merchant: fortunately, Napoleon came to power in time to annul the sale and ordered the edifice redecorated for his imperial coronation in 1804.

Restoration of Notre-Dame

Before abdicating in 1848, Louis-Philippe, a king whose architectural services to Paris have gone largely unrecognized, initiated restoration of the dangerously neglected church. The architect Eugène Viollet-le-Duc worked from 1845 to 1864 to restore the monument.

Like all cathedrals in France, Notre-Dame is the property of the state, although its operation as a religious institution is left entirely to the church.

A few 16th- and 17th-century buildings still remain north of the cathedral. They are remains of the cloister of the cathedral chapter, whose school was famous long before the new cathedral was built. Early in the 12th century, one of its theology professors, Abélard, upholder of the conceptualist thesis, left the cloister with his disciples, crossed to the Left Bank and set up an independent school in the open air near the present Place Maubert. After a prolonged struggle with the Notre-Dame authorities in matters of doctrine and academic freedom, the followers of Abélard (died 1124) won the right, from king and pope, to form and govern their own community in 1200. This was the start of the University of Paris. Abélard is best remembered for the romantic tragedy he acted out with Héloïse, niece of the vengeful Fulbert (Canon of Notre-Dame).

City walls. After the Roman town on the Left Bank was sacked by barbarians, some time between 250 and 275, the fire-blackened stones were freighted across to the Île de la Cité, where a defensive wall was constructed. Neglected in times of peace, it was rebuilt several times over the course of the centuries. The bridge to the Left Bank (Petit Pont, first of the Paris bridges) was guarded by a fortified gate, the Petit Châtelet, and that to the Right Bank (Pont au Change) by the Grand Châtelet, a fort, prison, torture chamber, and morgue, finally demolished in 1801.

In the years 1180 to 1225, Philippe-Auguste (Philip II Augustus) built a new wall that protected the settlements on both banks. In 1367–70 the Right Bank *enceinte* (enclosure) was greatly enlarged by Charles V, with the massive Bastille protecting the eastern approaches as the Louvre protected the west. By the late 14th century Paris was undoubtedly the largest town in France, with a population between 150,000 and 200,000. Louis XIV had the Charles V extensions transformed into the primitive form of the tree-planted Grands Boulevards in 1670, embellished at the Porte Saint-Denis and the Porte Sainte-Antoine with triumphal arches still standing. The word *boulevard* was a military engineering term for the platform of a defensive wall. As the city grew, so did its population, which is estimated to have reached half a million by the end of the 17th century.

Another century passed before a new wall was begun, enabling the *fermiers généraux* (farmer-generals) to collect customs duties on goods entering Paris (they gave the royal treasury a flat annual sum in anticipation of customs revenues and pocketed the difference, which made them all extraordinarily wealthy men). The tollhouses are still standing at Place Denfert-Rochereau. The pre-Revolutionary population of the city was probably around 600,000, although the first official census in 1801—taken after many citizens had fled during the Revolutionary turmoil—showed a figure of 548,000.

Table 3: Paris, Social Data

rooms	number of dwellings (000)	number of dwellings percentage
One	360.4	32
Two	383.2	33
Three	225.7	20
More than four	169.7	15
Total	1,139.0	100

Total number of rooms	2,643,528	Rooms per dwelling	2.32
Total number of persons	2,513,588	Persons per room	.95

	number of dwellings (000)	number of dwellings percentage
Equipment of Occupied Conventional Dwellings		
Running water (inside)	1,074	94
Toilet with flushing	630	55
Bath or shower	491	43
Bath or shower and toilet	431	38
Central heating	546	48
Telephone	372	33
Total	1,139	100

	live births	deaths	excess of births over deaths
Natality and Mortality			
City proper			
Absolute	37,140	28,800	8,340
Per 000 inhabitants	14.5	11.3	3.2
Metropolitan area			
Absolute	102,456	62,261	40,195
Per 000 inhabitants	15.9	9.7	6.2

Source: *Annuaire statistique de la France 1970–71; Annuaire de statistique internationale des grandes villes, 1968; Statistics of World Large Cities, 1971.*

The last wall

The last wall, built by Adolphe Thiers for Louis-Philippe, was a genuine military installation with outlying forts. By the time it was finished it enclosed more than a dozen hamlets outside Paris: Vaugirard, Grenelle, Auteuil, Passy, Les Batignolles, Montmartre, La Chapelle, La Villette, Belleville, Charonne, Gobelins and Observatoire. Post-Revolutionary rebuilding and post-Napoleonic economic recovery, with expansion of employment due to the Industrial Revolution, drew more and more people to Paris, and—as railways developed—with ever-increasing facility. In 1851 the city population had reached 1,053,000. Between 1852 and 1870 Haussmann razed the walls of the *fermiers généraux* and built the *boulevards extérieures*. Parisians numbered 1,852,000 in 1872, and 2,447,000 20 years later. The 1845 walls were finally knocked down and made into boulevards yet more exterior in 1925. By that year the population, after reaching a peak in 1921, had already begun the decline which has continued ever since.

Île Saint-Louis. In 1627, Louis XIII accorded a 60-year lease on two mudbanks behind the Île de la Cité to a contractor, Christophe Marie, and two financiers. It was 37 years before Marie was able to unite the islets, dike the circumference, lay out a central avenue with 10 lateral streets, and rent space to some householders. The Church of Saint-Louis-en-l'Île was begun the same year, 1664, but one of the finest houses, by Le Vau (who designed the Academy buildings), was completed in 1640. Another, the Hôtel de Lauzun, a few yards upstream on the Quai d'Anjou, was completed in 1657 and is used by the municipality to receive illustrious visitors. A great number of the houses on the island are originals. The Pont Marie to the Right Bank, completed as part of the contract, is—though modified for modern traffic—the original span. In 1740 it was the first of the bridges to lose the houses and shops along its back.

The Île Saint-Louis is one part of the mythical Paris that still exists in reality. Tranquil and lovely, it has matchless views of Notre-Dame, maintains its intimacy with the Seine, and constitutes a genuine neighbourhood with a disproportionate number of writers, artists, actors, and musicians among the residents.

The Bastille. The road off the upper end of the Île Saint-Louis leads to the Place de la Bastille on the Right Bank. From the river to the *place* runs a canal, the Bassin de l'Arsenal, which brought water to the moat around the Bastille. Under the Place de la Bastille the waterway continues underground for almost a mile, then emerges to form the Canal Saint-Martin, which, with its bridges and locks, and its barges sailing slowly down the centre of city streets, makes one of the least known and most picturesque sections of Paris. In 1972 traffic engineers were threatening to drain the water and make the canal into a highway.

The capture of the Bastille on July 14, 1789 (since annually celebrated), was a symbolic blow at tyranny rather than an act of liberation for tyranny's victims. The prison, virtually unused for years and scheduled for demolition by the monarchy, held on that July 14 but four counterfeiters, two madmen, and a young aristocrat who had displeased his father.

Napoleon had the *place* laid out in 1803. It was to be one of the stations on a triumphal way from the Place de la Nation (which lies to the east) to the Arch of Triumph at the Étoile. The neighbourhood between Bastille and Nation, the rue du Faubourg Saint-Antoine, has been one of skilled craftsmen since the mid-15th century, when the self-governing royal abbey gave space within its wide domains to those cabinetmakers who refused to abide by the restrictions of Paris guilds as to styles and types of wood to be used. The Faubourg was always among the first to revolt when revolution was in the air and was noted for the speed with which it raised barricades of impressive height. In the 1970s the Faubourg's character was changing because the small workshop, like the small grocery shop, was being squeezed to death by economic pressures.

The Marais. To the west of the Bastille lies a triangle of terrain which has its base along the river up to City Hall (Hôtel de Ville) and its apex just short of the Place de la République to the north. It keeps its name—*le marais* ("marsh")—from the Middle Ages, and because it became the market garden of Paris, it gave its name to all market gardens in France (*culture maraîchère*). Extension of the city walls along the Right Bank led to diking of the shore and drainage of the soil. In 1107 the Knights Templars established their Paris Temple, a vast fortified enclosure, at the top of the triangle. In 1360, Charles V moved into his new royal residence in the lower right-hand corner, where the rue de Lions marks the former location of the menageries.

Charles VII preferred to live just behind the Bastille, in the Hôtel des Tournelles, which Henry II had had enlarged and beautified by Philibert Delorme in 1550. Great nobles, such as the dukes of Guise and Lorraine, followed the King and had palaces built in the vicinity. When Henry II was killed in a joust on the rue Saint-Antoine in 1559, his widow, Catherine de Médicis, had the Tournelles razed. On the site, in 1607, construction began on the first residential square to be designed in Paris (others had arisen accidentally at wide crossroads). Henry IV had it built "to fill in the blanks" within the city's walls and reserved a house for himself among those he commanded "builded of a same cimettry for the decoration of the towne." They are three stories high, red brick with white-stone quoins (solid-corner angles) and window surrounds, the ground floors forming arcades over the sidewalks, still standing in full state and grace. Completed in 1612, the square was named la Place Royale, known to the thronging public simply as La Place. Another wave of building by the rich, anxious to be close to a royal project, endowed the Marais with 200 more private palaces. In 1800 the name was changed to Place des Vosges, which it is still called.

The first residential square

In 1792 the Knights Hospitallers of St. John of Jerusalem were turned out of the Temple that had been given to them in 1313 when the Templar order was dissolved. On August 13, 1792, the Royal family was incarcerated in the keep, the King was taken off to his death January 21, 1793, and the Queen removed to the Conciergerie August 2 (execution October 16). Two years later the death of the Dauphin—from maltreatment and malnutrition—was announced, but there are those who still believe that the 10-year-old child buried at the Church of Sainte-Marguerite, on the rue Saint-Bernard, was a substitute for the heir to the throne.

The Temple keep, a possible rallying point for royalists, was levelled in 1808, and in 1857 Haussmann installed on the site the *mairie* (town hall) of the third *arrondissement* and the Square du Temple.

After the 17th-century building boom the Marais remained virtually untouched, and three-quarters of the buildings there date from before 1870. Toward the end of the 19th century, while some of the oldest and most imposing of the palaces were being knocked down by private entrepreneurs, other private owners managed to restore a few mansions, and the nation and the municipality also restored a handful of fine buildings. The district became a dumping ground for Jewish refugees from eastern Europe, and to profit from the poverty-stricken newcomers scores of fine houses were subdivided into tiny apartments, with workshops on the lower floors and in courtyard sheds. In 1970 there were still 7,000 shops and workshops in the district, with 40,000 employees, although one of the prime industries, textiles and clothing, was rapidly diminishing. Through centuries of neglect the Marais became one of the worst slums in Paris: in the early 1970s there was no running water in 30 percent of the dwellings, 10 percent had no electric light, and 60 percent were without toilets. Cleaving to the pattern of bygone ages, buildings cover more than 85 percent of the area, compared with a city-wide average of 55 percent land occupation. Despite a population decline of 19 percent between 1962 and 1968, it remains the most densely inhabited part of Paris, some streets attaining a density of 5,800 inhabitants per acre (2,000 per hectare), the Paris average being 240 per acre.

Degeneration of the Marais into a slum

In 1969 the Municipal Council approved the creation

of a plan for the restoration of the locality. Not a "museum" plan, it is an urban renewal scheme for ending slum conditions while preserving the workaday life and animation and restoring the undeniable beauty of the quarter.

Among the restored ancient buildings open to the public are: the Museum of the History of Paris (Musée Carnavalet) 1545, enlarged by François Mansart 1645; the Museum of the History of France (National Archives, Hôtel de Soubise) parts from 1375 and 1553, main portions 1704–15; the Museum of Nature and the Hunt (Fondation Somer, Hôtel Guénégaud) by François Mansart 1648–51; and the Caisse Nationale des Monuments Historiques (Hôtel de Sully) by Androuet du Cerceau (Jean I). There are currently at least a dozen more.

Closer to the Hôtel de Ville is the Gothic Hôtel de Sens, built at the end of the 15th century for the bishops of Sens then also bishops of Paris. Restored after 40 years' work, it served by the 1970s as a city library of specialized collections. Nearby, behind facades of a much later date, two half-timbered medieval houses have been uncovered. Here and there through the Marais portions of the 13th-century city wall, including one of the watch towers, are still to be seen.

Hôtel de Ville. Three city halls have stood on this site, the last two grander, "improved versions" of their predecessor. The present building (1874–82) replaces the Renaissance structure in use from the 16th century until it was burned by the insurrectionary Communards in 1871. The first was the Maison aux Piliers (House with Pillars) used by the municipality from 1357 to 1533.

In the building are the official apartments of the *préfet de la Ville-de-Paris*, representing the national authority which has ruled the capital—with few brief and usually bloody exceptions—since the first kings appointed the counts of Paris as administrators of the city.

The three Paris-based prefects and diverse departments of the national government hold almost all the administrative and financial power. The municipal council serves the voters by airing public problems in debate and by obstructing or rejecting proposals they deem inimical to the city. Frenchmen at times complain that they are "the most governed people in the world"; Parisians, deprived of municipal self-rule, complain that they are the most governed of Frenchmen.

The municipal framework

The beginnings of the municipality were promising. In 1141 the crown sold the principal port (by the site of the Hôtel de Ville) to the river merchants guild, la Marchandise de l'Eau, whose ship-blazoned arms eventually were adopted as those of Paris. In 1171 Louis VII gave La Marchandise a charter confirming their "ancient right" to a monopoly of river trade. In 1190, when Philippe-Auguste went off crusading for a year, he entrusted the administration of the city, not to his Provost, but to the guild. In 1220, the crown ceded one of its own precious rights to the townsmen: the right to collect duty on incoming goods. The merchants were also made responsible for maintaining fair weights and measures. St. Louis recognized the Parloir aux Bourgeois as the municipal body with the Prévôt des Marchands heading an elected council of *échevins* (municipal magistrates).

The Provost of the Merchants in 1356 was Étienne Marcel, who wanted a Paris as rich and free as the independent cities of the Low Countries. It was he who gave the Maison aux Piliers to the municipal government, and it was he who slew the Dauphin's counsellors in the Palace throne room and took over the city. He showed great executive skill and equally great political stupidity. He allied himself with the revolting peasants (the Jacquerie), with the invading English, and with Charles the Bad, ambitious King of Navarre. While going to open the city gates to the Navarese in 1358, Marcel was slain by the citizens.

In 1382 a tax riot grew into a revolt called the "Maillotin uprising" because the rioters armed themselves with *maillets*, mauls. They were ruthlessly put down and the municipal function was suspended for the next 79 years. It was not until 1533, when Francis I ordered the teetering Maison aux Piliers replaced by a new building, that a monarch manifested an encouraging interest in municipal government.

The massacre of thousands of Protestants, on St. Bartholemew's Eve, 1572; the Day of the Barricades, May 12, 1588, when the Catholic League drove Henry III out of Paris; and the Fronde, 1648–53, all involved the Paris masses, but were palace revolutions rather than popular uprisings. When the people rose in their own cause, the focus of their action was inevitably the Hôtel de Ville, as was indeed the case in 1789, 1830, 1848, 1871, and 1944.

In July 1789, for example, after killing the Provost, the revolutionary mob took the Hôtel de Ville as it had already taken the Invalides and the Bastille. Three days later Louis XVI appeared on the balcony, wearing the tricolour cockade (royal white added to municipal red and blue), and was cheered by the crowd. Four years later the building was taken as headquarters for the fanatically extremist Commune, which directed mob action to control the Convention. On 9 Thermidor, year II (July 27, 1794), the Convention's guards entered the Hôtel de Ville and seized Robespierre and his Communards; all were executed soon after.

The Hôtel de Ville as a revolutionary centre

In August 1830, Louis-Philippe appeared on the balcony accompanied, as had been Louis XVI, by the Marquis de La Fayette, and he, too, was acclaimed by the revolutionary crowd. The name of the square was changed in that year to Place de l'Hôtel de Ville. As the Place de Grève (strand, or bank) it had been the principal port of Paris for centuries (the refusal of boatmen to put out gave the French worker the phrase for going on strike *faire la grève*). From 1310 to 1832 it was Paris' principal place of execution.

In 1871, after Napoleon III's defeat at Sedan, a new republic was declared from the Hôtel de Ville steps, but when the national government, in its turn, capitulated, Parisians refused to accept defeat and in March formed the Commune of Paris. In May, troops of the government entered the town and fought sharp engagements with the Communards, who set fire to the Hôtel de Ville, the Tuileries Palace, the Palais de Justice, the Préfecture de Police, the Arsenal, and other government buildings. The Communards executed 67 hostages, including the archbishop of Paris, and then went on to die after combat among the tombs of Père-Lachaise Cemetery.

In a bloodbath unparalleled in the history of France, the government subsequently rounded up and executed 20,000 Parisians, shipping thousands of others to penal colonies.

Finally, in 1944, in the bloody turmoil of World War II the National Council of Resistance made the Hôtel de Ville its headquarters during the Liberation combat. General Charles de Gaulle appeared on the balcony and was acclaimed by the crowd.

Les Halles. On the rue de Rivoli westward from the Hôtel de Ville is the Tour Saint-Jacques, the bell tower of a 1550 flamboyant Gothic church, the body of which was removed under the Directoire. A few streets further northwest toward the Louvre is the *quartier* of Les Halles, from 1183 to 1969 the central market of Paris. An incredible monument to human appetites, Les Halles burst with an infinity of colours, sounds, and odours. Its almost permanent traffic jam girdled some of the happiest *bistros*, some of the busiest streetwalkers, and some of the biggest fortunes in Paris. It was traditional for merrymakers to come down during the night to have onion soup, snails, or pig's feet and to depart at dawn with armloads of blossoms bought at the wholesale flower market. When the markets moved out to a new location, the picturesque, bawling, brawling district became a silent, empty, waiting stage.

The markets of Les Halles

To remake a unified space in the antique heart of Paris in any way that Paris wished was an opportunity that might not come again for another 100 years. After three years of public hearings and unpublic disputes, a general plan was agreed to by the municipality and Cabinet. Not one of the ten market pavilions—eight originals designed in 1866 by Victor Baltard, and two 1936 imitations—will remain here for posterity, even though the iron-and-glass "umbrellas" were admired by leading 20th-century arch-

itects and by simple Paris sentimentalists. While the meat market still waited for its new quarters in the scandal-stained La Villette slaughterhouses, the pavilions remained standing, the emptied ones used for exhibitions, concerts, and the like. The Church of Saint-Eustache (1532–1637) remains, as does the circular Halle au Blé (Grain Exchange) 1811–13, burned by the Commune and restored in the 1880s as the Bourse de Commerce (Commercial Exchange). Many of the old houses—a few are very old, with the majority somewhat elderly—will be renovated or restored to keep some of the Vieux Paris flavour. The central 25 acres (ten hectares) and an adjacent 12-acre (five-hectare) plot are being transformed. Around the edges will appear new structures, most of them low-built: the new Ministry of Education, an International Centre of Commerce with exhibition halls, a museum of contemporary art, a public library, an auction sale room, and several apartment houses and hotels. The empty space they border is simply a concrete skin over a huge underground installation. Its surface, something like a dried lake bed, will be varied by terraces of different levels, small gardens, some thin stands of trees, and light wells. Beneath, on several levels, a subterranean city will emerge. At the bottom will be the tracks of the east–west express subway (RER) and the extended north–south suburban Sceaux railway line; above that, the rail stations and waiting rooms; above that, on several floors around a central garden, the "Forum" through which 80,000 commuters a day will stream, containing movie theatres, sports installations, eating places, bars, and shops. Just under the surface will be the regular métro with its usual Les Halles station, a small cab and bus station, and some subterranean roads. There will be underground parking.

To the west of Les Halles is the Banque de France, which also has an important underground installation (the French national gold reserve), and, to the north of the bank, the Stock Exchange (Bourse). The bank is in close proximity to most of the banking and insurance headquarters offices in Paris. The city is the location of only 27 percent of all such headquarters in France, but those in Paris count 95 percent of all those with more than 1,000 employees, lending fuel to the provincial objection that "everything is run from Paris." What is not run from the city by the central government has a good chance of being run from there by some giant corporation. The last available statistics show that over 62 percent of all French corporate tax was paid by Paris-based companies. The region produced close to a third of the national product in services (exclusive of civil service) and the same amount of industrial product.

The city's financial section

Table 4: Paris, Economic Data

	persons employed	
	(000)	percent
Agriculture	1.9	0.1
Mining	1.5	0.1
Manufacturing	450.2	30.0
Construction	83.8	5.6
Electricity, sanitation	11.0	0.7
Commerce	365.3	24.4
Transport	78.0	5.2
Services (including administration)	483.8	32.3
Others and unknown	23.2	1.6
Total	1,498.7	100.0

Source: *Annuaire de statistique internationale des grandes villes*, 1968.

MODERN DEVELOPMENTS

Rue de Rivoli. The Louvre and the Tuileries Garden take up the south side of this street, and on the other side runs an arcade more than a mile long. Napoleon I opened up the street from the Place de la Concorde, Charles X continued it, as did Louis-Philippe, and Napoleon III carried it on down into the Marais.

Opposite the middle of the Louvre, the Place du Palais-Royal leads to the palace of Cardinal de Richelieu, built in 1624 and willed to the royal family. Louis XIV lived there as a child, and during the minority of Louis XV the kingdom was ruled from there by the debauched but gifted regent. Late in the 18th century Louis-Philippe d'Orléans, who became Philippe-Egalité after the Revolution, undertook extensive building around the palace garden. It was a commercial operation, and the prince hoped to pay his debts from the property rents. ("Well cousin," said Louis XVI, "so you're going to keep shop; we'll never get to see you except on Sunday.") Around the garden he built a beautiful oblong of colonnaded galleries, and at each end of the gallery farthest from his residence, a theatre. The larger playhouse has been the home of the Comédie-Française, the state theatre company, since Napoleon's reign. The princely apartments now shelter high state bodies such as the Conseil d'État.

The prince's financial success was modest, but the social impact was sensational. From the 1780s to 1837 the Palais-Royal was the local synonym for excitement. It was the centre of Parisian political and amorous intrigue and the site of the most celebrated gambling dens and popular cafes. Today the garden and its galleries are still beautiful but are wistfully deliquescent, a Pompeii where even the tourists are rare.

Just behind the garden is the Bibliothèque Nationale, the national library of deposit, with the expected enormous collections of books and prints, some 6,000,000 of each.

When Haussmann greatly enlarged the Place du Palais-Royal in 1852, he did not molest the palace when he pushed through the Avenue de l'Opéra. At the top of the new street, where the Grands Boulevards crossed an enormous new *place*, the new Opera House was built, pulling pleasure seekers further away from the Palais garden. The Opéra (1825–98), the neo-Baroque masterwork of Charles Garnier, is a splendiferous monument to the Second Empire. By acreage it is the largest theatre in the world, but so much space is devoted to such embellishments as the Grand Staircase that in seating capacity it is not the largest theatre in Paris. Just behind the Opera House the large department stores indulge in the same kind of uninhibited monumentality, the sort of thing they now avoid in their branches at suburban shopping centres springing up around the country.

The Opéra

On the rue de Rivoli the next *place* is the Place des Pyramides. The gilded equestrian statue of Joan of Arc stands not far from where she was wounded (at the Saint-Honoré Gate) in her unsuccessful attack on British-held Paris, September 8, 1429.

Farther along toward the Place de la Concorde the rue de Castiglione leads to the Place de Vendôme, an elegant octagonal *place*, little changed from the 1698 designs of Jules Hardouin-Mansart. In the centre, the Trajanesque Vendôme Column, 44 metres high and spiralled in the bronze of 1,200 captured cannon, bears the effigy of Napoleon, who had it erected in 1810. It was pulled down during the Commune and put back up by the Third Republic. The *place* and the gas-lit rue de la Paix have lost none of their discreet distinction, nor have their shops. The rue de Rivoli shops, once equally chic, have in many cases acquired a disguised but unmistakably vulgar accent. The street's hotels maintain their traditional high quality. The German commander of Gross Paris, Dietrich von Choltitz, who disobeyed Hitler's order to burn the city, was captured in his headquarters at the Meurice Hotel August 25, 1944.

La Voie Triomphale. From the Arc de Triomphe du Carrousel in the courtyard between the open arms of the Louvre, there extends one of the most remarkable perspectives extant in any modern city. It is called—though not in everyday speech—the Triumphal Way. From the middle of the Carrousel Arch, the line of sight runs the length of the Tuileries Garden, lines up on the obelisk in the Place de la Concorde, and goes up the Champs-Élysées to the centre of the Arch of Triumph.

The Louvre's modest triumphal arch stands in the open space where costumed nobles performed in an equestrian display—*un carrousel*—to celebrate the Dauphin's birth in 1662. The design of the arch, an affable imitation of

that of the Arch of Septimius Severus (Rome), was confected in 1808 by Percier and Fontaine, who also perpetrated a great deal of the Empire style's sphinx-fraught furniture and decoration. Napoleon, a record of whose victories is incised on the arch's flanks, decorated the summit with the famed four bronze horses from St. Mark's in conquered Venice. The Venetians had taken them from conquered Constantinople, which had acquired them in Rome, which, in turn, had looted them from, it is believed, the Temple of the Sun in Corinth. After Napoleon's defeat, France was forced to return them—to Venice.

The Tuileries Garden

The Tuileries Garden, which fronted the Tuileries Palace (looted and burned in 1871 during the Commune), has not altered much since André Le Nôtre laid it out in 1664. Le Nôtre, who was born and who died right in the garden, in the gardener's cottage, carried the line of his central *allée* beyond the garden and out into the country by tracing a path straight along the wooded hill west of the palace. On this hilltop, 170 years later, the Arch of Triumph was erected.

At the western edge of the garden, Napoleon III erected a hothouse—the Orangerie—and a court for real (or royal or court) tennis—the Jeu de Paume. The former is used for temporary art shows, the latter houses the Louvre collection of paintings by the Impressionists and their forerunners. Among the artists represented are: Cézanne, Degas, Gauguin, Manet, Monet, Renoir, Rousseau (*le douanier*), Toulouse-Lautrec, van Gogh. From the terraces on which these museums stand there are splendid views of the olympian traffic jams of the Place de la Concorde.

The formal exit gate from the Tuileries is flanked by two winged horses (17th century), and the entrance to the Champs-Élysées across the square is similarly garnished (horses, earthbound, 18th century), both pairs having been removed in turn from the water trough at Chateau de Marly.

The Place de la Concorde was designed as a moat-skirted octagon in 1755 by Jacques Ange Gabriel. He had won competition set by the *échevins* of Paris for a king-flattering "Place Louis XV." The river end was left open, and on the inland side, two matching buildings were planned. The ground floor was arcaded and the facade was nimbly adapted from the Louvre colonnade, all with a refinement typical of the era. Although Gabriel built eight giant pedestals around the periphery of his *place*, they remained untenanted until Louis-Philippe gave them statues representing provincial capitals going clockwise from the Navy Ministry (Ministère de la Marine): Lille, Strasbourg, Lyons, Marseilles, Bordeaux, Nantes, Brest, and Rouen. Louis-Philippe also had the Luxor Obelisk, a gift from Egypt, installed in the centre and flanked by two fountains. Later, the surrounding moat was filled in, and the Place de la Concorde took on its present geography.

Louis XVI was decapitated near the statue of Brest, January 21, 1793. The guillotine was brought back to the *place* four months later and erected near the gates of the Tuileries, and the executions went on for nearly three years, with a total of 1,343 deaths on this spot. An approximately equal number of persons perished in other parts of town.

La Madeleine

Between the twin buildings the broad rue Royale mounts to the Church of La Madeleine, a stern oblong, fenced with columns 20 metres high, ostensibly a Greek temple but closer to the Roman notion of Greek. From 1764 to 1806 three false starts were made on building a church on this site. Napoleon ordered Vignon to build a temple to the glory of the Grande Armée but then vanished from the scene, and La Madeleine was not consecrated until 1842.

The Place de la Madeleine is the western terminus of the Grands Boulevards, which imitate the arch of the river from there north and east to the Place de la République and the Bastille. The glittering chic of the Grands Boulevards, which flavoured Paris life from 1750 to the 1880s, is gone with the *boulevardiers*, gone with Café Tortoni, the Café Riche, the Maison Dorée. The state-run Opéra-Comique persists just off the boulevard des Italiens, the wax museum is still there on the boulevard Montmartre, and, a few doors away, Offenbach's theatre, the Théâtre des Variétés, still operates. The Théâtre de la Renaissance, where Coquelin created Cyrano de Bergerac in 1897, still functions on the boulevard Saint-Martin, as does the Théâtre de l'Ambigu, where Frédéric Lemaître, idol of boulevard melodrama, thrilled all Paris in the mid-19th century. Some of the movie palaces of the 1930s are still serving along the boulevards as well.

West off the rue Royale runs the rue du Faubourg Saint-Honoré, which, in addition to the British Embassy and the Élysée Palace (residence of the French president), has on its shop windows some of the most prestigious names in Paris retail trade and has as its window shoppers some of the most richly dressed and—at times—most beautiful women in the world.

Along the first 2,500 feet (800 metres) of the Champs-Élysées, between Concorde and the Rond-Point des Champs-Élysées, little has changed since the childhood of Marcel Proust. The 230-foot-wide avenue is bordered with chestnut trees, behind which on both sides are gardens, usually full of children, often with their nannies. Rides on donkeyback and in goat-drawn carts are still offered. Punch still punches Judy. The discreet pavilions in the gardens are tearooms, restaurants, and theatres. Along the paths it is still possible to get one's shoes dusty on the way to the office.

The Champs-Élysées today

From the Rond Point up to the Arch of Triumph just about everything has changed, and continues to change, along the Champs-Élysées. Under the Second Empire this was a street of luxurious town houses, of which but two were left in the early 1970s. Then came the cafés, night-clubs, luxury shops, and the cinemas, but the street retained its feeling of luxury and the tree-shaded sidewalks (wide as a normal street) offered promenades that were the pride of Paris. Since the 1950s, however, more buildings designed as corporate headquarters have evicted shops, tearooms, and hairdressers alike. Automobiles park on the sidewalks and in many places one rank of the double file of trees has been felled. Pedestrians seem to mill rather than stroll. In the 1970s this street, once one of the most justly famous in the world, was becoming another in the international series of "downtown prestige addresses" from which all that bestowed the prestige has been extirpated.

At the top of the Champs-Élysées is a circle 450 feet in diameter from which 12 imposing avenues radiate to form a star (*étoile*). From 1753 to 1970, it was called Place de l'Étoile, then was re-named Place Charles de Gaulle. In the centre of the circle is the Arch of Triumph, commissioned by Napoleon in 1806. After Napoleon's fall it stood unfinished until Louis-Philippe saw to its completion in 1832–36. At 50 metres, it is twice as high as the Arch of Constantine, which inspired it, and, at 45 metres, a little more than twice as wide. Jean Chalgrin was the architect and François Rude sculpted the frieze and the spirited group, "La Marseillaise" (real title, "The Departure of 1792"). On Armistice Day in 1920, the Unknown Soldier was buried under the centre of the arch, and each evening the flame of remembrance is rekindled by a different patriotic group.

La Défense

The westward thrust of the city was dramatically demonstrated in the 1970s when the biggest concentration of tall buildings in the nation arose over two miles beyond the Arch, on the far side of the rich little suburban wedge of Neuilly. The site, called Quartier de la Défense, was formerly just a place in the road where the depressed suburban municipalities of Puteaux, Courbevoie, and Nanterre listlessly touched. Now, 30 office towers, 30 stories tall, heated and air conditioned from a central plant, are the hub of a complex, balanced city plan. The "ground level" between buildings is a raised platform reserved to pedestrians, with roads and parking below. There will be shops, restaurants, cafés, a shopping centre, hotels, and apartment houses. Before the project was begun, the state had already constructed at La Défense its Centre National des Industries et des Techniques, an exhibition hall with 90,000 square metres of floor space. Nanterre became the site of a campus of the

University of Paris in the 1960s, and in the 1970s specialized schools of university level were installed in the new centre: a National School of Architecture, the National School of Decorative Arts, the National Conservatory of Music, and the Institute of Advanced Cinematographic Studies. The three municipalities benefit directly by acquiring low-built public housing in park settings, a large park, day child-care centres, and new schools.

New business quarters. As a counterbalance to the westward march of office buildings and as part of the effort to limit the obliteration of residential quarters around the business centre, five "poles of attraction" were instituted in other parts of Paris in the late 1960s and early 1970s. Two of these are directly on the waterfront at each end of Paris, Front de Seine in the southwest corner, and Austerlitz-Bercy-Lyon in the southeast corner.

The Front de Seine covers 70 acres (29 hectares) on the Left Bank, between the Eiffel Tower and the southern city limits. A neighbourhood of factories and substandard housing is replaced by a spread of 16 towers of 32 stories and four buildings 15 stories high. One-quarter of the space is used for offices, the rest for apartments, some of which are low-cost housing.

The project at the opposite end of the river straddles the Seine, putting office buildings around the Gare d'Austerlitz (Left Bank) and the Gare de Lyon (Right Bank). Bercy, which lies directly on the river (Right Bank) between the two stations, was until this development one of the secret cities of Paris. Fenced and guarded, its chalets lined cobbled lanes named for the great vineyard districts of France. The great oaks, it was said, flourished because their roots were soaked in wine. This was the village of vintages, where merchants stored and sold their stocks of liquid treasure. Now the dealers, abandoning folklore for modern conveniences, have retreated into a modern block on the city line at the river. The rest of their domain is occupied by a 20-acre (eight-hectare) park, around which are 14 buildings, which range from 16 to 40 stories in height. Half the 6,000 apartments are low-rent public housing, and the rest of the tower space is for offices. Schools and shops complete the project.

Bercy

At the Place d'Italie, 5,000 feet south of the Gare d'Austerlitz, there is developing what the planners call "an urban strong point" midway between the city centre and Orly airport. The project was begun in 1969 and should be terminated in 1989. On 215 acres (87 hectares) there will finally be 14,000 new apartments, 200,000 square metres of office space, and 150,000 square metres of retail shops.

The run-down areas around the Gare de l'Est and the Gare du Nord (the East and North railway stations), high on the Right Bank, are also being developed as another commercial complex.

The project that literally overshadows the others is Maine-Montparnasse, where the centre piece of the composition is a 56-story (200-metre-high) office tower. Planned for completion in 1973, it was designed as the tallest in Europe. It stands on the site of the old (1850) Montparnasse railway station. A more compact station has been built one street away on the avenue du Maine, where the rails are hidden on three sides by three buildings 15 to 18 stories high, which include 30 floors of offices and 1,000 apartments. The units are joined by a raised platform that serves as "ground level" above the street. Excluding employees of the retail shops, there is room for 21,000 jobholders in the project.

The effect on the surrounding neighbourhood of Montparnasse is prodigious. Always fairly seedy, it was nevertheless dear to many as the playground, classroom, and workshop of the 1920–35 "lost generation." The population of painters and writers declined after World War II, but it has been further reduced by the effects of real-estate euphoria provoked by the big tower. Some favourite old *beehives of Bohemia* have been replaced by new speculative buildings, and the remaining studios have property values now as absurdly high as most artistic ambitions.

Montparnasse

Factories and workshops disappear from Paris at an even faster rate than the vanishing artist's *atelier*. Permits for expansion or even rebuilding such premises are rare, taxes are purposely inhibitory, and concessions and premiums for relocating in the new regional industrial zones are attractive. During debate on the city plan, some municipal councilmen expressed apprehension that eventually there would be no more "blue collar" workers in Paris, a state of affairs that would change Paris far more profoundly than any number of skyscrapers.

The ministry quarter. The government, both local and national, is a gargantuan occupier of Paris office space and at an accelerating pace: in 1970 there were four times as many public agents per inhabitant as there were ten years earlier. One civil servant out of every six in France works in Paris. The seventh *arrondissement*, a delta-shaped district the base of which runs along the river from the Eiffel Tower past the Esplanade des Invalides to the Pont du Carrousel, is known as the *quartier des ministères*. Ten of the 15 ministries are located there along with the prefecture of the Paris Region and the National Assembly.

Most of the territory of the seventh *arrondissement* is government property. This is the old *Faubourg Saint-Germain*, an impeccable address since the early 18th century. As such, it was subject to heavy expropriation during the Revolution, and ministries are lodged mostly in splendid old mansions and convents. Properly imposing, these are never large enough and always difficult to adapt to modern administration. When it proves impractical to spread into adjacent buildings or to construct quite dreadful annexes in the garden, branches are installed wherever space can be found—in this case a total of 115 different locations. The Ministry of Finance has 25 different buildings; Social Affairs has 25; and the Prime Minister, who lives and works in the Hôtel de Matignon (1722) on the rue de Varenne, also uses 18 other scattered buildings.

Probably the best known of all ministries is the low-built, ornate Ministry of Foreign Affairs on the Quai d'Orsay between the Esplanade des Invalides and the National Assembly. The address "Quai d'Orsay" has become a synonym for the ministry.

The Quai d'Orsay

The National Assembly is housed in the 1722–28 Palais-Bourbon, seized during the Revolution. Succeeding regimes have tacked bits and pieces onto the old palace, including the Greek peristyle facing the river as ordered in 1807 by Napoleon.

The Pont de la Concorde, just below the gates of the Palais-Bourbon, was built 1787–91, partly with stones from the demolished Bastille.

Quartier Latin. At the Pont de la Concorde the Boulevard Saint-Germain begins, curving two miles (three and one-fourth kilometres) eastward to find the river again at the Pont de Sully. A little less than halfway along the boulevard is the pre-Gothic Church of Saint-Germain-des-Prés. Belonging to a Benedictine abbey founded in the 8th century, it was sacked four times by the Vikings, and rebuilt between 990 and 1201, parts of the present church dating from that time.

The area has long been a meeting place for artists. Racine died there in 1699; Delacroix had his studio in the still-winsome little Place Fürstenberg; publishing houses moved in during the 19th century; and the principal cafés have been meeting places for artists, writers, and publishers since. From 1945 to 1955 it was the hub of the "Existentialist" movement, which stemmed from Jean-Paul Sartre's adaptation of Kierkegaardian philosophy but which is more generally remembered as the excuse for a revived Bohemia and a renewed expatriate flood into Paris. The first *discothèque*—an institution which gained international currency in the 1960s—was founded here. In the 1970s it was still a very lively centre for literature, food, and talk.

Straight north from the Saint-Germain-des-Prés crossroads is the École des Beaux-Arts on the Quai Malaquais, the state school of painting and sculpture. For many generations the students' annual *Bal des Quatz'Arts*, usually concluding with a nude dawn dip in the Concorde fountains, was one of the memorable wild youthful moments of the Paris calendar. In recent years it has gained

as little attention as the works of art produced during the academic year.

Two streets south of the crossroads is the Church of Saint-Sulpice (1646–1780), the work of six successive architects who, in 134 years, scarcely had a good day among them. The street alongside the church is sprinkled with shops specializing in devotional statuary, much of it on the aesthetic level of seaside souvenirs and known in France as "Saint Sulpicerie."

Eastward to the boulevard Saint-Michel, the area toward the river from boulevard Saint-Germain is a tangle of narrow, animated streets, "typically Parisian" in a noisy, happy way that now, alas, typifies so little of Paris.

East of the boulevard Saint-Michel is the university precinct, self-governing under the kings, where, in class and out, students and teachers spoke Latin until 1789. On the corner of the boulevards Saint-Germain and Saint-Michel are the remains of one of the three baths of the Roman city. These are in the grounds of the Cluny Museum, a Gothic mansion built 1485–1500, which now houses a collection of 20,000 objects and works of art of the Middle Ages, including the renowned six-panel unicorn tapestry "La Dame à la Licorne."

The wide, straight boulevard Saint-Michel (no longer the Boul Mich in the student lexicon) is the main street of the student quarter. The university buildings are curiously little visible to the casual visitor who sees a mile of bookshops, cafés, cafeterias, and cheap movie houses.

La Sorbonne

The university was built up of colleges, each founded and supported by a donor, often a prelate or a religious order. Robert de Sorbon, chaplain to St. Louis, established a college in 1235 that eventually became the centre of administration for the university to which it gave its name, La Sorbonne. The liberal arts faculty off the Place Sorbonne, which is the administrative seat for the whole University of Paris, is housed in large, anonymous 1900 buildings tucked away behind small streets. Some faculties, schools, and institutes have been moved out of Paris, some to the new towns, and others are dispersed at some distance. The Science faculty has done both, with new buildings (1968) in the town of Orsay and also on the site of the old Halle aux Vins, still in the fifth *arrondissement* but down on the river next to the Jardin des Plantes, a 1626 botanical station to which are attached the Museum of Natural History and a small, charming, old fashioned *ménagerie*.

This decentralization, plus the creation of new "campuses" (the concept has been so completely forgotten in France that the American word is used) in the provinces as well as in the Île de France, should eventually end classroom overcrowding and the student housing shortage in Paris. Application of reforms promulgated in the 1960s should provide curricula as modern as the new buildings.

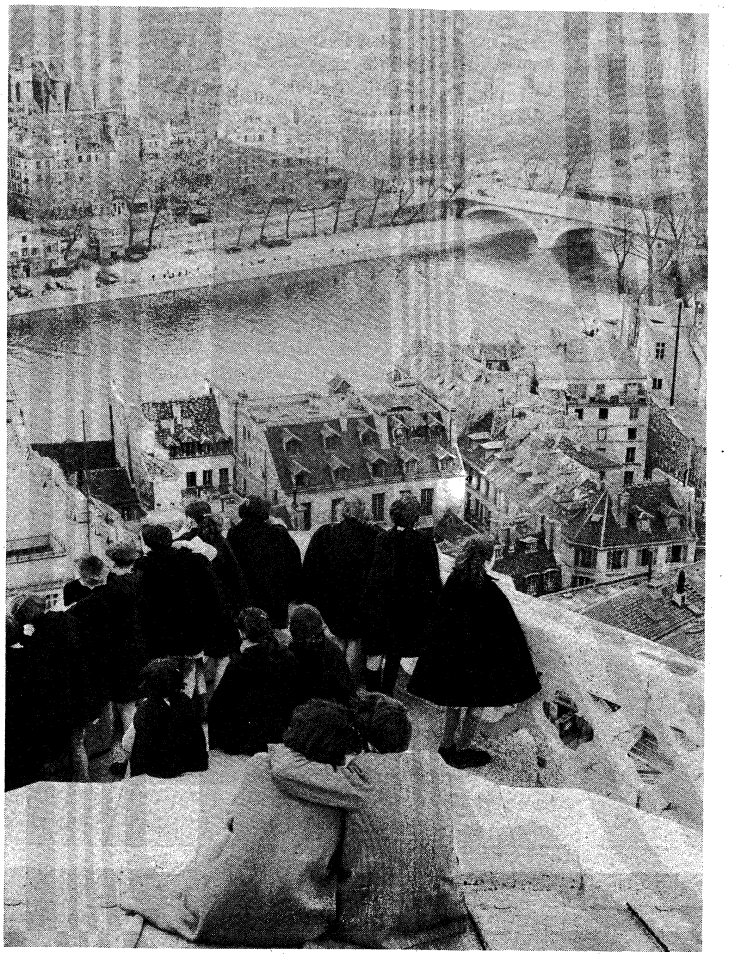
The oldest Sorbonne building is the chapel (1635–42), the gift of Cardinal de Richelieu, who is buried there. Designed by Lemercier, it was another of the new domed "Jesuit" churches of the period.

The independent Collège de France was set up a few steps from the University by Francis I in 1529 to offer a more liberal, modern curriculum than the narrow theology and limited Latin of the Sorbonne. Bestowing no degrees, it has always had a superb faculty of famous specialists, especially in philosophy, literature, and the sciences.

At the top of the hill rising from the river the boulevard skirts the Luxembourg Garden, the remains of the park of Marie de' Medici's Palace (1616–21), which now houses the French Senate. The chestnut-planted garden, with its pond for toy sailboats, its marionette theatre, and its ridiculous statuary, has a curious power of persuading many visitors that they are 19th-century persons of quality.

The Panthéon

Across the boulevard at the end of the rue Soufflot stands the Panthéon, 1755–92, by Jacques-Germain Soufflot. The colonnaded facade, the great dome, the barrel-vaulted interior make for power and dignity. Cured of an illness, Louis XV built the church as a votive offering to Ste. Geneviève, in replacement of the moldering 5th-century abbey named for her. When the church was finished, the Revolutionary authorities decreed it to be the Panthéon, final resting place for heroes of the Revolution. Napoleon restored it to the church, which lost it to the next revolution, and so on until the 1885 burial of Victor Hugo, after which it remained the Panthéon, but with the cross still atop the dome. Voltaire, Jean-Jacques Rousseau, La Fayette, Hugo, Zola, Felix Éboué and the 1943 Resistance leader Jean Moulin are among those buried under the inscription, "Aux grands hommes, la Patrie reconnaissante" ("To great men, [from] the grateful Motherland").



The Seine River and its embankments.
Henri Cartier-Bresson—Magnum

Behind the Panthéon is a steep street named La Montagne Saint-Geneviève. It was the Caesars' paved road to Italy. The hill leads down to the lively market square of Place Maubert and a tangle of ancient, picturesque, riverside streets. The best known of these is the medieval rue de la Huchette, from which the Street of the Fishing Cat (rue du Chat-qui-Pêche)—a blank, sweaty-walled airshaft—leads to the Quai Saint-Michel. Two Gothic churches, modest Saint-Severin (1489–94) and humble Saint-Julien-le-Pauvre (1165–1220), nest in this urban bramble patch. Saint-Julien-le-Pauvre was cut in half when collapsing portions were removed in 1651. The Sorbonne's annual assembly met in the church for 250 years, until 1524, when the meetings were forbidden because of student rowdiness. The square in front of the church offers one of the finest views of Notre-Dame de Paris.

The Buttes. The river valley of Paris is almost entirely circled by high ground. Upon the heights of Passy, on the Right Bank between the southern and western city limits and the Arch of Triumph, perch the *beaux quartiers* of the 16th *arrondissement*. The population contains a preponderance of persons with above-average incomes and, in many cases, of above-average pretensions.

The old-shoe amiability of Montparnasse (14th and a little of the 15th *arrondissement*) has been tainted by an injection of modishness provoked by the presence of the skyscraper and its attendant structures.

The Butte-Montmartre (18th *arrondissement*) and the Buttes-Chaumont (19th *arrondissement*), which rise along the northern rim of the city, are still unaffectedly "popular" in the sense of "populi." The oblong 18th *arrondissement* may perhaps be characterized as an urban pudding full of ill-assorted lumps. The range runs from the beautiful to the quaint, through the tawdry to the sordid. It has many seductive village corners and even more brutal city traps. It has broad avenues, but it also has winding lanes, some of which transform themselves into five-flight stairways. From the early 19th century until the 1920s' migration to Montparnasse, it was the great art colony of Paris. Some sections are mercilessly gewgawed for the tourist trade and others are unselfconsciously picturesque. It is the home of hundreds of thousands of hardworking, ordinary folk and one of the main centres of crime and prostitution. Montmartre itself is strewn with cheap little nightclubs, which often prove to be menacingly expensive.

Montmartre

The most noted landmark of Montmartre, one of the most indelibly Parisian, was inserted into the landscape only in 1919: the Basilica of the Sacré Coeur (Sacred Heart), more fully, of the National Vow to the Sacred Heart, paid for by national subscription after the French defeat by the Prussians in 1870. The work began in 1876 but was delayed by the death of the architect, Paul Abadie, who took inspiration from the 12th-century five-domed Romanesque Church of Saint-Front in Perigueux, itself inspired either by Venetian or Byzantine churches. Alongside the monumental terraced stairway of the garden-planted Square Willette below the church entrance runs the only funicular railway in Paris.

On the Buttes-Chaumont, just to the west of Montmartre, are the gasworks, the canal, the abbatoirs, and one of the most engaging parks in Paris. Called the Parc des Buttes-Chaumont, it was yet another Haussmann enterprise (1864–67). A bare hill, half hollowed out by abandoned tunnel quarries, stinking with the dumped refuse of generations, and the lair of army—especially Foreign Legion—deserters turned thug, it was turned into a romantic landscape with a lake, a waterfall, a grotto (false stalactites), winding woodland paths, and picturesque bridges. It is the largest park within the walls of Paris, 60 acres (25 hectares), and the only one to stay open as late as midnight.

This portion of the 19th *arrondissement* is known as Belleville, and just below on the slope toward the river is Ménilmontant, in the 20th *arrondissement*. Together, these two sections produce the Brooklynites or Cockneys of Paris, who have their own nasal, buzz-saw manner of speech and a hardheaded, warm-hearted approach to life. Part of their domain is Père-Lachaise Cemetery, site of the wall—le Mur des Fédérés—against which the last of the Communard fighters were shot, and to which pilgrimages are still made. Chopin, Marshal Ney, Baron Haussmann, Alfred de Musset, Balzac, Delacroix, Bizet, Rossini, Sarah Bernhardt, Isadora Duncan and Colette are among the noted persons buried here.

Green spaces. Paris appears to have much more green, open space than it really has. This optical-psychological illusion is fostered by the garden effect of the Seine's open waters and its tree-lined banks (2,300 of its own trees, in addition to nine major parks and gardens along its route) and by the parklike effect of 87,000 trees (mostly plane trees, only 10 percent chestnuts) planted along the streets. One earnest civil servant, wishing to demonstrate that the illusion was no illusion but a reality, showed in 1970 that—counting everything, including cemeteries and large private gardens—there is one tree for every 6.6 inhabitants of Paris, far above international minimum standards.

The city's abundance of trees

Every plan for improving Paris has called for more parks and playing fields, and every plan has fallen short of its goals.

The large parks and gardens of the central city were on the outskirts when first created by royalty, and the city grew around and beyond them. The hero of the green-space movement remains Napoleon III. During the years of his exile in England he had been impressed by London parks and resolved to give Paris the same joys. Two ancient royal military preserves at the approaches to Paris were made into "English" parks: the Bois de Boulogne, 2,380 acres (962 hectares) on the west, and the Bois de Vincennes, 2,460 acres (995 hectares) to the east. Also under the Second Empire, 173 acres (70 hectares) were given to promenades and to 24 garden-planted city squares, and the number of trees planted along the streets was almost doubled.

In the century from 1870 to 1970 a total of 252 acres (102 hectares) was gained for the population. From 1953 to 1967 successive plans provided for the addition of two dozen new squares to the "green patrimony" of the citizens, but financial and administrative difficulties proved too great for more than token progress. Administrative complexities are suggested by a short list of the institutional organisms responsible for greenery: public squares by the City Parks Department; street trees by the City Streets Department; riverbank trees by the Navigation Service; Tuileries and Palais Royal by the Cultural Affairs Ministry; Luxembourg by the Senate; and Jardin des Plantes by the Ministry of National Education.

Just outside the city three enormous recreational facilities have been established, with facilities for most sports, stadia, green fields, and woodlands. One of these, at Issy-les-Moulineaux, immediately beyond the city limits, was borrowed by the army in 1890 and returned to the city in 1970.

The number of public swimming pools in Paris intramurals was increased from 39 to 54 in the years 1962 to 1972.

SOME PARISIAN FACTS AND FIGURES

Cultural facilities. There are 60 theatres and 200 motion picture theatres, 15 concert halls, and 65 museums in Paris. The city operates 75 neighbourhood lending libraries and two specialized collections. The libraries of the Arsenal, Saint-Geneviève, and Bibliothèque Nationale are research libraries requiring special readers' cards.

Consumption. Paris consumes one-quarter of all the commercially prepared foodstuffs in France. More than 400,000 metric tons of meat and a thousand million litres of wine are sold in the capital every year. Every day the inhabitants and industries use 4,000,000 cubic metres of water. Treatment stations cleanse 1,200,000 cubic metres of water and 3,000,000 cubic metres of effluent each day. Each inhabitant throws away one kilogram of household refuse per day.

Paris underground. From Roman times to the end of the 19th century, Paris was tunnelled for the extraction of gypsum, limestone, and brick clay. Notre-Dame, Saint-Germain-des-Prés and Saint-Severin are among the churches built of local stone. The galleries, which extend under seven *arrondissements*, were filled in many years ago, save for the catacombs of Place Denfert-Rochereau and 90 miles (150 kilometres) of inspection galleries.

The 1,200 miles (2,000 kilometres) of sewer tunnels serve as host for drinking-water ducts 987 miles (1,588 kilometres), industrial and street-hydrant ducts 1,048 miles (1,686 kilometres) and also for the telephone, *télex*, telegraph circuits, the system of fire and police alarms, and the traffic signals. Also hung in the sewer conduits are the pipes of a system—unique in the world—for furnishing compressed air to homes and workrooms. Installed by an Austrian, Victor Popp, in 1881, the system was designed to furnish motive power for clocks and elevators. The compressed air still finds many applications, and three stations service 600 miles (1,000 kilometres) of tubes.

Electricity 6,000 miles (10,000 kilometres) of cable and gas 1,400 miles (2,300 kilometres) of pipe are both laid under the sidewalk. Some 89 miles (144 kilometres) of separate tubing deliver compressed gas to 2,000 street lamps and 1,800 homes.

Steam for central heating, furnished to 2,200 subscrib-

ers, has been circulated under the city streets since 1928. The steam comes from two rubbish incinerators, an electrical generating plant, and two steam generators. The subway uses 105 miles (169 kilometres) of tunnels.

BIBLIOGRAPHY

General description: P. COHEN-PORTHEIM, *The Spirit of Paris* (1937), provides a short and agreeable introduction that is supplemented rather than superseded by J. RUSSELL, *Paris* (1960). For a rather fuller treatment the English reader may consult: A.H. BRODRICK (ed.), *Paris* (1950) and *Greater Paris and the Île-de-France* (1952); G.R. MARTINEAU (ed.), *Paris: Its Environs* (1954); H.P. CLUNN, *The Face of Paris*, new ed. (1958); B. EHRLICH, *Paris on the Seine* (1962); or V. CRONIN, *The Companion Guide to Paris* (1963). There are also two very well-illustrated volumes with English versions of text by distinguished French writers: ANDRE GEORGE, *Paris* (1952; Eng. trans., new ed., 1957); and MARCEL BRION, *Paris* (1962; Eng. trans., *Paris in Color*, 1964). Famous writers whose personal presentations of Paris must be read in the original French include: LEON-PAUL FARGUE, *Le Piéton de Paris* (1939); JEAN COCTEAU, *Paris tel qu'on l'aime* (1950); HENRY DE MONTHERLANT, *Le Fichier parisien*, augmented edition, especially interesting on the life of the city (1955); PAUL MORAND, *Paris*, sumptuously illustrated with photographs (1970); and GEORGES SIMENON, *Le Paris de Simenon*, with sketches by FREDERICK FRANCK (1970).

History: Good general histories available in English are H. BIDOU, *Paris* (1937; Eng. trans., 1939); and R. LAFFONT (ed.), *Illustrated History of Paris and the Parisians* (1958). Major works in French include: M. POETE, *Une vie de cité: Paris de sa naissance à nos jours*, 3 vol. and album (1924–31); L. DUBECH and P. D'ESPEZEL, *Histoire de Paris*, 2 vol. (1931); and PHILIPPE LEFRANCOIS, *Paris à travers les siècles*, 10 vol. with photographs (1948–56). For the history of the town's growth, see particularly R. HERON DE VILLEFOSSE, *Construction de Paris* (1939); or the handier surveys by M. RAVAL, *Histoire de Paris*, 6th ed. (1953); and by P. LAVEDAN, *Histoire de Paris*, new ed. (1960). Studies of special periods range from P.M. DUVAL, *Paris antique: des origines au troisième siècle* (1961), to B. CHAMPIGNEULLE, *Paris de Napoléon à nos jours* (1969); and there is also a remarkably documented forecast of future development by P. MERLIN, *Vivre à Paris 1980* (1971).

Antiquities: The monumental work of F. DE ROCHEGUDE and M. DUMOLIN, *Guide pratique à travers le vieux Paris* (1923, rev. 1925), has been adapted by J.P. CLEBERT under the title *Les Rues de Paris*, 4 pts. (1958), which may be supplemented by G. PILLEMENT, *Paris inconnu: itinéraires archéologiques* (1965), and *Les Environs de Paris inconnus . . .*, 2 vol. (1961). Also by Pillement is a comprehensive series of works on the old houses or palaces: *Les Hôtels de Paris*, 2 vol. (1945); *Les Hôtels du Faubourg Saint-Germain* (1950); *Les Hôtels de l'Ile Saint-Louis, de la cité, de l'université, et du Luxembourg* (1951); *Les Hôtels d'Auteuil au Palais-Royal* (1952); *Les Hôtels des boulevards à Charonne* (1953); and *Les Hôtels du Marais*, 3rd ed. (1955). For the churches, see A. BOINET, *Les Églises parisiennes*, 3 vol. (1958–64). Of sentimental interest are L.P. FARGUE *et al.*, *Dans les rues de Paris au temps des fiacres* (1950); R. HERON DE VILLEFOSSE, *Histoire et géographie galantes de Paris* (1957); J. GALLOTTI, *Le Paris des poètes et des romanciers* (1964); and G. PILLEMENT, *Paris disparu* (1966) and *Les Environs de Paris disparus* (1968), with copious engravings and photographs.

Economy: R. MINGUET, *Géographie industrielle de Paris* (1957) and *Paris au travail* (1959).

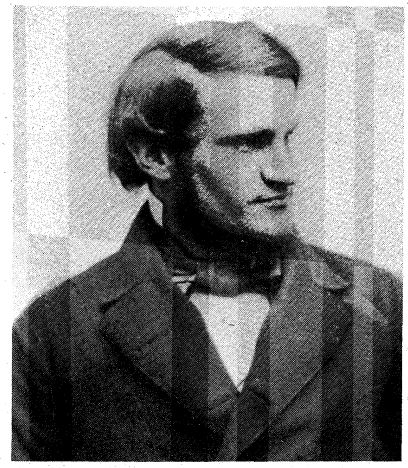
(B.E.)

Parkman, Francis

Francis Parkman, often called America's greatest historian, is best known for his seven-volume history of the Anglo-French struggle for North America, in which he displayed a rare talent for dramatic exposition combined with the strict historical accuracy of the scholar. He was born in Boston on September 16, 1823, the son of Francis Parkman, a leading Unitarian minister of Boston. As a boy, he met many of his father's literary friends and read widely in the family library. He spent many summers at his grandparents' farm in nearby Medford, while during the rest of the year he was drilled in Greek, Latin, and mathematics at the Chauncy Place School, in Boston, in preparation for Harvard.

Early years

At college Parkman, a talented linguist, read almost as many books in foreign languages as in English, including the original texts of great historians of antiquity and Italian writers such as Niccolò Machiavelli. He also devoured the major works of French literature and history. The romantic themes of many of the works of Chateaubriand and of Jules Michelet, the historian, attracted him, as had the novels of Sir Walter Scott. In serious archival studies he was encouraged by his Harvard teacher, the renowned historian Jared Sparks. Sparks, a man drawn to adventure and exploration, who wrote about the French explorer Robert Cavalier, sieur de La Salle, in North America, and at one time planned a safari to Timbuktu, exerted an enormous influence on an impressionable undergraduate such as Parkman.



Parkman.
By courtesy of the Metropolitan Museum of Art, gift of I.N. Phelps Stokes, Edward S. Hawes, Alice Mary Hawes, Marian Augusta Hawes, 1937

Though teachers and books helped to shape Parkman's thinking in his formative years, he gathered data, as indicated by his letters and journals, through direct observation. During his college years he exhausted friends who struggled to keep pace with him on woodland expeditions through New England and southeastern Canada. Yet he did not neglect to participate in whiskey punch and Indian war cries that sometimes followed dormitory suppers. Pretty girls and horses, he concluded, were "the 'first-ratest' things in nature." After a breakdown in health during his last year in college, he made a grand tour of Europe in 1844. His particular interest in the Roman Catholic Church prompted him to observe it at close range, even to the extent of living for a short time in a monastery in Rome. In the following year, he toured historic sites in the northwest of America and, to please his father, completed requirements for a law degree at Harvard. In the summer of 1846 he embarked on a journey to the Great Plains in which he travelled a portion of the Oregon Trail to Ft. Laramie.

Beginnings of literary career

Parkman's literary career had its real beginning after he returned from the West. Despite temporary illness and partial loss of sight, he managed to write a series of Oregon Trail recollections for the *Knickerbocker Magazine*. Published in 1849 as *The California and Oregon Trail*, the book's title was misleading because Parkman had ventured nowhere near California. He keenly regretted the "publisher's trick" of the mention of California as a stimulus to better sales. The book, in later editions called *The Oregon Trail; Sketches of Prairie and Rocky-Mountain Life*, became one of the best selling personal narratives of the 19th century.

The Oregon Trail served notice that a new writer, at home on the frontier as well as in staid, provincial Boston, had appeared. *History of the Conspiracy of Pontiac*, completed just before his marriage to Catherine Scollay Bigelow in 1851, was his first historical work, a comprehensive survey of Anglo-French history and Indian affairs in North America, culminating in the great Ottawa chief's "conspiracy" and Indian war of 1763. In the "dark years" of illness following the death of his young son (1857) and his wife (1858), Parkman entered a period of depression and semi-infirmity. His complaints of heart trouble, insomnia, painful headaches, semiblindness, wa-

ter on the knee, and finally arthritis and rheumatism, which fill his correspondence, were probably the result of what modern physicians have diagnosed as an underlying neurosis. By personalizing his illness and calling it the "enemy," Parkman seems to have forced himself to play the role of a man of action at the cost of great tension. His struggle against the "enemy" enabled him to maintain his self-respect and appears to be at least partly responsible for the powerful drive and creative force behind his writings.

History of France and England in North America

By the time the U.S. Civil War ended, Parkman had at least partly overcome his personal "enemy" of illness to complete his *Pioneers of France in the New World,* a vivid account of French penetration of the North American wilderness that created a setting for his later volumes. In the 27 years following the Civil War, Parkman (who had to content himself with writing militant, patriotic letters to the press during the conflict) completed his elaborate series by writing six more historical works in addition to the *Pioneers. The Jesuits in North America in the Seventeenth Century* (1867) is a powerful narrative of the tragedy of the Jesuit missionaries whose missions among the Hurons were destroyed by persistent Iroquois attacks, and his *La Salle and the Discovery of the Great West,* first published in 1869 as *The Discovery of the Great West* but later revised after French documents were made available, is in many respects one of the best one-volume biographies in the English language. La Salle, a hardy, gallant figure who overcame almost every obstacle in his path, was a heroic figure almost made for Parkman's pen. *Count Frontenac and New France Under Louis XIV* (1877) tells the story of New France, the early French settlement in Canada, under its most formidable governor, a man of vanity, courage, and audacity. Yet it was in *Montcalm and Wolfe* (1884)—a true biography of the French general Marquis de Montcalm and the English general James Wolfe, both of whom died at the Battle of Quebec in 1759—that Parkman not only reached his highest achievement in character portrayal but also showed how great biography can be used to penetrate the spirit of an age. By contrast, Parkman's *The Old Régime in Canada,* published in 1874, provides a sweeping panorama of New France in her infancy and youth, a pioneer work in social history that holds the interest of the reader no less than his narrative volumes. Parkman's literary artistry is perhaps best studied in his *Half-Century of Conflict,* completed shortly before his death which occurred in Jamaica Plain, Massachusetts, on November 8, 1893. This final link in his history *England and France in North America* is a fascinating but complex account of events leading up to the French and Indian War.

Assessment

Parkman's weaknesses as a writer have often been exposed. There is little question that he tended to stress dramatic qualities of his narrative and thus to give individuals an importance in events that they may not have deserved. His unbrotherly depiction of Indians as conspirators and ferocious savages, criticized by many of his contemporaries, obscured the fact that the Indians under Pontiac were fighting for self-determination. Parkman's belief that leadership should be in the hands of the cultured and capable was projected into his work to explain, for example, that La Salle accomplished as much as he did because he was "a man of thought, trained amid arts and letters." And there are passages that show that Parkman, in some cases, was a spokesman for Anglo-America who saw in French Canada a source of spiritual and political despotism that would attack society like mildew and stifle its growth. Yet because he showed that authoritarianism has its evils, he was not necessarily an exponent of democratic Anglo-Saxonism. Rather, he portrayed the Anglo-French and Indian wars as part of a struggle between contesting civilizations, in which the interior wilderness acted as a modifying force on rival colonial cultures. Perhaps Parkman's greatest achievement was his skill in recognizing the dramatic potentials in the raw materials of history, so that he could create a narrative both historically accurate and, as he said, "consistent with just historic proportion." When he wrote that his aim was "to get at the truth," he explained the lifelong search for factual data that underlies his entire work. Not all of his interpretations have been accepted unquestioningly, but Parkman's genius with the pen was such that his main figures—Frontenac, Montcalm, Wolfe, La Salle, and Pontiac—are not so much remembered today because of what they did but because Parkman made them the heroes of his history of Anglo-French rivalry in North America.

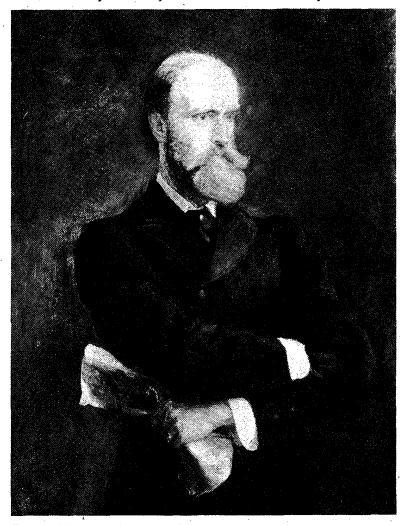
BIBLIOGRAPHY. WILBUR R. JACOBS (ed.), *Letters of Francis Parkman,* 2 vol. (1960), includes a short biography with an analysis of Parkman's illness. LOUIS CASAMAJOR, a physician, probes the psychological implications of Parkman's neurosis in "The Illness of Francis Parkman," *Am. J. Psychiat.,* 107: 749–752 (1951). MASON WADE, *Francis Parkman: Heroic Historian* (1942), is an excellent one-volume biography; and SAMUEL ELIOT MORISON (ed.), *The Parkman Reader* (1955), contains representative selections from Parkman's historical writings. Parkman's literary techniques and skills as a writer are explored by WILBUR R. JACOBS, "Some of Parkman's Literary Devices," *New England Quarterly,* 31:244–252 (1958); DAVID LEVIN, *History As a Romantic Art: Bancroft, Prescott, Motley, and Parkman* (1959); and HOWARD DOUGHTY, *Francis Parkman* (1962). The criticism of Parkman's work as a historian of the Indian is well exemplified by FRANCIS P. JENNINGS, "A Vanishing Indian: Francis Parkman Versus His Sources," *Pennsylvania Magazine of History and Biography,* 87:306–323 (1963); and ROBERT SHULMAN, "Parkman's Indians and American Violence," *Massachusetts Review,* 12:221–239 (1971). Further criticism and appraisal of Parkman's significance as a historian is in DAVID E. GRIFFIN, "'The Man for the Hour,': A Defense of Francis Parkman's Frontenac," *New England Quarterly,* 43:605–620 (1970); and in WILBUR R. JACOBS, "A Message to Fort William Henry," *Huntington Library Quarterly,* 16:371–380 (1953).

(W.R.J.)

Parnell, Charles Stewart

An Irish nationalist and the leader of the struggle for Irish Home Rule in the late 19th century, Charles Stewart Parnell, by his radical tactics and "new departure," upset the land settlement that had been imposed by the British on the Irish for three centuries. He took the question of self-government for Ireland out of its academic domain and made it a practical political issue, converting one of the two great British parties to its cause. British authority in Ireland, after Parnell, never again rested on so solid a foundation. Both Sinn Féin republicans and Irish literary revivalists saw him as a hero and a precursor.

By courtesy of the National Gallery of Ireland



Parnell, painting by Sydney Prior Hall, 1892. In the National Gallery of Ireland.

Parnell was born at Avondale, County Wicklow, on June 27, 1846. His mother was a daughter of Commo. Charles Stewart of the United States Navy, a hero of the War of 1812 whose parents had emigrated from Belfast before the American Revolution. The anti-British tradi-

tions and atmosphere of Parnell's home were significantly different from those of the majority of the Anglo-Irish Protestant landowning class to which he belonged. They did not, however, prevent his parents from giving him an education normal for his class. He went to three English boarding schools, where he seems to have been unhappy, and to Cambridge, where in 1869, after an undistinguished career, he was suspended for a relatively minor breach of discipline and decided not to return.

The Home Rule League and the Land League. The Ireland to which Parnell returned was in ferment (see BRITAIN AND IRELAND, HISTORY OF). The government's oppressive measures against the revolutionary Irish Republican Brotherhood (the Fenians) aroused intense national feelings among even the moderate Irish. In 1870 a new political group, the Home Rule League, was set up to press for Irish autonomy in local government; in 1874 it returned 56 candidates to Parliament, where they formed a party under the nominal leadership of Isaac Butt. Though socially conservative and deferential to the opinions of the Roman Catholic hierarchy, all appealed in some degree to the national sentiments of the electorate. Parnell, an eminently suitable Home Rule candidate, was elected to Parliament for Meath in April 1875. Within two years he distinguished himself by his indifference to the opinion of the House of Commons and his sensitivity to Irish nationalist opinion. He embraced the policy of obstructing English legislation to draw attention to Ireland's needs, and his handsome presence and commanding personality gave him a powerful appeal. In September 1877 the Home Rule Confederation of Great Britain elected Parnell its president; he had become, at the age of 31, the most conspicuous figure in Irish politics.

Policy of obstructionism

In 1878 an agricultural crisis in Ireland seemed to threaten a repetition of the terrible famine and mass evictions of tenant farmers of the 1840s. To resist eviction and make Irish landlordism unworkable, the Irish Land League was founded in 1879 by a Fenian, Michael Davitt. Many moderates condemned the league, but Parnell identified himself with it and became its first president, thus becoming the centre of the great "new departure" national movement in which revolutionary devotion was combined with agrarian agitation and was supported by the obstructionist tactics of the "active section" in Parliament. Soon after the general election of 1880, Parnell was elected chairman of the Home Rule group in the new Parliament. After the rejection by the House of Lords of a moderate measure for Irish land reform, Parnell organized a massive land agitation, for which he then won the support of the clergy and of "moderate" opinion. It was combined with parliamentary obstruction on so large a scale that ultimately 36 Irish members were suspended. At this time Parnell rejected a policy of secession from Parliament, put forward by the Land League.

The passage in 1881 of Gladstone's Land Act, which conceded the principle that fair rents could be judicially determined, presented Parnell with a serious test of statesmanship. Its passage was unquestionably a great achievement for the Land League, but the most active Land Leaguers were not content and a split in the movement seemed likely. This Parnell avoided by pursuing a policy moderate in substance—testing the act by bringing selected cases before the land commission—but making speeches couched in violent language. As a result, probably in accordance with his wish, he was, on October 13, 1881, lodged in Kilmainham jail, Dublin. This assured his continued popularity and absolved him of responsibility for subsequent events.

Imprisonment in Kilmainham

Parnell's arrest was followed by the suppression of the Land League and a winter of sporadic local terror. It became clear to the government that only Parnell could restore order. In the spring of 1882 Parnell began negotiations for his release, conducted in the main through Capt. William O'Shea, a "moderate" Home Rule member, whose wife had been Parnell's mistress since 1880. A settlement was reached, the so-called Kilmainham Treaty, whereby tenants were to obtain substantial concessions and Parnell was to use all his influence to decrease further agitation.

The murders by nationalists in Phoenix Park, Dublin, of the chief secretary and the permanent undersecretary, which occurred within a few days of Parnell's release (May 2, 1882), caused a general revulsion against terrorism, and Parnell had little difficulty in bringing the nationalist movement again under firm discipline, subordinating the Irish National League (the successor to the Land League) to the Home Rule Party in Parliament.

Parliamentary manoeuvres. The Kilmainham Treaty ended the revolutionary phase of the "new departure." The results of by-elections showed that Parnell's leadership was unquestioned, except in eastern Ulster, and, after the Reform Bill of 1884 extended the franchise to agrarian workers, it became apparent that Parnell was likely in the next Parliament to lead a party of between 80 and 90 members. With this potential strength Parnell became a force to be reckoned with. He contemptuously refused overtures made for his support by the radical wing of the Liberal Party led by Joseph Chamberlain and Charles Wentworth Dilke.

Alliance with Tories

The Tory advances to him led very quickly to a combination in which Tories and Irish voted together to defeat the Liberal government (June 1885), and in the election campaign that followed (November–December 1885), Parnell, having failed to get a satisfactory Home Rule statement from Gladstone, issued the "vote-Tory manifesto." Although the Irish could put the Liberals out, they could not keep the Tories in. In these circumstances, the Tories immediately broke with them and announced the intention of reintroducing coercion in Ireland. Parnellites and Liberals voted together to bring down the government, and Gladstone took office in February 1886. For his continuation in office he depended on Irish support.

There followed the curious and ominous episode of the Galway election. Parnell, under pressure from the O'Sheas and Joseph Chamberlain, put forward Captain O'Shea as Home Rule candidate, although he had refused to take the pledge "to sit and vote with the party." The evidence suggests that Chamberlain was attempting to undermine Parnell's authority and split his party. If so, he failed. A mutiny of a small faction was quelled and O'Shea was elected.

Although Gladstone's Home Rule proposals—involving a wide measure of autonomy—fell short of nationalist aspirations, Parnell accepted them as a basis of settlement and enlisted public opinion in their support. The introduction of the bill, though it was later rejected by the Commons on the second reading (June 1886), was regarded as his personal triumph. When the Conservative Lord Salisbury succeeded Gladstone as prime minister, Parnell withdrew to some extent from active political life. This was partly due to ill health but also to political reasons. With the Irish party firmly allied to the opposition, there was now no room for parliamentary obstruction. Parnell would neither challenge Gladstone's leadership nor appear as his henchman. He also held aloof in Ireland from the ingenious rent-withholding combination known as the plan of campaign, devised by William O'Brien.

The Times forgery

Despite his relative inactivity Parnell was kept before the public through the efforts of his enemies. On April 18, 1887, *The Times* published a facsimile of a letter purporting to be written by Parnell condoning the Phoenix Park murders of May 1882. Parnell immediately denounced it as a forgery. Nearly two years later the forger, a journalist named Richard Pigott, collapsed under cross-examination before an investigating commission. Parnell, after Pigott's suicide in Madrid soon afterward, was transformed in the eyes of the English liberals from a dubious ally into a hero and martyr. This brief period was the peak of Parnell's career.

Parnell's fall. On December 24, 1889, Captain O'Shea filed a petition for divorce, naming Parnell as co-respondent. Although Parnell's liaison had been known to some members of the Irish party, nationalist Ireland in general took it that the proceedings represented another attempt to wreck Home Rule. This was given colour by the fact that O'Shea was a follower of Joseph Chamberlain. The theory that there were political motives behind the di-

vorce proceedings is not necessarily false. The suit being undefended, the court returned a verdict against Parnell and Mrs. O'Shea on November 17, 1890.

The initial reaction of the Irish public was to uphold Parnell. In Britain, however, Nonconformist opinion was so hostile that the Irish parliamentary party found itself in an agonizing dilemma. Parnell was determined to hold the leadership and defy Gladstone. If the party upheld Parnell they would be destroying the Liberal alliance, and with it the hopes of Home Rule in their generation. If they rejected Parnell, they would be turning against him at the bidding of an Englishman. After a long and emotional debate, the majority rejected his leadership; a sizable minority remained with him.

There followed a series of bitter electoral campaigns. The Catholic hierarchy, although slow to pronounce, now declared Parnell morally unfit for leadership. His marriage to Mrs. O'Shea in June 1891 exacerbated Catholic opposition. He himself displayed feverish energy and increasing recklessness, directing his appeal more and more to the revolutionary elements. This appeal left a deep impression on the young but was rejected by the majority of the nation. When his principal ally, the nationalist *Freeman's Journal*, fell to his enemies shortly after his marriage, his cause was clearly lost. He died at his wife's home in Brighton on October 6, 1891, and was buried in Glasnevin Cemetery, Dublin. The city, Parnellite to the end, gave him a magnificent funeral.

BIBLIOGRAPHY. R.B. O'BRIEN, *The Life of Charles Stewart Parnell, 1846–1891*, 2nd ed., 2 vol. (1899), has not been superseded, although several lives of Parnell have been published since. CONOR CRUISE O'BRIEN, *Parnell and His Party, 1880–90* (1957, corrected impression 1964), is a study of Parnell's political leadership and of the political machine that he controlled. F.S.L. LYONS, *The Fall of Parnell, 1890–91* (1960), is an excellent detailed study of the divorce crisis and its sequel; and the same author's *Parnell* (1963), a very useful short study.

(C.C.O'B.)

Particles, Subatomic

Subatomic particles, often called elementary particles, are the fundamental units of matter and energy.

The growth of knowledge in the present century has profoundly altered the accepted view of the nature and properties of subatomic particles. As of the early 1970s, the existence of more than 200 such particles has been firmly established, and a number of laws governing the relations among them had been stated. No complete theory as yet exists, but experimental and theoretical research is being carried out on a large scale at the major universities of the world and at laboratories organized especially for this purpose such as the CERN (European Organization for Nuclear Research) laboratory at Geneva, the National Accelerator Laboratory near Chicago, and the Serpukhov laboratory near Moscow. (For a more detailed account of the research tools used in this work, see ACCELERATORS, PARTICLE; RADIATION DETECTION AND CHARACTERIZATION.)

This article is divided into seven parts. The first section embraces the concepts and theories of subatomic particles developed in historic sequence before 1960. During this period relatively few particles were known, but there were complicated developments, both conceptual and experimental, that make this section more difficult to understand than those that follow it. The second section, *Four basic forces and four classes of particles*, which may be read independently, describes the hundreds of new particles discovered since 1960. The next section, *Conservation laws and quantum numbers*, treats the principles of basic symmetry, by which the new multiplets of particles are organized. The fourth section, *Spectroscopy of the hadrons; the quark model*, develops a table of particles. The fifth section, *Experiments*, returns to a historical description, starting in 1960. The section on *Relations between the interactions* is concerned with the concepts of currents and the algebra of currents. The final section, *Undiscovered particles*, deals with theories that assume the existence of certain undiscovered particles.

HISTORY UP TO 1960

Atomic structure and nuclear forces. This section reviews developments of particle physics up to the early 1960s. The reader who does not need this detailed history, but is more interested in the hundreds of new particles discovered in the 1960s and the new concepts such as the quark model, by which the new periodic table of particles is organized, can omit this section.

The idea that matter is composed of elementary building blocks goes back at least as far as Democritus (500 BC). Until the discovery of the electron and the structure of the atom (see the articles ELECTRON; ATOMIC STRUCTURE), atoms could be and sometimes were regarded as elementary particles in this sense, and, indeed, the 100 or so different atoms are still spoken of as chemical "elements." After the discovery of the "electron" (e) in 1897 and of atomic nuclei in 1911, however, it was clear that an atom had to be regarded as a composite system consisting of a cloud of electrons surrounding a tiny heavy central core, the atomic nucleus. The electron cloud was found to obey quantization rules in that it can exist only in discrete, or definite, energy states (also called energy levels). The state of lowest energy, the ground state, is stable, whereas those of higher energy, the excited states, can make transitions to lower states with the emission of light quanta ("photons," symbol γ) the energy of which makes up the difference in energy between the initial and the final states involved in the transition. This concept was not at first supported by a systematic quantitative theory, though the behaviour of the electron clouds could be semiquantitatively described by a combination of the classical theory of electricity and magnetism with the quantization rules of Niels Bohr, a Danish physicist. A far-reaching improvement in the theoretical situation resulted from the discovery in 1926 of the true quantum mechanics (see MECHANICS, QUANTUM). Once the quantum mechanics of the electromagnetic field had been worked out in 1927 by P.A.M. Dirac, an English physicist, a quantitative theory of photon emission and absorption that agreed with experiment became possible.

Early notions of elementary particles

In summary, in the theory of atoms of the late 1920s, electrons, nuclei, and photons were all understood as elementary particles, and the electrons and nuclei were understood as bound together to form atoms by the electromagnetic forces between them.

That the treatment of the atomic nucleus in this theory of atoms is only schematic was recognized from its inception. Experiments on nuclei had shown that, like atoms, nuclei can exist in excited states. Thus an extension to atomic nuclei of the point of view that had been successful in the theory of atoms led to a picture in which the constituent protons (the nuclei of hydrogen atoms) and other (then unknown) particles are also governed by the laws of quantum mechanics and are bound together by (then unknown) forces. Within a few years experimental discoveries confirmed this general atomic picture, and it was reduced to a concrete quantitative theory. In 1932, James Chadwick, an English physicist, discovered the neutron (n), a particle slightly heavier than the proton and of zero electric charge, and evidence grew that the constituents of all atomic nuclei are "nucleons," nucleon being a generic term meaning either proton or neutron. The nucleons are bound together, according to the laws of quantum mechanics, by a nuclear force acting strongly only when the particles are separated by extremely small distances (less than a few times 10^{-13} centimetre). Nuclear force refers to those forces other than electrical that act between nucleons (see NUCLEUS, ATOMIC).

Nuclear forces

Quantum mechanics also made possible a quantitative description of the phenomenon of beta (β-) radioactivity, in which a nucleus, through the emission of both a negative electron and a neutral particle of small mass (usually taken to be zero), undergoes a transition to a lower energy state, thus becoming a nucleus with one more proton and one less neutron. This neutral particle, later named the neutrino, is difficult to detect because its interaction with matter is so weak. The first experiment in which a neutrino beam produced an observable effect was reported in 1956.

The three phenomena mentioned above illustrate a classification of interactions that is basic for an understanding of the remainder of this article: strong, electromagnetic, and weak interactions. Nuclear forces are strong interactions, stronger than electromagnetic. The difference in strength implies that nuclear binding energies are much larger than atomic energies. On the other hand, weak interactions, such as those responsible for beta radioactivity, are much weaker than electromagnetic.

At first sight, this theory of atomic nuclei would seem to be a natural extension of the older view that matter is composed of an elementary building block, the atom: a list of elementary particles should include electrons, protons, neutrons, and neutrinos. The theory of atomic nuclei, however, contained some radically new ideas about transformations of elementary particles. The neutron can transform into a proton and emit an electron and a neutrino. This transmutation contrasts with the behaviour of the electrons in atomic theory; in the atom, electrons retain their identity while interacting with photons. To develop the full consequences of such predictions, it is necessary to make an apparent digression to describe developments in the relativistic theory of electrons.

Negative energy states and quantum field theory. The first quantum-mechanical theories of the electron, those used in the theory of atoms in the mid-1920s, were nonrelativistic and did not take into account the effects of the finite propagation velocity of light. In 1927 Dirac introduced a relativistic theory of the electron that was successful in accounting for the spin of the electron and for the fine structure of the energy levels of hydrogen (*i.e.*, some energy states have a substructure). Unfortunately, the theory also predicted the existence of states the properties of which did not make sense, the so-called negative energy states. All attempts to amputate these negative energy states from the theory proved unsatisfactory.

Negative energy states

In 1930 Dirac made a bold proposal to put the negative energy states to good use to explain the existence of the proton. He noted that, according to the so-called Pauli exclusion principle (see MECHANICS, QUANTUM), two electrons in an atom cannot occupy the same state. He then made the hypothesis that *in vacuo* all the negative energy states predicted by his theory are occupied by electrons but that the resulting charge distribution is unobservable. Thus a positive-energy negatively charged electron (e^-) moving in the vacuum cannot jump into a negative energy state with the emission of a photon because there is no negative energy state that is not occupied. Furthermore, the removal of an electron from a negative-energy state would result in an observable "hole," which, Dirac showed, would behave like a normal positive-energy particle of opposite charge (e^+). His initial proposal to interpret the proton as this hole, however, turned out to be unsatisfactory for two reasons. First, it was found that a hole in the sea of negative-energy electrons would have to have the same mass as an ordinary electron, in disagreement with the observation that protons are 1,836 times heavier then electrons. Second, the hydrogen atom formed of an electron and a hole would annihilate itself in a small fraction of a second leaving nothing but photons. In 1931 Dirac concluded that, if holes existed, they would describe a new positive particle.

About a year later the positive electron was discovered by Carl D. Anderson, a U.S. physicist, who named it the positron. (There is a standard ambiguous usage of the word electron. Sometimes it stands for the negatively charged, e^-, and sometimes it is used as a generic term for the positive or negative particles, $e^\pm$. The symbols e^+ and e^- should be used when any significant ambiguity arises. A similar ambiguity, in the case of neutrinos, is resolved by distinguishing between them as neutrinos and antineutrinos.) Dirac's prediction, one of the most spectacular achievements of theoretical physics, was comparable to the prediction by Hideki Yukawa, a Japanese physicist, of the existence of another elementary particle, the meson (see below *Yukawa mesons and strong interactions*).

It was later shown that Dirac's hole theory is essentially equivalent to a theory in which electrons and positrons appear symmetrically with positive energies only, so that there is no unobservable sea of negative-energy electron states. The natural language for such a description is that of the quantum theory of fields. Dirac's relativistic theory of the electron was reinterpreted as a theory of two interacting fields, the electron–positron field and the electromagnetic, or photon, field.

The notion of the particle has quite different statuses in the relativistic quantum theory of fields and the nonrelativistic theory of particles. A leading idea of quantum field theory is that all particles of the theory appear as quanta of some field. For example, photons are particles associated with the electromagnetic field, whereas electrons and positrons are quanta of the Dirac field. This appears to be a perfectly innocent requirement, and indeed it is, until the additional requirement is made that the theory be consistent with the special theory of relativity. Then, remarkably, it is found that, if the particles are to interact with one another at all, it must also be possible for them to be created and annihilated. Thus, if electrons and photons are to be described by a relativistic field theory and interact at all, the creation and annihilation of electron–positron pairs must be possible. The fact that particle creation and annihilation necessarily takes place in such a theory makes the definition of an elementary particle essentially more complicated. Even if in a given physical situation no actual particle creation is possible without violation of the laws of conservation of energy and momentum (*i.e.*, neither energy nor momentum—mass times velocity—can be destroyed), nevertheless the interaction responsible for pair creation and annihilation has indirect effects, the so-called virtual creation and annihilation processes. In such a theory, for example, a single electron moving through the vacuum polarizes it (*i.e.*, the electron may be thought of as surrounded by a cloud of virtual photons and electron–positron pairs). Under such circumstances there are at least two possible ways (and actually many more) to define an electron: with its cloud (the dressed electron) or without (the bare electron). The situation is even more complicated in the theory of beta radioactivity put forth by Enrico Fermi, an Italian physicist, in which a proton has to be considered as existing virtually as a neutron plus a positron plus a neutrino, as well as having a cloud of photons, e^+e^- pairs, and nucleon–antinucleon pairs surrounding it. Such a notion of an elementary building block is remote from the pictures of particles as little billiard balls that had motivated the work only a few years earlier, and it was not the end of the evolution of the idea of elementary particles.

Annihilation of particles

Yukawa mesons and strong interactions. In 1934 Yukawa proposed a theory of nuclear forces that turned out to be of great significance. Yukawa's ideas were based on an instructive analogy with the theory of electrical forces, which will next be explained. Coulomb's law of electrostatics states that the force that two charges, e_1 and e_2, exert on one another when they are separated by a distance r is given as the product of the two charges divided by the distance of separation squared, e_1e_2/r^2. There is an equally accurate but somewhat less familiar description of this force (shown schematically in Table 3) that attributes it to the exchange of virtual photons between the two charges. The factors e_1e_2 in Coulomb's law then comes about because the ability of the first charge to emit and absorb photons is proportional to e_1 and that of the second charge is proportional to e_2. The dependence of the force (F) on the separation turns out to be the simple geometrical factor, $1/r^2$. Yukawa pointed out that, if two nuclear particles, instead of emitting and absorbing massless photons, emitted and absorbed certain particles of finite mass, $m \neq 0$ (later called mesons), then (see again Table 3) the only essential change in the results is the replacement of quantities e_1 and e_2 by two other numbers, g_1 and g_2, which are measures of the ability of the particles to emit and absorb mesons, and the appearance of an exponential factor, $\exp[-r/(\hbar/mc)]$, in Yukawa's force law, in which the symbol $\hbar$ is Planck's constant divided by two times pi ($\hbar = h/2\pi = 1.06 \times 10^{-34}$ joule-second) and c is the speed of pho-

tons in vacuo (3×10^8 metres per second—*i.e.*, the velocity of light). This new factor is important because it means that the force between two particles falls off much faster with increasing separation than it would according to the Coulomb's law. Stated briefly, the force that could arise from the exchange of virtual mesons of mass m has a range that is approximately inversely proportional to that mass—*i.e.*, a length called the Compton wavelength of the meson. For a mass of the order of magnitude of a few hundred electron masses, the quantity $\hbar/mc$ is of the order of magnitude of 10^{-13} centimetre, the range observed for nuclear forces. The strength of the forces is determined by the parameter, or quantity, g ($g_{NN\pi}$ in Table 3) for a nucleon, called the meson–nucleon coupling constant (see below *Four basic forces and four classes of particles*).

Problem of the "divergencies"

Two decades of experimental and theoretical effort were necessary before it became clear that Yukawa was essentially correct. The situation was much complicated by the fact that, during the 1930s and '40s, it was widely believed that quantum electrodynamics is a defective theory. The prime pieces of evidence for this belief were the "divergencies"—*i.e.*, the infinitely large or ambiguous answers to some simple physical questions that follow from the equations of the theory. These difficulties were almost universally attributed to the physical assumption underlying the theory, viz., that electrons are point particles. It was well-known that difficulties of a similar type occur in the classical theory of electrons and are overcome by attributing a spatial extension to the electron. In the classical theory that predates the era of quantum physics, this spatial extension led to modifications not only of the divergent quantities of the theory but also of the quantities for which the point electron theory predicts finite answers. Similar effects were expected in quantum theory. Thus, for example, a breakdown in the theory was expected in the description of processes in which an electron, colliding with a nucleus, loses energy by the emission of photons. When, between 1933 and 1936, preliminary evidence appeared in cosmic rays for particles with anomalously high penetrating power, it was tempting to explain them as electrons the energy of which was sufficiently high that the predictions of quantum electrodynamics break down. This explanation was short-lived because more detailed experiments showed there were two groups of particles, one of which lost energy by radiation precisely according to the predictions of quantum electrodynamics as worked out in 1934 by Hans Bethe and Walter Heitler, whereas particles of the other group were far more penetrating. By 1937, the conclusion was inescapable; the discovery of a new particle was announced, first called a mesotron, known later as a meson, and now called the muon (μ) the mass of which is 207 times that of the electron. The laws of electrodynamics had not broken down in the anticipated way at all, and Yukawa's predicted meson had appeared—or so it seemed.

The two-meson hypothesis

It was ten years later, in 1947, that Italian physicists Marcello Conversi, Ettore Pancini, and Oreste Piccioni found that those same cosmic-ray mesons were captured in solid materials at a rate 10^{12} times slower than Yukawa's theory predicted. Numerous suggestions of varying plausibility were made to save the theory. The one that proved to be correct (1947) was a two-meson hypothesis: the mesons observed at sea level are not Yukawa mesons but weakly interacting decay products (muons; μ) of the true Yukawa particles, themselves produced copiously in nucleon–nucleon collisions high in the atmosphere. Shortly thereafter, C.M.G. Lattes and co-workers at the University of California, Berkeley, turned a plausible explanation into an important discovery by publishing pictures of the real positive Yukawa particles produced in a cyclotron, later called pions (π^+) that decay into one positive particle, later called a muon (μ^+) and one neutral particle, later called a μ-neutrino (ν_μ). The decay proceeds according to the following schematic representation: $\pi^+ \rightarrow \mu^+ + \nu_\mu$. Cloud-chamber experiments later showed that the muon in turn decays into a positron (positive electron) and two light neutral particles by a reaction that can be written as $\mu^+ \rightarrow e^+ + \nu_e + \bar{\nu}_\mu$, in which the symbol $\bar{\nu}_\mu$ represents a μ-antineutrino and ν_e a neutrino. The discovery of pions, π^+, and muons, μ^+, was followed shortly by systematic studies of their negative counterparts, the negative pion, π^-, and the negative muon, μ^-, by means of the nuclear disruptions or "stars" that they cause when absorbed in photographic emulsions.

Ignorant of the fact that the experimental meson was not the Yukawa meson, theoretical physicists systematically developed a theory of mesons in the period from 1934 to World War II. The mesons introduced by Yukawa were positively or negatively charged. Charged and neutral mesons of spin zero and one were subsequently considered. One of the most significant developments of the period (1938) was the creation by the English physicist, Nicholas Kemmer, of the charge-symmetrical theory, in which the interaction of nucleons with charged mesons is so related to their interaction with neutral mesons that the resulting nuclear forces are the same between proton and proton, proton and neutron, and neutron and neutron. What was emerging was a basic symmetry, which, in a terminology to be explained below, is isospin, or $SU(2)$, symmetry (see below *Spectroscopy of the hadrons; the quark model*). Applied to pions the charge-symmetrical theory led to the prediction of a neutral pion (π^0), which was found in 1950 by the detection of photons resulting from the decay of the neutral pion into two gamma ray photons, $\pi^0 \rightarrow \gamma + \gamma$.

Isospin

If the state of elementary-particle theory had been assessed, say in 1950, a long list of fundamental building blocks, would have been encountered: photon (γ), negative electron (e^-), positron (e^+), proton (p), neutron (n), negative pion (π^-), neutral pion (π^0), positive pion (π^+), negative muon (μ^-), positive muon (μ^+), neutrino (ν), and antineutrino ($\bar{\nu}$). They are shown in Table 1.

Table 1: Stable and Metastable Particles

symbol	J^P	helicity	mass (MeV)	mean life (sec)	typical decays (or name)
Quanta					
γ	1^-	± 1	0	stable	photon
g	2^+	± 2	0	stable	graviton
Leptons					
ν_e, ν_μ	$\frac{1}{2}$	$-\frac{1}{2}$	0	stable	neutrino
e^-	$\frac{1}{2}$	$\pm\frac{1}{2}$	0.5	stable	electron
μ^-	$\frac{1}{2}$	$\pm\frac{1}{2}$	106	2×10^{-6}	$\rightarrow e^-\bar{\nu}_e\nu_\mu$
Mesons (B = O)		Y^*, I^*			
$\pi^\pm$	0^-	0, 1	140	3×10^{-8}	$\rightarrow \mu^\pm\nu_\mu$
π^0	0^-	0, 1	135	1×10^{-16}	$\rightarrow \gamma\gamma$
$K^\pm$	0^-	$\pm 1, \frac{1}{2}$	494	1×10^{-8}	$\rightarrow \mu^\pm\nu_\mu$
K^0	0^-	$\pm 1, \frac{1}{2}$	498	1×10^{-10} 5×10^{-8}	$K_S \rightarrow 2\pi$ $K_L \rightarrow 3\pi$
η	0^-	0, 0	549	3×10^{-19}	$\rightarrow 3\pi$
Baryons (B = 1)					
p	$\frac{1}{2}^+$	$+1, \frac{1}{2}$	938	stable	
n	$\frac{1}{2}^+$	$+1, \frac{1}{2}$	940	1×10^{3}	$\rightarrow pe^-\bar{\nu}_e$
Λ	$\frac{1}{2}^+$	0, 0	1,116	3×10^{-10}	$\rightarrow p\pi^-$
Σ^+	$\frac{1}{2}^+$	0, 1	1,189	$.8 \times 10^{-10}$	$\rightarrow p\pi^0$
Σ^0	$\frac{1}{2}^+$	0, 1	1,193	$\sim 10^{-19}$	$\rightarrow \Lambda\gamma$
Σ^-	$\frac{1}{2}^+$	0, 1	1,197	1×10^{-10}	$\rightarrow n\pi^-$
Ξ^0	$\frac{1}{2}^+$	$-1, \frac{1}{2}$	1,315	3×10^{-10}	$\rightarrow \Lambda\pi^0$
Ξ^-	$\frac{1}{2}^+$	$-1, \frac{1}{2}$	1,321	2×10^{-10}	$\rightarrow \Lambda\pi^-$
Ω^-	$\frac{3}{2}^+$	$-2, 0$	1,672	1×10^{-10}	$\rightarrow \Lambda K^-$

* Y = hypercharge, I = isotopic spin.

There are only five stable particles; the rest are metastable, *i.e.*, immune to strong decay. About 25-meson resonances and 50-baryon resonances are omitted. The particles listed in Table 1 are plotted in Figure 1. Furthermore, there was clear experimental evidence that this list was not complete. In 1947, two clear pictures made with a cloud chamber (a device that shows particle tracks in a gas) showed the decays of two new particles, one neutral and one charged, which came to be known as V particles because of the characteristic appearance of their

decays in a cloud chamber. What emerged after some years of painstaking study was the addition to the list of elementary particles (see Table 1) of four "strange" *K*-mesons with masses between those of the pion and the proton—two neutral (K^0 and $\overline{K}^0$) and two charged ($K^{\pm}$)—and six "strange" baryons with masses greater than that of the proton, called Lambda neutral (Λ^0), Sigma minus, neutral, and plus (Σ^-, Σ^0, Σ^+), and Xi zero and Xi minus (Ξ^0, Ξ^-). (For an explanation of "strangeness," see below *Strangeness, or hypercharge.*) To regard all these particles as fundamental building blocks was to invite comparison with the Ptolemaic theory of the solar system (with its many epicycles).

Elementary and "strange" particles

Meanwhile, theoretical developments cast a new light on the conceptual foundations of the theory.

Theoretical development. As was pointed out above in connection with the discovery of mesons, "divergences" can appear in the answers given by electrodynamics to very simple questions. In the late 1940s great progress in overcoming these difficulties was made by the theoretical physicists Shin'ichiro Tomonaga in Japan, Julian Schwinger, Richard P. Feynman, and Freeman J. Dyson of the United States, and others. Their contribution was twofold. First, they reworked the formalism of electrodynamics, using methods that made calculation and insight much easier. Second, they showed that the divergences appearing in physical quantities are all direct or indirect consequences of two fundamental divergences, those in the corrections to the mass and to the charge of the electron arising from its interaction with photons. If the corrected mass and charge are set equal to the observed mass and charge of the electron (a process called mass and charge renormalization), all answers to the physical questions become finite.

Actually, much of this advance in theory was stimulated by two fundamental experimental discoveries explained below: the anomalous gyromagnetic ratio of the electron and the Lamb shift. In experiments carried out in 1947 in the United States, physicists Polykarp Kusch and Henry M. Foley showed that the value of the spin magnetic moment (a quantity associated with the torque experienced by a rotating charged body in a magnetic field) of the electron is not $e\hbar/2mc$ (e is the charge on the electron, m is its mass, c is the velocity of light, and $\hbar$ is Planck's constant divided by 2π), as had been predicted by Dirac's 1927 theory of the electron but, rather, this quantity multiplied by a factor, 1 plus a small constant, or $(e\hbar/2mc)\ (1 + \varepsilon)$, in which ε is small but nonzero. (The quantity $2[1 + \varepsilon]$ is called the gyromagnetic ratio.) Later measurements by others gave a value of the constant ε equal to $0.001\ 159\ 622 \pm 0.000000027$. At about the same time Willis Eugene Lamb, Jr., a United States physicist, measuring the position of the energy levels of hydrogen, found a shift from those predicted by Dirac's 1927 theory. From the 1930s the hole theory had predicted corrections to Dirac's original values for both quantities, but the corrections were, unfortunately, divergent. The existence of these experiments was an enormous incentive to make some sense out of the predictions of the theory. After this had been done in a more or less ad hoc way for these particular experiments, it was extended to the general theory described above. The detailed comparison with experiment gives excellent agreement. For example, theory gives an equation for the quantity ε as equal to $(\alpha/2\pi) - 0.328\ \alpha^2/\pi^2$, in which α, the fine structure constant, is equal to $1/137.03602 \pm .00021$. Substitution in this equation gives the value ε equal to 0.001 159 641, as compared with the experimental value above.

The Lamb shift

The renormalization process was extended to meson theory in 1951, when it was shown that renormalization of nucleon mass, meson mass, meson–nucleon coupling constant g, and a meson-scattering constant is sufficient to yield finite answers for all physical quantities. It turned out, however, that these results were of little practical value because the entire method was based on expansions of the quantities involved as a sum of terms proportional to powers of the quantity $g^2/\hbar c$ (a so-called perturbative expansion); for the large values of this quantity $g^2/\hbar c$ necessary to describe mesons, it is doubtful whether the sum is a useful expression. The situation is different in quantum electrodynamics in which the fine-structure constant $e^2/\hbar c$ is small and the first few terms of the expansion are a good approximation.

The remarks of the preceding paragraph are indications of a general problem. The success of renormalization theory represented great progress in quantum field theory but did not go far enough. The infinities could not be eliminated from the theory without resort to perturbative expansions. The response to this problem was twofold: the development of a general theory of quantized fields and dispersion theory. Since the latter became immediately an important practical tool, it will be described first in the following discussion.

Dispersion theory

The quantitative description of the collisions of subatomic particles is in terms of their scattering and production "amplitudes." These are complex-valued functions (such a function contains a real term and an imaginary term—*i.e.*, one containing the factor $\sqrt{-1}$) of the parameters describing the collisions (energy of particles, angle through which they are scattered, and so on). The square of the absolute value of the function gives the probability of the collision. Dispersion theory makes two kinds of assertions about such amplitudes. First, it makes the qualitative assertion that they are analytic functions (see ANALYSIS, COMPLEX). This is a strong restriction on the kind of behaviour the function can have. Second, it makes quantitative assertions—*e.g.*, that the real part of the function is equal to a certain integral transform (a mathematical operation) of the imaginary part. Such an equation is usually called a dispersion relation because the first examples in physics occurred in the optics of dispersive media. In that case, the function in question is the complex refractive index of the medium regarded as a function of the frequency of light; and the dispersion relation connects the real part of the refractive index with the imaginary part, which is essentially the absorption coefficient of the medium. In 1927, Hendrik Anthony Kramers, a physicist in The Netherlands, showed that this relation is a consequence of the principle that signals cannot be propagated faster than the speed of light. In 1946 a similar relation was suggested for the amplitudes describing collisions of the particles; but it was not until 1954 that physicists were successful in linking dispersion relations directly to quantum field theory. Within a few years, dispersion relations for a larger number of collision processes were found. Although in its modern form it was sired by quantum field theory, dispersion theory soon developed a technique and an outlook independent of its starting point. From the middle 1950s on, much of the quantitative description of particle reactions was worked out in the language of dispersion theory.

Most dispersion relations contain quantities (the analytic continuation of amplitudes) that are not directly observable. There are dispersion relations, however, that contain only observable quantities. Since they are in good agreement with experiment, these provide a valuable support for the ideas of dispersion theory.

Although the general theory of quantized fields eventually provided a derivation of dispersion relations starting from first principles, its initial objectives were conceptual; it sought to locate those general properties of a field theory indispensable for a reasonable description of particles, and then to see what consequences followed. A clear cut example of this kind was the proof in 1957 of the so-called *CPT* theorem. The theorem says that under very general assumptions, the equations that describe the laws of nature possess a certain symmetry; *i.e.*, they remain unchanged in form under a certain transformation denoted *CPT*. (The letters indicate that the transformation behaves like an inversion of space–time coordinates, *PT*, followed by an interchange of particle and antiparticle, *C*; *C* is sometimes called charge conjugation, whence the letter *C*; *P* recalls the parity operation, which replaces the space coordinate x by $-x$; *T* stands for time inversion, which replaces the time coordinate t by $-t$.) As a consequence of *CPT* symmetry, for example, the masses of particle and antiparticle have

The *CPT* theorem

Table 2: The Four Basic Forces

force	acts on:	"quanta" exchanged	range	strength (*see* Table 3, col. 3)	examples: stable systems	examples: reaction induced by force
Gravity	all particles*	graviton, g	long, *i.e.*, $F \propto 1/r^2$	$\sim 10^{-42}$	solar system	object falling
Weak interaction	all particles except γ and g	proposed weak boson, W	$<10^{-14}$ cm	unknown	none	radioactivity: *e.g.*, uranium→lead
Electromagnetism	particles with electric charge†	photon, γ	long, *i.e.*, $F \propto 1/r^2$	1/137	atoms, rocks	chemical reactions
Strong interaction	hadrons: mesons (π, K, . . .) baryons (N, Λ, . . .)	the hadrons themselves	10^{-13} cm	10	hadrons, nuclei	πp scattering, nuclear reactions

*Because all particles have energy or rest mass. †Here electric charge includes magnetic properties and (if they exist) magnetic monopoles.

to be exactly equal, and so do their total lifetimes against decay if they are unstable. The theorem holds even when space inversion, P, or particle–antiparticle transformation, C, do not separately define symmetries of the system. That is fortunate because one of the great discoveries (the parity revolution) of the 1950s was that neither parity P nor charge C is a symmetry in weak interactions such as those that cause decay of the neutron and muon (United States physicists Tsung-Dao Lee, Chen Ning Yang, Chien-Shiung Wu, and others). See CONSERVATION LAWS AND SYMMETRY.

Dawn of a new age. Rather than continuing the above history of particle physics up to the present, a description of events since the parity revolution of 1956 will be incorporated in the systematic account of present knowledge, found below. To complete this section, however, a few general remarks are in order. In the decade of the '60s, there was a continuation of the steady progress that had marked earlier decades, but in another respect it was a new age. The new age came about because of the discovery of literally hundreds of new particles, the resonances. It became clear once and for all that not every such particle should be regarded as an elementary building block. It was and is much more natural to regard the particles as different states of excitation of matter and to seek to classify them by methods analogous to those of the spectroscopy of atoms and atomic nuclei. Much of the theoretical work of the 1960s can be regarded as an effort to provide such a classification.

FOUR BASIC FORCES AND FOUR CLASSES OF PARTICLES

Four basic forces. The particles observed in nature are acted on by, and at the same time give rise to, various forces. There are four basic forces, differing greatly in strength. Listed in order from the weakest to the strongest, they are gravitation, weak interaction, electromagnetic interaction, and strong interaction. Some of the salient features of each are listed in Table 2.

Force of gravity. The force of gravity is unique among forces because, as Sir Isaac Newton said, it is universal; all objects that have mass or energy attract each other with a gravitational force. Two particles of masses m_1 and m_2 separated by a distance r attract one another with a force proportional to the product of the masses divided by the distance squared, or in equation form, $F = Gm_1m_2/r^2$, in which $G = 6.67 \times 10^{-8}$ centimetre^3-gram^{-1}-second^{-2} is the Newtonian constant of gravitation. This is called a long-range force (or sometimes infinite-range force) to contrast its $1/r^2$ dependence with the much more rapidly decreasing $1/r^2 \exp[-(r/a)]$ dependence of strong interactions, which are said to have range a. Although, compared with all other forces, the force of gravity is extraordinarily weak and therefore normally negligible, there are circumstances in which it dominates. For example, the motion of the planets in the solar system is essentially determined by their gravitational attraction because they are so far outside the range of strong forces and they are approximately neutral electrically, so that electric forces are unimportant.

The Newtonian constant of gravitation

Weak interaction. The weak interaction is more conspicuous for the transmutations it produces (the slow decays of particles) than for the force to which it gives rise. In fact, the force has never been detected directly. At short distances ($\sim 10^{-13}$ centimetre) it is swamped by the much larger strong interaction, at atomic distances ($\sim 10^{-8}$ centimetre) it is unappreciable compared with Coulomb forces, and (at still larger distances where Coulomb forces are shielded) the weak interaction is still smaller than the intermolecular forces arising from the electrical polarizability (charge separation) of molecules, forces that are difficult to treat theoretically and measure experimentally.

Electromagnetic interaction. The electromagnetic interaction gives rise to the forces between electrical charges and currents. The simplest example is that already described above, Coulomb's law for the force between charges e_1 and e_2, in which the force is equal to the product of the charges divided by the square of the separation distance (r); *i.e.*, $F = e_1e_2/r^2$; for two charges of the same sign the force is repulsive and for unlike charges the force is attractive. All charges occurring in nature are integer multiples of the basic unit e, the magnitude of the charge on the electron. Neutral (charge zero) particles have no Coulomb force between them, but they may have magnetic interactions arising from electrical currents they contain. The electromagnetic interaction is responsible for the binding of electrons around atomic nuclei—*i.e.*, for atomic structure. Its laws thus provide a theoretical foundation for chemistry and biology.

Strong interaction. The strong interaction is responsible for the nuclear force that binds neutrons and protons together to form atomic nuclei, for nuclear scattering and reactions, and for creation and annihilation processes occurring in the collisions of high-energy particles. As was stated above, in connection with Yukawa's contribution to an understanding of strong interactions, a typical contribution to the force between two nucleons is $g^2/r^2 \exp(-r/[\hbar/mc])$, in which g is a meson-nucleon coupling constant and $\hbar/mc$ is the Compton wavelength of that meson.

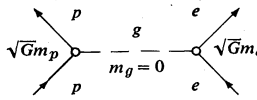
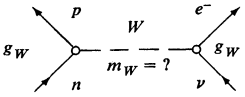
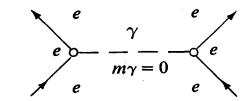
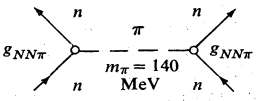
Comparing the forces

A standard elementary way to compare the gravitational, electromagnetic, and strong forces is to compare Gm_em_p/r^2, e^2/r^2, and $g^2_{NN\pi}/r^2 \exp[-r/\hbar/m_\pi c]$ for values of r sufficiently small that the exponential factor in the last expression is approximately one, or unity. Here m_e is the electron mass and m_p is the proton mass; $g_{NN\pi}$ is the coupling constant for the nucleon–pion interaction as shown in row four of Table 3. This is equivalent to comparing the three dimensionless coefficients shown in Table 3:

$$\frac{Gm_em_p}{\hbar c} = 3 \times 10^{-42} \qquad \frac{e^2}{\hbar c} = \frac{1}{137} \qquad \frac{g^2_{NN\pi}}{\hbar c} \approx 10.$$

For the weak force the comparison is more involved. If it arises from the exchange of a virtual "intermediate" (so far undiscovered!) W-meson as indicated in row two of Table 3, and, if the coupling constant (the analogue of e for electric forces and $g_{NN\pi}$ for strong forces) is g_W, then the analogue of the above three dimensionless numbers is $g_W{}^2/\hbar c$. What experimentation has determined so far, however, is the so-called Fermi constant G_F given by the expression $G_F = g_W{}^2 (\hbar/m_Wc)^2 \sqrt{2} = 1.42 \times 10^{-49}$ erg-centimetre^3, in which m_W is the mass of the hypothetical W-meson. In the absence of knowledge of this mass, one cannot determine the desired number $g_W{}^2/\hbar c$, from Fer-

Table 3: Strength and Range of the Four Basic Forces

force and Feynman diagram explaining the force via exchange of virtual quanta	force law	dimensionless force coefficient	typical reaction and mean life
Gravity 	$F = Gm_p m_e \frac{1}{r^2}$	$\frac{Gm_p m_e}{\hbar c} = 3 \times 10^{-42}$	—
Weak 	$F = g^2_W \frac{e^{-r/\lambda_W}}{r^2}$	unmeasurable*	$\Lambda \rightarrow p\pi^-$ 2.5×10^{-10} sec
Electromagnetic 	$F = e^2 \frac{1}{r^2}$	$\frac{e^2}{\hbar c} = \frac{1}{137}$	$\Sigma^0 \rightarrow \Lambda\gamma$ 10^{-19} sec
Strong 	$F = g^2_{NN\pi} \frac{e^{-r/\lambda_\pi}}{r_2}$	$\frac{g^2_{NN\pi}}{\hbar c} \approx 10$	$\pi p \rightarrow \Lambda K$ $\sim 10^{-23}$ sec

*The weak force is unmeasurable outside a range $\lambda_W = \frac{\hbar}{m_W c} < 10^{-14}$ centimetre. From neutron decay, $n \rightarrow pe^-\bar{\nu}$ (explained by same diagram, but with ν *out*going), one measures instead $\sqrt{2}\, g_W^2 \left(\frac{\hbar}{m_W c}\right)^2 \equiv G_F = 10^{-5} \frac{\hbar^3}{m_p^2 c}$.

mi's constant. The standard way of comparing weak forces to other forces is to compute the dimensionless quantity $G_F/\hbar c(m_p c/\hbar)^2 = 1.02 \times 10^{-5}$. This is indeed much smaller than the fine-structure constant $e^2/\hbar c$ and much larger than the gravitational constant $Gm_e m_p/\hbar c$, both shown above (for further discussion of the intermediate meson, see below *Undiscovered particles*).

Four classes of particles. The classification of interactions by strength is associated with a classification of particles by the forces that influence them. Particles that are directly acted upon by the strong interactions are called hadrons or strongly interacting particles; those that are acted on by the three weaker interactions, the nonhadrons, are called the leptons, the graviton, and the photon. Most particles react to several forces (see Table 4). The strongest force they experience usually controls

Table 4: Relation of Forces to Particles

particle	forces that the particle experiences			
	gravity	weak interaction	electromagnetism	strong interaction
Graviton g	X			
Photon γ	X		X	
Lepton ν	X	X		
Leptons μ^-, e^-	X	X	X	
Hadrons	X	X	X	X

their behaviour, but weaker forces may govern their decay. Furthermore, even electrically neutral hadrons respond to electromagnetic forces because the hadrons interact so very strongly that pure, untarnished hadrons are not found in nature, rather each hadron is partially a mixture of other hadrons. Thus, the neutral neutron behaves part of the time as if it were a mixture of a proton (of positive electric charge) and a π^- meson.

Hadrons. Hadrons, the strongly interacting particles, are numerous (see below *Spectroscopy of the hadrons*). Here, only the classification according to baryon number will be introduced. The baryon number, B, is an integer, positive, negative, or zero, assigned to each hadron. The particles of baryon number 0, ± 1, have special names: those of baryon number 0 are mesons, those of baryon number 1 are baryons, and those of baryon number −1, antibaryons. Particles of baryon number 1 and larger include ordinary atoms. For example, hydrogen has baryon number 1, deuterium baryon number 2, helium baryon number 4, etc. For such atoms the baryon number coincides with the mass number used in chemistry and nuclear physics—*i.e.*, the baryon number is the number of nucleons in the nucleus. Anti-atoms (composed of antinuclei with positron clouds surrounding them) then have negative baryon numbers.

The baryon number

The nonhadronic particles, those not subject to strong interactions, deserve special remarks:

Gravitons. The long-range gravitational interaction between two masses may be thought of as arising from the exchange of particles (see the top sketch of Table 3). The exchanged particles are called gravitons; they are massless and have spin 2 in units of $\hbar$. Individual gravitons are undetectable (see below *Undiscovered particles*).

Photon. The photon plays the same role for electromagnetism as the graviton does for gravitation; its exchange is responsible for electromagnetic forces. The photon is massless and has spin 1 in units of $\hbar$.

Leptons. The only other massless particles besides the graviton and photon are the neutrinos, ν_e and ν_μ, both of which have spin ½ in units of $\hbar$ and negative helicity; *i.e.*, their spin is pointed opposite to their velocity; the antineutrinos, $\bar{\nu}_e$ and $\bar{\nu}_\mu$, are also of spin ½ but have positive helicity. All known massless particles are electrically neutral and move with speed c.

The remaining leptons are the negative electrons, e^-, and muon, μ^-, and the antiparticles, e^+ and μ^+, respectively. All have spin ½, and, as will be shown in the section that follows, there are two conservation laws linking the leptons.

CONSERVATION LAWS AND QUANTUM NUMBERS

Particles are classified by their quantum numbers. It is the purpose of this section to list the 11 most important quantum numbers (see Table 5) and to describe how they arise from conservation laws (see CONSERVATION LAWS AND SYMMETRY).

Some conservation laws are exact; some are approximate. The typical form of an exact conservation law is an assertion that some quantity does not change, whatever interactions are taking place (*e.g.*, in classical mechanics, mass and energy in any closed system do not change). For an approximate conservation law the corresponding assertion is that the quantity is conserved in, say, strong

Table 5: Conservation Laws

conservation law	quantum numbers and observed values*	holds for				example		
		strong	electro-magnetic	weak	gravity†	p	π^0	e
1. Four-momentum	mass, not quantized: unit of mass is MeV	X	X	X	†	938	139	0.51
2. Angular momentum	$J = 0, \frac{1}{2}, 1, \frac{3}{2}, 2, \ldots$ [$\hbar$]	X	X	X	†	$\frac{1}{2}$	0	$\frac{1}{2}$
3. *CPT*	none	X	X	X	†			
4. Charge	$Q = 0, \pm 1, 2, \ldots$ [e]	X	X	X	†	+1	0	−1
5. Baryon number	$B = 0, \pm 1$	X	X	X	†	+1	0	0
6. Lepton number	$l_e = \pm 1, l_\mu = \pm 1$	X	X	X	†	0	0	$l_e = 1$
7. Charge conjugation	$C = \pm 1$	X	X		†	−1	+1	−1
8. Parity	$P = \pm 1$	X	X		†	+1	−1	+1
9. Time inversion	no quantum number							
10. Isospin symmetry	$I = 0, \frac{1}{2}, 1, \frac{3}{2}$	X			†	$\frac{1}{2}$	1	‡
11. $SU(3)$ symmetry	$Y = -2, -1, 0, +1$	X§			†	1	0	‡

*The list of the observed values is restricted to "elementary" particles; nuclei and antinuclei are excluded. †The gravitational force introduces no new explicit violations of conservation laws; it reflects the symmetries and absences thereof occurring in the other three interactions. ‡Inapplicable. §Approximately.

interactions but not in weak and electromagnetic. Of particular importance are the additive conserved quantities. Such a quantity takes a definite value for each particle, and its value for a system of particles is the sum of the values for the individual particles.

Meaning of quantum number. The phrase quantum number requires explanation. It originated in the Bohr theory of the atom, in which certain quantities must take only integral values, whereas according to classical mechanics, they could have any values. Most of the quantum numbers that are dealt with in this article take integral or half-odd integral values. For exceptional cases such as mass, energy, or momentum, which can take a continuum of values, it is customary to use the same terminology; they are also referred to as quantum numbers.

The list of exact conservation laws starts with those associated with space-time symmetries.

Energy and linear momentum. In the absence of outside forces, the energy, E, and the linear momentum, p (mass times velocity; vectors—quantities having magnitude and direction—are designated in this article by italicized boldface characters), of a system are conserved. This conservation law can be shown to follow from the invariance of the dynamical laws describing the evolution of the system under translation (motion) in time and space, respectively. Two observers in relative motion with a fixed velocity with respect to one another will attribute different energy and different momentum to a given system, but the connection between these descriptions is given by the special theory of relativity. In particular, that theory says that the rest mass, m, of the system is the same for all observers; *i.e.*, the rest mass is an invariant. For brevity, the adjective "rest" is usually omitted, and, for convenience, mass is measured in units of energy (ergs, joules, but, most frequently, electron volts). When a system is at rest with respect to an observer, momentum p is zero, and its energy is equal to its mass times the velocity of light squared ($E = mc^2$).

If a subatomic particle is completely stable (as are the electron and the proton, for example), then it has a precisely determined mass. Particles that decay last only a finite length of time and therefore (by a fundamental law, the so-called uncertainty relation) cannot have a definite energy and mass. Instead, they are characterized by an "average mass" and a mass breadth Δm related to their mean life at rest, τ, by the equation $\Delta mc^2 = \hbar/\tau$. Here c is the velocity of light and $\hbar$ is Planck's constant divided by 2π. The quantity mean life is introduced because an unstable particle does not always survive for the same time. The mean life is an average time of survival.

Angular momentum, spin, and helicity. In the absence of outside torques (twists), the angular momentum, J, of a system is conserved. The law of conservation of angular momentum is a consequence of the invariance under rotations of the dynamical law of evolution of the system. Associated with the angular momentum, J, is a quantum number, J, the total angular momentum in units of $\hbar$, taking only integer and half-odd integer values: $J = 0, 1/2, 1, 3/2, \ldots$.

Particles themselves have an intrinsic angular momentum, called spin, but massless particles (γ, ν, and graviton) differ from particles with nonzero rest mass in the way that their spin behaves and even in the way that it is defined. To be specific, when an observer moves with the same speed as a massive particle, the particle appears to be at rest; then its spin is defined as its angular momentum, J, as viewed by that observer. Its component J_z along the direction of motion z is called its helicity. This component ranges in integer steps from $+J$ to $-J$; *i.e.*, it can take on the $2J + 1$ different helicity values J, $J - 1$, $J - 2$, . . . $- J$. Examples are given, as heavy dots, in Figure 1. Many massive particles of low spin

Helicity

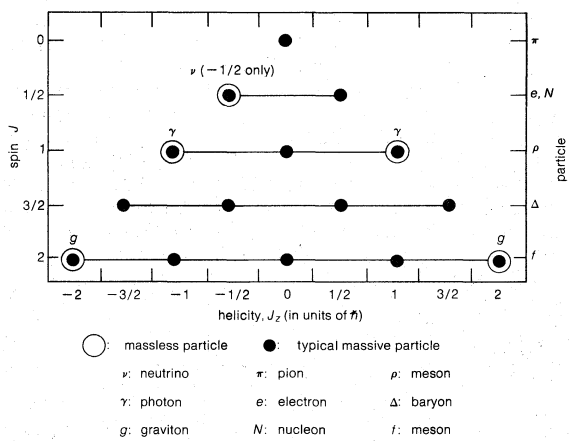


Figure 1: Allowed helicities for massless and typical massive particles (see text).

have been found; 0, ½, 1, 3/2, 2, 5/2, 3, for example, and it is likely many of the resonances the spins of which have not been measured have much higher spins than these. Massless particles are different. They move with the speed of light, and no observer can move fast enough to catch up with them. The helicity for a massless particle is defined directly as the component of the angular momentum (in units of $\hbar$) along the direction determined by the momentum. The possible values of the helicity for a massless particle are 0, ± ½, ± 1, . . . just as for a massive particle. If one of these values occurs, however, *no other need be possible.* This statement has to be qualified in a theory invariant under P the operation of space inversion. Then if s is a possible helicity, so is $-s$. In any case, the magnitude, $|s|$, of the helicity is called the spin. This situation is illustrated with open dots in Figure 1. The photon has spin 1 and helicities ± 1. The graviton has spin 2 and helicities ± 2. The neutrinos ν_e and ν_μ have spin ½ and negative helicity, whereas their antiparticles $\bar{\nu}_e$ and $\bar{\nu}_\mu$ also have spin ½ but positive helicity.

CPT symmetry and statistics. Two examples of conservation laws that do not fit precisely into the previous mold are those associated with *CPT* symmetry and with statistics. As already remarked (see above *History up to 1960*), *CPT* symmetry is a symmetry operation that exists and is compatible with any dynamical law having certain general properties. The only quantum number associated with *CPT* is $(CPT)^2$, which takes the values ± 1. It is $+1$ for states of integer angular momentum and -1 for states of half-odd integer angular momentum. The phrase "statistics of particles" refers to the symmetry properties of states involving several indistinguishable particles. In quantum mechanics every particle and every system composed of particles has a wave function that describes its state. Two laws of transformation of wave functions under interchange of particles occur for the particles observed in nature: the law of Fermi–Dirac statistics in which the wave function is antisymmetric (*i.e.*, changes sign under interchange of any pair of particles) and the law of Bose–Einstein in which it is symmetric (*i.e.*, is unchanged by the interchange of any pair of particles). The latter law is named after the Indian physicist Satyendra Nath Bose and Einstein. According to the Pauli exclusion principle, electrons satisfy Fermi–Dirac statistics. Any admissible dynamical law of evolution must have the property that it preserves the symmetry or antisymmetry of states of identical particles, *i.e.*, preserves their statistics.

It is a notable experimental fact that particles satisfying Fermi–Dirac statistics always have half-odd integer spin, whereas those satisfying Bose–Einstein statistics have integer spin. One of the most important general theorems of quantum field theory asserts that this law of the connection of spin with statistics is a consequence of general assumptions of quantum field theory. The customary terminology for particles, fermions for half-odd integer spin particles, and bosons for integer spin particles, takes the law of the connection of spin with statistics for granted. Since the possibility of a violation of this law seems remote, for the rest of this article it will be assumed valid. Thus, no quantum number describing the statistics of a particle need be listed in the Tables.

The connection of spin and statistics

The remaining exact laws of conservation seem to be connected not with space-time symmetries but rather with some internal characteristics of particles.

Electric charge. The electric charge, Q, is not only conserved; it is quantized. Every system has a charge that is a whole multiple of charge, ne, in which n is an integer (positive, negative, or zero). The charge Q is an additive quantum number, and so the electric charge of a system of particles is the sum of the charges on the individual particles. All the long-lived subatomic particles have charges 0 or $\pm e$, but, as will be shown, there are excited states of baryons that are doubly charged. The fact that most subatomic particles have charges 0 or $\pm e$ is, thus, happenstance.

Baryon number. The baryon number, B, is a conserved quantity similar to the charge in being quantized (taking integer values) and in being additive. As already remarked above, the number B is $+1$ for baryons, -1 for antibaryons, and 0 for mesons, leptons, photons, and gravitons. Baryon number B is called atomic mass number A in the areas of chemistry and nuclear physics. Conservation of baryon number is responsible for the stability of the proton, reactions such as the decay of a proton into a positron plus photon ($p \rightarrow e^+ + \gamma$) being forbidden according to the theory. Only in collisions of matter and antimatter can baryons annihilate, this energy going into the production of mesons; *e.g.*, $p + \bar{p} \rightarrow \pi^+ + \pi^-$.

CPT symmetry an aid in classification

The validity of *CPT* symmetry permits simplification of the classification of particles because it shows that antiparticles have characteristics deducible from those for particles. They must have the same mass, the same spin, opposite charge, opposite baryon number, opposite magnetic moment, and the same lifetime. Consequently in the tables that follow, the antiparticles of the baryons will not be separately listed.

It should be remarked that, although *CPT* symmetry is a generally accepted symmetry, experimental physicists are constantly searching for tests of its validity as well as that of other conservation laws. Such tests serve to strengthen support for the basic principles underlying present ideas about particles.

Lepton, electron, and muon numbers. The lepton number is an additive quantum number, l, assigned as follows:

$l = +1$ for ν_e, e^-, ν_μ, μ^-, (*i.e.*, all leptons);
$l = -1$ for ν_e, e^+, ν_μ, μ^+, (*i.e.*, all antileptons);
$l = \ \ 0$ for all other particles.

The electron number is an additive quantum number, l_e, assigned as follows:

$l_e = +1$ for ν_e, e^-;
$l_e = -1$ for ν_e, e^+;
$l_e = \ \ 0$ for all other particles.

The muon number, l_μ, is an additive quantum number, l_μ, assigned as follows:

$l_\mu = +1$ for ν_μ, μ^-;
$l_\mu = -1$ for ν_μ, μ^+;
$l_\mu = \ \ 0$ for all other particles.

Clearly, the lepton number for any particle is equal to the sum of the corresponding electron and muon numbers ($l = l_e + l_\mu$); so only two of the three quantum numbers are independent.

Conservation of muon number *or* electron number explains the absence of the decays $\mu^+ \rightarrow e^+ + \gamma$ or $\mu^+ \rightarrow e^+ + e^- + e^+$. Similarly, their conservation explains why the neutrinos ν_μ produced in π^+ decays ($\pi^+ \rightarrow \mu^+ + \nu_\mu$) are observed to interact with nucleons only via the interaction $\nu_\mu + n \rightarrow p + \mu^-$, but not $\nu_\mu + n \rightarrow p + e^-$.

The remaining conservation laws are approximate, being valid only for some subset of the four kinds of interactions.

Particle–antiparticle conjugation, C. This operation is often called charge conjugation, recalling its discovery in the theory of electrons in which it interchanges the positron and the negative electron, e^+ and e^-. That is to say, charge conjugation is a mathematical operator that transforms a state describing a particle into a state describing its antiparticle. In general, it interchanges particles and antiparticles without affecting space–time coordinates. The strong and electromagnetic interactions are invariant under charge conjugation, but the weak interaction is not. This implies that, to the extent that weak interactions may be neglected, the probability of a reaction remains the same if all particles are replaced by their corresponding antiparticles. Thus, e^+e^+ scattering (the deflections of one beam of positrons by another) has the same probability as e^-e^- scattering. Under charge conjugation, baryons go into antibaryons and leptons into antileptons, but mesons and photons go into mesons and photons. For example, charge conjugation interchanges π^+ and π^- mesons. Neutral particles may be their own antiparticles (as in the case of π^0 and γ) or not (as in the case of K^0 and $\bar{K}^0$). When a neutral particle is its own antiparticle, operation on it by charge conjugation amounts to multiplication by a plus or minus sign called the charge conjugation, or C, quantum number. For example, C, operating on a gamma-ray state changes its sign and a γ is said to be "odd" under C; a π^0 is "even" under C.

Space and time inversion, P and T. The operation of space inversion, P, acts on space–time coordinates by leaving t alone, $t \rightarrow t$, and transforming x into its negative, $x \rightarrow -x$. There is an associated quantum number, the parity, which is equal to $+1$ for a wave function that is unchanged under P inversion, and equal to -1 for a wave function that is multiplied by -1. States with these transformation laws are said to be of even and odd parity respectively. Strong and electromagnetic interactions are invariant under space inversion, but the weak interaction is not. The parity revolution of 1956 resulted from the realization that, in neutron decay and muon decay, the neutrinos and antineutrinos appearing experimentally have definite helicity, in violation of the law of conservation of parity. The parity, symbolized as P, is a multiplicative quantum number; the parity of a composite sys-

Parity violation

tem is determined by multiplying the intrinsic parities of the particles with the parity of any orbital angular momenta the system may have.

The result of the parity revolution, that neither space inversion, *P*, nor charge conjugation, C, is an exact symmetry of the weak interaction, led to a quantitative theory in which, nevertheless, the combined operation, symbolized as *CP*, is an exact symmetry. (This phenomenon can be thought of as being analogous to the conservation of baryon number $n = n(B) - n(\bar{B})$. In spite of the fact that the number of baryons $n(B)$ can change with time and so can the number of antibaryons $n(\bar{B})$, their difference is conserved.) A few years after the parity revolution, however, it was found that the long-lived decay mode, K_L^0, of the K^0 meson can decay into $\pi^+ + \pi^-$. This process violates *CP* symmetry, and hence the weak interaction has none of the operations *P*, *C*, or *CP* as an exact symmetry. The only discrete symmetry that appears to be exact is the combination of *CP* with time reversal *T*—that is, *CPT*.

Isospin (I) and hyperchange (Y). In Table 5 there are two more quantum numbers, *I* and *Y*, but they are defined only for hadrons and so are explained in the section below.

SPECTROSCOPY OF THE HADRONS; THE QUARK MODEL

Classes of hadrons. It is notable that there is only a superficial difference between the "old" metastable hadrons of Table 1 and more recently discovered unstable "resonances."

In terms of stability there exist three classes of hadrons:

Unstable "resonances." An example of an unstable resonance is the omega (ω) meson. It has a mass of 784 million electron volts (MeV) and decays into three pions (3π, each of mass 140 MeV) with a surplus energy of nearly 400 MeV. If the ω meson were not a resonance, it would fall apart immediately in $\frac{1}{3} \times 10^{-23}$ second (this is the time needed for a pion, moving with nearly the speed of light to move 10^{-13} centimetre and thus escape beyond the range of nuclear forces). But the omega meson *is* a resonance, the name given to long-lived states, and it actually holds together for a surprisingly long time: 10^{-22} second.

Metastable hadrons. Metastable hadrons eventually decay via weaker interactions. If the omega meson weighed less than three pions it would be unable to decay by the strong interaction and thus would be called a "metastable" hadron. The ω meson would eventually decay via electromagnetism into a pion and a photon (in 10^{-21} second) or by weak interaction into two leptons (in 10^{-10} second).

Stable hadrons. There is only one absolutely stable hadron, the proton. The examples above should show that stability is really an accidental property (depending on the availability of final states), although, of course, it is certainly important to the experimenter who has to deal with the particle. In what follows below, metastable and resonant hadrons will seldom be distinguished from each other.

Charge multiplets and isospin. Figure 2 shows the known stable and unstable particles and as many resonances as fit into the picture. As may be noted, hadrons occur in "charge multiplets" of one to four particles, each with the same mass. In fact, it turns out that hadrons share all properties except the electric charge, which seems to have no role in the strong interaction (it is as though it had been sprayed on superficially). Thus the neutron and the proton—the members of the $N(939)$ "nucleon" doublet (see below)—share practically the same mass and nuclear interactions and the same spin and parity. This explains why it is possible to give a family name (N, π, . . .) to a multiplet and to indicate the charge of one of its members with a separate superscript; *e.g.*, (π^+, π^0, π^-).

Although the nuclear force barely distinguishes between differently charged members of a multiplet, it is much concerned with the total size of the family; *i.e.*, the total "multiplicity" (*M*) is a most significant quantum number. To give an example, there are some reactions for

Total "multiplicity"

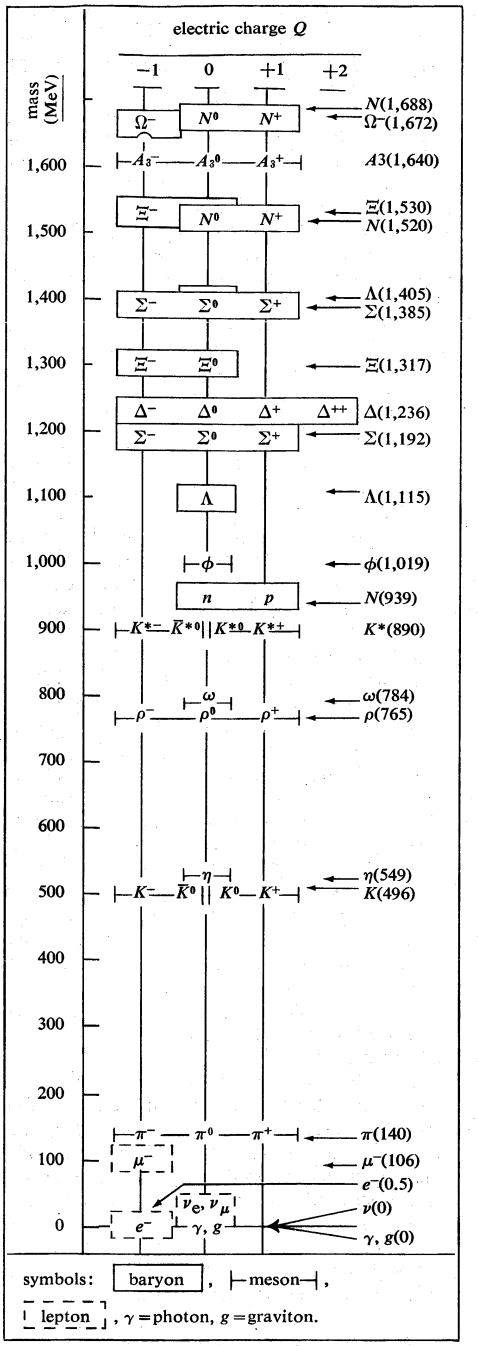


Figure 2: Metastable particles and some resonances with mass up to 1,700 MeV. The horizontal lines and bars link members of the same charge multiplet. Among the hadrons (mesons and baryons), no distinction is made between metastable states and resonances. Details of all the metastable particles are given in Table 1.

which hadrons with multiplicity equal to zero ($M = 0$) are forbidden to participate, while there are also others for which a multiplicity equal to one or greater than one ($M \geqq 1$) is forbidden, etc.

Often, for convenience, multiplicity is translated into a closely related vector quantity, *I*, according to the relation $M = 2I + 1$.

Experimentally it is found that the quantity *I* is conserved in the strong interaction. The multiplets of hadrons suggest the multiplets of states arising from the existence of angular momentum and the formalism of isospin is based precisely on this analogy. One introduces an isospin vector operator, *I*, analogous to the angular momentum operator, *J*, but acting on some internal coordi-

nates of the hadron rather than on its space coordinates. In this hidden internal coordinate system, electric charge increases along a particular axis conventionally labelled the z-axis, and is equal to the sum of two terms: $Q = \langle Q \rangle + I_z$, in which the second term, I_z, can take on $2I + 1$ different values, and the first term, the average charge $\langle Q \rangle$ (later called $Y/2$), is discussed in the following section on hypercharge Y.

Strangeness, or hypercharge. The baryon multiplets plotted in Figure 2 are not all centred symmetrically about the charge = 0 axis. The last quantum number Y, the hypercharge, is a measure of their displacement. To define this quantum number precisely, the centre of each multiplet is defined by its average charge $\langle Q \rangle$. This is literally the value of charge Q at which each of the multiplets drawn in Figure 2 can be balanced. Thus, for the box in Figure 2 containing n and p, (the nucleon doublet) $\langle Q \rangle$ is $+ \frac{1}{2}$, and, for the pion triplet, $\langle Q \rangle$ is zero. To avoid dealing with half-integral values, the hypercharge Y is defined as twice the average charge, $2\langle Q \rangle$. The strong interaction conserves hypercharge. This completes the list of conservation laws (see Table 5).

The hypercharge can be related to an older measure of the displacement called strangeness, S, which is twice the displacement of some multiplet measured not from $Q = 0$ but from the most familiar multiplet (π for mesons, N for baryons). For mesons, in which the reference pion has $Y = 0$, Y and S are the same; but, for the baryons, in which the reference nucleon has $Y = 1$, strangeness S is $(Y - 1)$. As mentioned above, the first strange particles were discovered in 1947. They were called "strange" because they did not decay by the strong interaction, and at the time that seemed strange. Accordingly, physicists Murray Gell-Mann and Abraham Pais, in the United States, and Kazuhiko Nishijima, in Japan, introduced a law of conservation of strangeness (now called conservation of Y). Thus, the K^+ meson would "normally" be expected to decay strongly in 10^{-23} second into π^+ and π^0, but the meson has $S = Y = 1$, and the pions have $S = Y = 0$; and so the reaction is forbidden to the strong interaction. The weak interaction finally leads to this and other decays, but it takes 10^{-10} second, or 10^{13} times "too long."

Conservation of strangeness

Multiplets and super-multiplets. The symbol (N, π, . . .) was used above to describe a multiplet, and an explanation of this nomenclature is in order: when a resonance is first discovered, the most apparent quantum numbers are B (baryon number), Y (hypercharge), and I (isospin) and the superficial one Q (charge); hence a symbol is invented that represents the (B,Y,I) for the particle, and the superscript Q is appended (*e.g.*, N^+ for proton), and this is followed by another symbol (m, J^P), representing (mass, angular momentum parity), so that the full specification for the proton is, technically speaking, $N^+(939, \frac{1}{2}^+)$. Sometimes an asterisk is used to show that a multiplet is an "excited state" of a multiplet of lower mass. Thus, there is the $K(496)$ meson and the next one found, $K^*(890)$, is often called simply the K^*.

In January 1961, Gell-Mann and physicist Yuval Ne'eman of Israel independently proposed an "eightfold way" [$SU(3)$ invariance] to explain the jumble of hadron' multiplets. This explanation starts with the eight stable baryons that were known at the end of 1960. In Figure 3A each of these eight stable baryons has been plotted as a black dot. The horizontal-scale is almost the same as that of Figure 2: charge Q still increases to the right, but the displacement of the multiplets has been corrected by plotting charge *minus* average charge, $Q - \langle Q \rangle$. Thus, the top entry is the N doublet, with the neutron (n) at the left and proton (p) at the right. To describe each particle there remain six quantum numbers and the mass. Many physicists had tried plotting many different combinations, but the most satisfying is to collect hadrons with the same J and P as is done here, and choose hypercharge Y for the vertical axis; a striking hexagonal snowflake configuration then emerges, with six corner dots and two more at the centre. Thus the first complete supermultiplet, an octet, those baryons with $J^P = \frac{1}{2}^+$, has been plotted. Unfortunately this plot has no way to display mass.

Ne'eman's "eightfold way"

If the same process is tried with the seven stable mesons that were known at the beginning of 1961, another hexagon emerges—the one shown in Figure 3B; these mesons all have zero angular momentum and negative parity ($J^P = 0^-$). In 1961 only one more meson doublet, the K^*, was known, but its J^P were unknown, and so no more octets could be assembled.

Returning to the baryons, by 1961 two unstable baryon multiplets had been discovered. One of them, the Δ (1,236, $3/2^+$), had been known since the '50s. It is a quartet, appearing as Δ^-, Δ^0, Δ^+, and Δ^{++}; so it is clear that it cannot fit into a hexagon, which seats at most three abreast. It has $Y = 1$ and is plotted at the top of Figure 3D. Another baryon doublet had just been discovered and named Y_1^* (1,385). It is now called $\Sigma(1,385)$, and, since it has $Y = 0$, it is plotted below the multiplet $\Delta(1,236)$. This was the last clue necessary for the formulation of the eightfold way.

The quark model. Figure 3 strongly suggests that hadron multiplets occur in groups (called supermultiplets) sharing the same J^P. These regular patterns in turn suggest that hadrons must be composed of even more primitive elements fitted together in an orderly way. They were named "quarks" by Gell-Mann and are drawn in Figure 3 as a triangular array of three asterisks (perhaps question marks would have been better). Despite many attempts to find them, they have not shown up, but they are essential for organizing knowledge; hence particle physicists all speak of the "quark model" of hadrons.

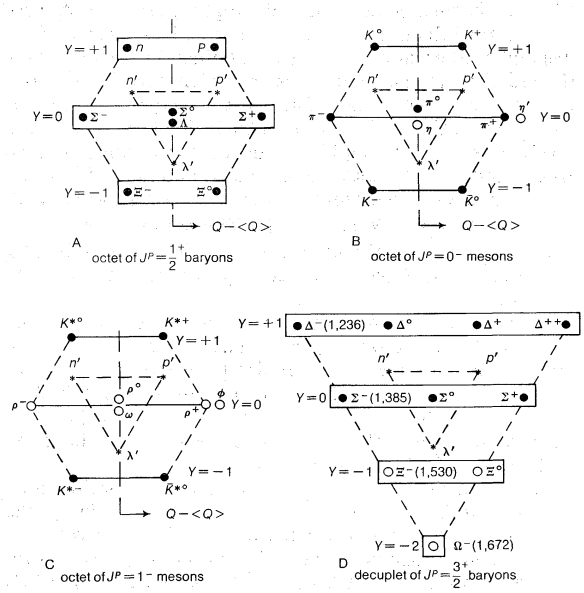


Figure 3: Supermultiplets of hadrons—octets and decuplets. The horizontal lines link members of the same charge multiplets in a fashion similar to that of Figure 2; solid dots represent particles that were known when the eightfold way was formulated in 1961; open dots are particles predicted and later discovered; asterisks labelled n', p', and λ' are the possible set of primitive particles called quarks; Q stands for electrical charge; $\langle Q \rangle$ indicates average charge of the multiplet taken as a whole.

The quark model is thus oversimplified.

To clarify the Periodic Table of the chemical elements, atoms are explained as being composed of more primitive units called nuclei and electrons, and, just as the electrons group into shells of 2, 8, 18, 32, . . . , similarly the quark model can be used to explain the supermultiplets of hadrons: 1's, 8's, and 10's. Historically the eightfold way had a more mathematical formulation, involving group theory, and the emphasis on the physical role of the quark (by Gell-Mann and another United States physicist, George Zweig) came later. The nonmathematical presentation used here is thus oversimplified.

Three quarks (q) are postulated, (but because all particles have antiparticles, the existence of three antiquarks [$\bar{q}$] is implied). The quarks are assigned baryon number

B = ⅓ ($\bar{q}$ then has $B = -⅓$). Mesons can then be considered as $q\bar{q}$ pairs bound together by the strong interaction, and baryons as bound states of three quarks (qqq). Returning to mesons, since there are three quarks (called n', p', and λ'—see Figure 4) and three antiquarks,

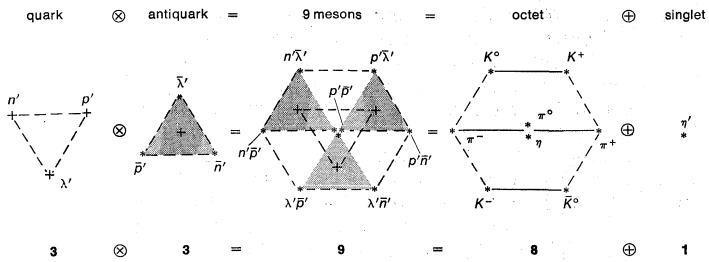


Figure 4: Combining quarks. Three hypothetical quarks and three antiquarks combine to form nine mesons, which group into an octet plus a singlet. Two such real octets + singlets are shown in Figure 3B and 3C.

then $q\bar{q}$ might be expected to combine into $\mathbf{3} \otimes \bar{\mathbf{3}} = \mathbf{9}$ different mesons ($n'\bar{n}'$, $n'\bar{p}$, $n'\bar{\lambda}'$, $p'\bar{n}$, . . . $\lambda'\bar{n}'$ . . . $\lambda'\bar{\lambda}'$). But the hexagons contain only eight particles! A way out of this dilemma came from a 19th-century Norwegian mathematician, Marius Sophus Lie, who investigated the symmetrical ways in which groups of basic objects might combine. One of the ways—the special Lie group called $SU(3)$—breaks $\mathbf{3} \otimes \mathbf{3}$ up into an **8** (hexagons) and a singlet. Some of these singlets have since been found—they are called η' and ϕ' in Figure 3.

The Lie group $SU(3)$

Mathematicians summarize the discussion above with equation $\mathbf{3} \otimes \bar{\mathbf{3}} = \mathbf{9} = \mathbf{8} \oplus \mathbf{1}$. (The symbols ⊕ and ⊗ come from group theory.) Figure 4 shows graphically how the **3** and $\bar{\mathbf{3}}$ triangles actually combine to generate a hexagon.

It was stated above that *baryons* are postulated to be formed of three quarks (qqq). The computation of the equation $\mathbf{3} \otimes \mathbf{3} \otimes \mathbf{3} = \mathbf{27}$ remains to be explained; group theory "algebra" of $SU(3)$ gives the answer as $\mathbf{3} \otimes \mathbf{3} \otimes \mathbf{3} = \mathbf{27} = \mathbf{1} \oplus \mathbf{8} \oplus \mathbf{8} \oplus \mathbf{10}$. Furthermore, $SU(3)$ predicts correctly that the **8** will again look like the hexagon and the **10** will look like the triangle of Figure 3D.

While theorists were making predictions, experimenters working with hydrogen bubble chambers (notably the Berkeley, California group of Luis W. Alvarez) were discovering new particles. The stable meson octet was completed with the discovery of the eta (η) particle in 1961. In the same year the rho (ρ) and omega (ω) mesons were discovered and found to have $J^P = 1^-$; and in 1962 the spin of the K^* was found to be 1^- also; this completed the 1^- octet. The first baryon "decuplet" was not completed until 1964, when the Ω^- was rediscovered. (A single probable Ω^- had been observed in 1954, but its discovery had attracted little attention.) The twin discoveries of η and Ω were triumphant vindications of $SU(3)$.

The 1973 particle data tables (see bibliography) list approximately 25 meson multiplets (in contrast to seven in Figure 2) and 50 baryon multiplets (ten are in Figure 2), and many of these have already been fitted into supermultiplets. By now, quarks, as concepts, are well established. Not so the real, free quarks. Physicists are always looking for them, but so far none has been established, despite perennial rumours of their discovery (see below *Undiscovered particles*).

In the presentation of the quark model here, geometrical arrays have been emphasized; but actually this is but a small part of the predictive power of the model, which relates the masses of all those multiplets within a supermultiplet and also their decay rates into various other hadrons, and even their electromagnetic decays.

Without entering into the technical details of the quark model or the eightfold way, the main features of these theories may be sketched. There are two essential ideas involved, the first of which is that the fundamental strong interaction consists of a dominant part that is very symmetrical and a weaker part that is less symmetrical. If the dominant part alone were present, all members of the same supermultiplet would have the same mass. For example, the members of the octet consisting of $N(939)$, $\Lambda(1{,}115)$, $\Sigma(1{,}192)$, and $\Xi(1{,}317)$ in Figure 2 would coincide in mass. The weaker part of the strong interaction is less symmetrical, and its lack of symmetry results in the splitting of the supermultiplet into the observed mass spectrum. (The quark model and eightfold way have this idea in common with half a dozen previous theories that failed to account for the facts.) The second essential idea is that the symmetry of the interaction is expressed in a specific way; there is a family of transformations that leave the dominant part of the strong interaction invariant and form a group called $SU(3)$. (See TOPOLOGICAL GROUPS AND DIFFERENTIAL TOPOLOGY and CONSERVATION LAWS AND SYMMETRY.) The $SU(3)$ group consists of all 3×3 unitary matrices of determinant 1. The weaker part of the strong interaction has a specific transformation law under $SU(3)$, which determines how the supermultiplets split into multiplets.

The dominant and weak parts of strong interaction

EXPERIMENTS

The article so far has dealt mainly with history, classification schemes, and theoretical explanations. But on a world-wide basis, several hundred million dollars are spent each year on active, experimental programs. This section, therefore, is devoted to some pictures of interacting subatomic particles and a plot of their cross section versus energy.

Formation of neutron and proton resonances. In some 1952 experiments, pions from the University of Chicago cyclotron were directed at protons (in liquid hydrogen), and the "scattering cross section" was measured for different energies of the pion beam. Scattering refers to the change of direction when two particles collide; scattering cross section is the probability that scattering will occur. When the probability is large, the two particles act as if they were big, with a large cross section.

Scattering experiments

The "effective" mass of the pion–proton system is calculated for each setting of the pion beam. (The effective mass is the sum of the rest masses plus the kinetic energies of all the particles in a system, as viewed from the system's centre of mass.) When the effective mass is plotted along the x-coordinate of a graph and the scattering cross section is plotted along y, it is found that the cross section peaks sharply when the system has an effective mass of about 1,236 million electron volts (see Figure 5).

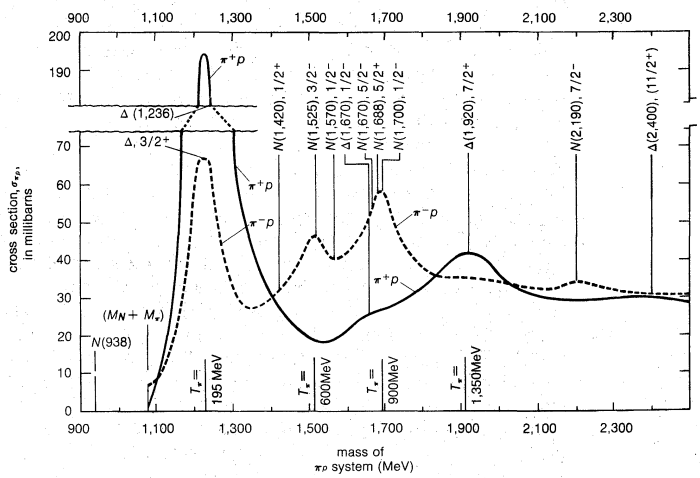


Figure 5: Total cross section for π^+ and π^- mesons incident on protons. Bumps in cross section represent formation of resonances.

Figure 5 shows cross-section plots for both π^+p (solid curve) and π^-p (dashed). Both cross sections show the peak at 1,236 MeV, although it is threefold higher for π^+ than for π^-. Such a peak is the signature of a resonance. This one is called $\Delta(1{,}236)$; it is a charge quartet (see Figure 2), meaning that it can be formed in four charge states: Δ^{++}, Δ^+, Δ^0, Δ^-. By shooting both π^+ and π^- beams into hydrogen, two of these states can be formed: $\pi^+p \rightarrow \Delta^{++} \rightarrow \pi^+p$ and $\pi^-p \rightarrow \Delta^0 \rightarrow \pi^-p$. Figure 5 shows

many resonances at higher mass that are not as prominent as the $\Delta(1,236)$ and that cannot be detected merely by hunting bumps in the cross-section plots. It is then necessary to analyze the angular distribution of the scattered pions, and it may even be necessary to "polarize" the protons (*i.e.*, align their spins with their direction of motion) in the target hydrogen and to look for resulting asymmetries in the angular distributions.

Production of the $\Sigma(1,385)$ resonance. Next to be examined is how the experimentally more remote resonance $\Sigma(1,385)$ is studied. This resonance has two decay modes: $\Sigma(1,385) \rightarrow \Lambda\pi$ and $\Sigma\pi$. The direct approach would be to study a bump in the $\Lambda\pi$ scattering curve at mass 1,385 MeV, but a Λ decays via the weak interaction in 10^{-10} second, and a π decays in 10^{-8} second; so a Λ beam on a π target is impractical.

The best that can be done is to study a reaction such as $K^-p \rightarrow \Lambda\pi^+\pi^-$, in which $\Lambda\pi$ "diparticles" are produced over a large region of $\Lambda\pi$ mass, and see whether or not some masses in particular are produced copiously. Figures 6 and 7 show several steps in a hydrogen-bubble-chamber experiment that yields $\Sigma(1,385)$.

Principle of a bubble chamber

Figure 6 shows a hydrogen chamber exposed to a beam of K-mesons. A bubble chamber consists principally of a volume of superheated, but not yet boiling, liquid hydrogen. Any charged particle, passing through the liquid,

By courtesy of Lawrence Radiation Laboratory

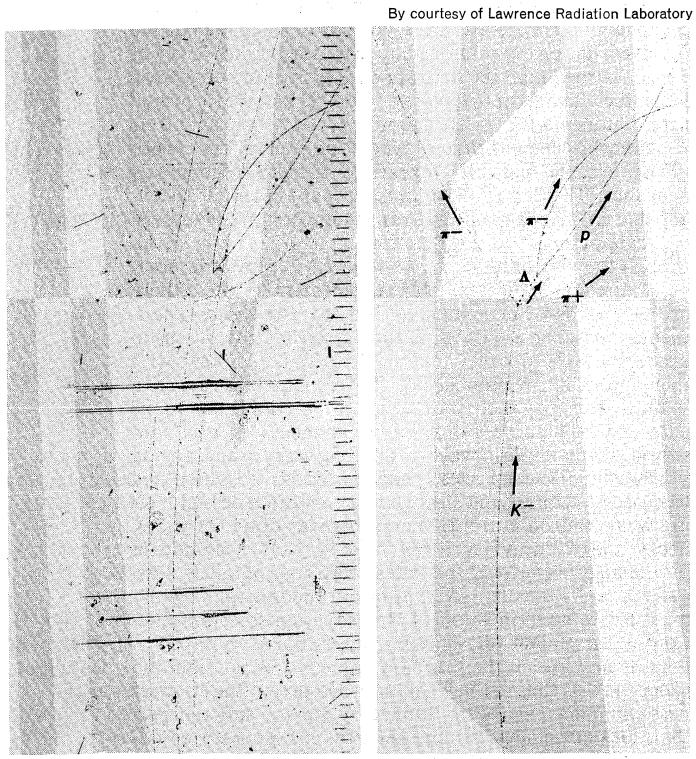


Figure 6: (Left) Photograph of a K^- meson interacting in a hydrogen bubble chamber. The interaction at the centre is shown and labelled at right.

agitates the molecules and heats the liquid. Boiling starts at these hot spots, and the path is recorded as a string of tiny bubbles that can be photographed. The chamber is in a magnetic field that bends the charged particles. The curvature of the tracks then tells the speed of the particle (see RADIATION DETECTION AND CHARACTERIZATION).

Three K^- mesons enter the bottom of Figure 6 (left) and travel through the chamber. Two pass through uneventfully, while one hits a proton to produce the reaction $K^-p \rightarrow \Lambda\pi^+\pi^-$. In its brief life of $\sim 10^{-10}$ second, the neutral Λ flies about one centimetre, then decays into two charged particles, $\Lambda \rightarrow p\pi^-$. Thus the Λ is seen as a gap of one centimetre, ending in a V formed by the p and π^-. For clarity, Figure 6 (right) repeats the tracks of interest, separated from the rest and labelled. An automatic film reader then measures the lengths and positions of labelled tracks to an accuracy of a few thousandths of a centimetre, and their coordinates are transmitted to a computer. If the tracks have been measured using two (or three) stereoscopic (*i.e.*, photographed from different viewpoints) pictures, the computer can then reconstruct their position, direction, curvature, and velocity.

Once the momentum of the Λ and of each of the π's has been computed, the computer can easily compute the effective mass $M_{\Lambda\pi^+}$ of the $\Lambda\pi^+$ diparticle and the effective mass $M_{\Lambda\pi^-}$ of the other. Further, it can make a "scatter plot" of one mass versus the other. This is shown in Figure 7B, in which each of the 5,640 dots represents an entire $\Lambda\pi^+\pi^-$ event like the one just described. For technical reasons it is more convenient to plot not the mass M but its square, M^2. For this particular choice of coordinates, it turns out that, if Λ and π did not attract one another to form resonances, the dots would be uniformly scattered over the allowed region. Instead, an almost black vertical band is seen, representing a considerable excess of events, at a value of $M^2_{\Lambda\pi^+} = (1{,}385\ \text{MeV})^2$, equal to $1.9\ (\text{BeV})^2$. This shows that there is a $\Lambda\pi^+$ resonance called $\Sigma^+(1{,}385)$. It can be shown that, if there is such a $\Lambda\pi^+$ resonance, then invariance of the strong interaction to transformations in isospin space requires that there should exist another resonance in the minus charge state, $\Lambda\pi^-$ of the same mass; in other words, the scatter plot of charge should show both the Σ^+ and Σ^- members of the $\Sigma(1,385)$ triplet. Although $\Sigma^-(1,385)$ is produced less intensely, a horizontal excess of events is indeed seen near the bottom of the allowed region, near $M^2_{\Lambda\pi^-} = 1.9\ (\text{GeV})^2$. It is conventional to display the projections of the dots onto the horizontal and vertical axes as histograms (bar graphs). The $\Lambda\pi^+$ and $\Lambda\pi^-$ enhancements can then be seen as sharp bumps, even sharper than the $N\pi$ bumps that are seen in Figure 5.

The "scatter plot"

In looking along the $-45°$ line of the scatter plot (parallel to the long straight "skin" of the allowed region), yet another enhancement will be seen, the projection of which is shown as the third histogram in Figure 7B. This turns out to be the signature of a $\pi^+\pi^-$ resonance known as the ρ meson.

Figure 7B, with 5,640 events, was published in 1967, and Figure 7A, earlier, in 1960. Comparison of these figures shows the rapid increase in technical capability in the intervening seven years. Figure 7A is taken from a paper that announced the discovery of the $\Sigma(1,385)$ as the first strange particle resonance. It displays 141 events that were measured and plotted with relatively little automation. This number of events was sufficient to show the existence of a bump (resonance) at 1,385 million electron volts but was insufficient to permit determination of the spin or parity of the resonance or even to hint at the production of the ρ meson in the same reaction.

Discovery of first strange particle resonance

RELATIONS BETWEEN THE INTERACTIONS

Currents. In the section above *History up to 1960*, it has been pointed out that the proposal of Lee and Yang about the violation of the law of conservation of parity in K-meson decays led to the discovery of its violation in muon decay ($\mu^- \rightarrow e^- + \nu_e + \nu_\mu$) and nuclear beta decay ($n \rightarrow p + e^- + \nu_e$). Further experiments have led to a much deeper understanding of the weak interactions and their relations to strong and electromagnetic interactions. The starting point was the discovery in the late 1940s that meson decay, nuclear beta decay, and meson capture ($\mu^- + p \rightarrow n + \nu_\mu$) can be described by Fermi interactions with nearly the same basic rate constant (the Fermi constant, G_F, of Table 3). The near equality of these three Fermi constants led to speculations about the possibility of a universal Fermi interaction among spin ½ particles. The question of the detailed structure of the interaction was not settled at that time. Ten years later (1956–58), after a series of definitive experiments, it became possible to single out a specific type of Fermi interaction. In particular, it was shown that the massless particle appearing in the reaction $e^- + p \rightarrow n + \nu_e$, which by convention is called an e-neutrino (ν_e), has negative helicity (as shown in Figure 1). The e-antineutrino (ν_e) turns out to have positive helicity.

Universal Fermi interaction

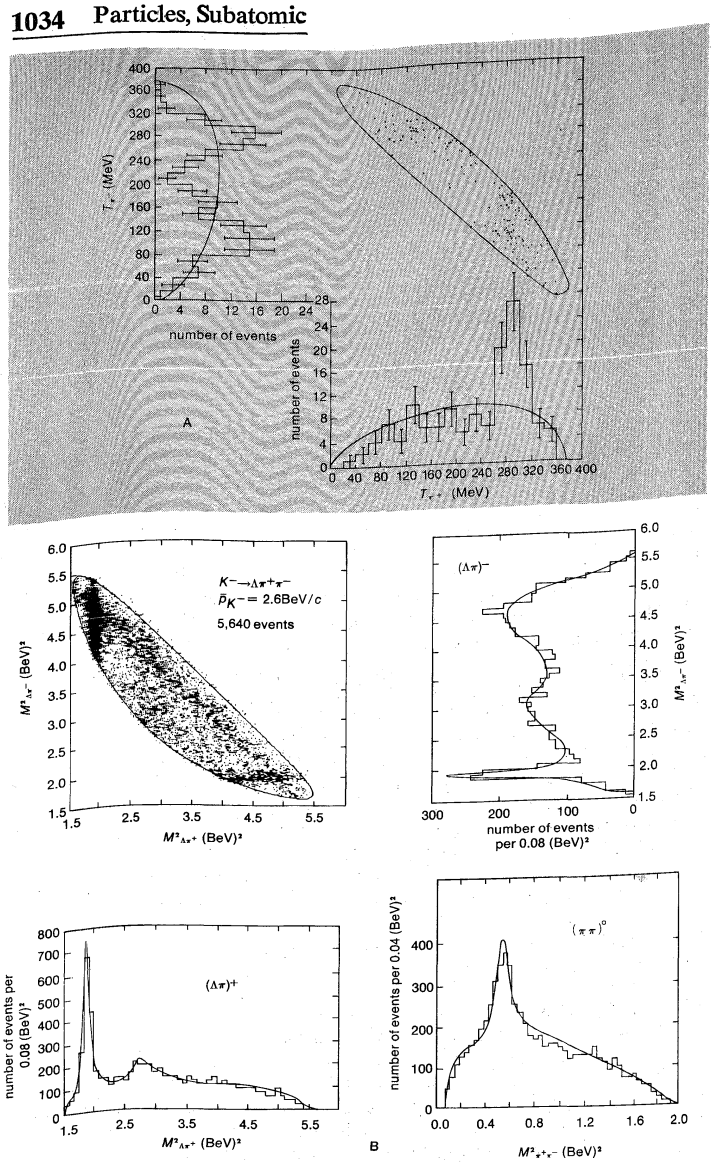


Figure 7: (A) Discovery, in 1960, of Σ(1,385), based on 141 events. This same reaction is pictured in (B) with 5,640 events.
By courtesy of (B) Lawrence Radiation Laboratory, Berkeley, California; (A) M. Alston, *et al.*, *Physical Review Letters*, vol. 5, no. 11, p. 520 (1960)

An important notion that played a considerable role in this clarification of the structure of the weak interactions is that of a conserved current. Current conservation is really a shorthand for conservation of charge. The reason for a dual notation of conserved current and conservation of charge is that for charges in motion the expression for the conservation law of electric charge in a small region involves both the density of charge and the density of current. It has become conventional to lump these quantities together and call them current (or charge-current). The notion of conserved current was first used to explain how it is that the strong interactions (which affect the n and p in the reaction $n \to p + e^- + \bar{\nu}_e$ but none of the particles in the decay $\mu^- \to e^- + \bar{\nu}_e + \nu_\mu$) do not make the rate of the two reactions different. The explanation given can be understood by analogy with the case of the electric charges of particles. That the charges of all elementary particles are integer multiples of the same basic charge (even though some of the particles are strongly and some only weakly interacting) is a consequence of the conservation law of electric charge. Similarly, if the e^- and $\bar{\nu}_e$ produced in the decay of the neutron are coupled to a conserved current, the near equality of the rates is at least in part accounted for.

In the detailed application of this idea to the weak interactions, it is necessary to introduce in addition to the electromagnetic current, a so-called weak current of the hadrons and a weak current of the leptons. These currents then interact, for example, via the exchange of a W-meson, an intermediate boson of the weak interactions. (This exchange is indicated in Table 3, in which the weak current of the nucleons is indicated at the left and that of the leptons at the right.)

The *CVC* hypothesis

Now it turns out that the weak hadron current splits into two parts, one of which, the so-called vector current, is conserved like the electromagnetic current. (This is the so-called conserved vector current, or *CVC*, hypothesis.) The other part of the current is called the axial vector current. There are certain nuclei whose beta radioactivity involves only the vector part of the weak hadron current. According to the *CVC* hypothesis, the rate of these decays can be compared directly with the rate of muon decay, since there are no corrections to either rate attributable to the strong interactions. The comparisons of the late 1950s gave fairly good agreement under the assumption of the equality of the Fermi constants. On the other hand, for reactions involving the axial part of the current, corrections to the rate of nuclear beta decay are expected arising from the strong interactions, and, in the early 1960s, no one knew how to calculate them. Empirically the corrections seemed to be of the order of magnitude of 20 percent.

The work described in the two preceding paragraphs left completely open the question whether or not the neutrino involved in negative muon (μ^-) capture or the decay of the π^- is the same as that appearing in neutron decay. With the advent in the early '60s of the great proton accelerators at the CERN laboratory in Geneva and in Brookhaven, New York, a whole new class of experiments became possible: experiments with high-energy neutrinos. This led to the important experimental discovery that neutrinos resulting from pion decay ($\pi^- \to \mu^- + \bar{\nu}_\mu$) cannot produce the reaction $\bar{\nu}_\mu + p \to n + e^+$, although they do produce $\bar{\nu}_\mu + p \to n + \mu^+$. As the notation already indicates, this experiment shows there are two kinds of neutrinos, ν_e and ν_μ. The same experiments were consistent with the idea that the neutrino appearing in K decay ($K^- \to \mu^- + \bar{\nu}_\mu$) is identical with that appearing in pion decay. To explain this fact in a natural way, the theory of weak interactions has to be extended to decays in which strangeness is not conserved. The simplest way to do this was to add to the previously discussed weak hadron current (which led to processes conserving strangeness) a strangeness-changing weak hadron current. When this was done for the K decay and Λ decay ($\Lambda^0 \to p + e^- + \bar{\nu}_e$), a significant discrepancy appeared; the rates were anomalously low.

The Cabbibo angle

Algebra of currents. An important theoretical development led to a clarification of this situation, the introduction by the Italian physicist Nicola Cabbibo in 1963, of what is now called the Cabbibo angle, θ_c. What Cabbibo proposed was that in the weak hadron current the strangeness-changing part be multipled by sin θ_c, and the part that does not change strangeness, by cos θ_c. As a consequence, the rate of neutron decay is reduced relative to the rate of muon decay by a factor $\cos^2\theta_c$, while the rate of Λ^0 decay is reduced by a factor sin $^2\theta_c$. A value $\theta_c = 13°$ eliminated all major discrepancies involving the vector part of the weak hadron currents.

There remained the problem of computing the corrections arising from the strong interactions to the decays involving the axial vector part of the weak hadron current. To solve it one needs relations between strong and weak interactions. Here, it was the idea of current algebra that led to decisive progress. According to Gell-Mann, the same hadron currents that appear in the weak interactions are basic dynamical variables of the strong-interaction theory and appear directly in the quantities describing the symmetry of the strong interactions. In the idealized case in which $SU(3)$ is an exact symmetry, there are several strictly conserved hadron currents, and the corresponding charges define quantum numbers. Even if, as is indicated by the mass splitting between the multiplets in Figure 2, $SU(3)$ symmetry is only approximate, certain relations among the hadron currents still persist. It is

these that are known as current algebra. These relations enable one to express the strong corrections to those weak decay rates that involve the axial part of the weak hadron current in terms of the scattering of pions on nucleons (a strong interaction effect). Experimental values of the latter are in good agreement with experiments on the weak interactions. Current algebra provides one of the few quantitative links among the strong, weak, and electromagnetic interactions. Many authors took up its methods, and, by the late 1960s, it had become a valuable tool in many branches of particle physics.

With the arrival of the Cabbibo theory, the structure of the weak interactions was largely fixed and in good agreement with experiment. There were only a few indications that essential mysteries remained. One is the phenomenon of CP violation in K^0 decays described above. The other, the very existence of the intermediate W-meson, will be discussed in the next section.

UNDISCOVERED PARTICLES

The particles known as quarks appeared above as a convenient mathematical tool in the description of hadrons. The question naturally arises whether, in fact, free quarks occur in nature. Such quarks, if discovered, would be fundamental building blocks for the particles so far seen; the particle physics of today would become the chemistry of quarks. It would be necessary to go to even higher energies and shorter distances to find out whether or not the quarks themselves are in turn complex objects. That this possibility is taken seriously is indicated by the fact that, as each new accelerator comes into operation, a search is made to see if it is producing quarks.

Whether quarks are naturally occurring

A position somewhat similar to that of the quarks is occupied by the undiscovered W-meson drawn in Table 3, and often called the intermediate (or weak) boson. So far no free W-mesons have been seen, and the principal argument for their existence is that they would complete the analogy between the weak and electromagnetic interactions, with the W-mesons playing a role for the weak interactions analogous to that of the photon for the electromagnetic interactions. Because, in the low-energy experiments to date, it is the ratio $G_F = \sqrt{2}g_W^2 (\hbar^2/m_W c)^2$ that is determined, it is not yet known whether or not the W-meson is really weakly coupled (then $g_W^2/\hbar c$ is much less than one, and m_W is not far from the proton mass) or much more strongly coupled (then m_W would be much larger than the proton mass). A favourite speculation is $e^2/\hbar c = g^2/\hbar c$; then $m_W c^2$ would be about 40 GeV (40 giga-electron volts, *i.e.*, 40,000,000,000 eV). The feasibility of detecting intermediate mesons is strongly dependent on their mass. As of 1971, experiments showed that the mass is necessarily greater than 4 GeV, and new experiments promised to raise this limit considerably.

Magnetic monopoles

The next undiscovered particle, the magnetic monopole, has quite a different place in particle physics. A quantum theory of a particle carrying magnetic instead of electric charge was first constructed by Dirac in 1931. The existence of such a particle in nature is so intriguing a possibility that much experimental effort has gone into its detection, all without success. The Dirac magnetic monopole is a theoretical possibility seemingly uncalled for by any known experimental facts.

From a theoretical point of view, magnetic monopoles are also intriguing. It was pointed out by Dirac that, if a particle of electric charge, e, and a second particle of magnetic charge, e', are present, the product, ee', is quantized so that $ee'/2\hbar c$ is a nonzero integer. The reason for this is that the charges generate an electromagnetic field that carries a nonvanishing angular momentum, ee'/c, and that has therefore the magnitude ½, 1, 3/2, . . . in units of $\hbar$. This fact that $e'e/\hbar c$ is at least ½ implies that the fundamental magnetic charge e' is at least 137/2 times larger than e, the smallest unit of electric charge. This can be seen by invoking the fine-structure constant $e^2/\hbar c = 1/137$ and writing

$$\frac{e'e}{\hbar c} = \frac{e'e^2}{ehc} = \frac{e'}{e} \cdot \frac{1}{137} \geq \frac{1}{2}.$$

consequently the monopole is so strongly coupled to the electromagnetic field that no existing theory can treat it reliably.

Although the dynamical theory invoking concepts of force and motion of magnetic monopoles is shaky, various alternative kinematical possiblities for their description have been explored. For example, Schwinger's "dion" is a hypothetical particle capable of four states of respective electric and magnetic charge (e, e'), $(-e, e')$, $(e, -e')$, and $(-e, -e')$. Dions are used to build up observed particles by a process somewhat analogous to the construction of hadrons from quarks.

The above list of undiscovered particles does not include the graviton, and that for a peculiar reason. Although any reasonable theory of gravitation predicts gravitons, individual gravitons are extraordinarily difficult to detect. Thus it is natural to adopt the attitude of physicists toward neutrinos that prevailed in the early 1950s. Although neutrinos had not been directly detected, the indirect arguments for their existence were so compelling that there were few who doubted it. Similarly, it seems very likely that gravitons exist.

Gravitons

In the late '60s, Joseph Weber, a physicist in the United States, performed experiments to detect gravitational waves—not individual gravitons but coherent bundles of many gravitations. These experiments are controversial, but if they have been interpreted correctly, they constitute a check on the existence of the graviton. What is remarkable about the experiments is that they indicate that the universe contains an enormous amount of energy in gravitational waves. Thus the experiments are more significant for their cosmological implications than for their impact on particle physics.

A similar remark applies to the question of the presence of antimatter in the universe. (As remarked above, antimatter consists of atoms in which the atomic nuclei are composed of antinucleons instead of nucleons and the surrounding cloud consists of positrons instead of electrons.) Particle physicists are already convinced that for every particle there must exist, in principle, an antiparticle. But it will be of great significance for cosmology when and if antimatter is found to occur naturally in some regions of the universe. (A clue would be the discovery of nuclei of antimatter in some of the cosmic rays that arrive at the Earth from distant galaxies.)

Anti-universes

These statements about what particle physics has to learn from cosmology may be too pessimistic. Perhaps the study of the extreme conditions prevailing in very dense stars will lead to some unexpected insights into the role of gravitation in particle physics. Perhaps it is more likely that progress will come from trying to understand what happens to gravitational forces at short distances; this would enable one to establish a fundamental connection between the fourth basic force and the other three. All that is really clear in the early 1970s is that the physics of subatomic particles is still far from being a closed subject.

BIBLIOGRAPHY. There are many texts and journals devoted to particle physics, but most of them assume a knowledge of quantum mechanics. Non-quantum-mechanical treatments may be found in *Scientific American* magazine, specifically M. GELL-MANN and E.P. ROSENBAUM, "Elementary Particles," 197:72–88 (1957); and G.F. CHEW, M. GELL-MANN, and A.H. ROSENFELD, "Strongly Interacting particles," 210:74–93 (1964). For a historical review, see the Nobel Prize Lecture of LUIS W. ALVAREZ, "Recent Developments in Particle Physics," reprinted in *Science*, 165:1071–91 (1969). Other journals that give nontechnical articles are *Physics Today* and the CERN *Courier*, the latter issued in Geneva, Switzerland, by the European Organization for Nuclear Research. There is an annual "Review of Particle Properties," alternating each April between *Reviews of Modern Physics* and *Physics Letters*, which contains all the latest tables of data and references. Some books for the layman are: A.H. ROSENFELD and J.G. GOLDHABER, "Elementary Guide to Elementary (?) Particles," in A.B. BRONWELL (ed.), *Science and Technology in the World of the Future* (1970); R.K. ADAIR and E.C. FOWLER, *Strange Particles* (1963); K.W. FORD, *The World of Elementary Particles* (1963); R.D. HILL, *Tracking Down Particles* (1963); and C.N. YANG, *Elementary Particles* (1962).

(A.H.Ro./A.S.Wi.)

Parturition, Human

Parturition, or labour, is the process of bringing forth a child from the uterus, or womb. The prior development of the child in the womb is described in the article EMBRYOLOGY, HUMAN: *The acquisition of external form;* the changes that take place in the mother during the development of the child in the uterus, in the article PREGNANCY.

At the termination of pregnancy in the human female, the irregular, intermittent contractions of the uterus that began in the early months of pregnancy become more regular and increase in frequency and intensity. This assumption of a rhythmic character by the uterine contractions marks the beginning of the process by which the maternal organism separates and expels the mature products of conception.

Labour is divided into three stages: the first stage, dilatation, having to do with the opening up of the neck of the womb, begins with the onset of labour and ends when the cervix, or neck of the womb, is dilated sufficiently (*i.e.*, nine to ten centimetres, or about four inches, in diameter) to permit the passage of the child's head; the second stage, expulsion—the passage of the child through the maternal birth canal—begins when the cervix is sufficiently dilated to permit the passage of the child's head and terminates with the expulsion of the child; the third, or placental, stage, related to the separation and extrusion of the placenta (afterbirth), begins with the birth of the child and ends when the placenta is expelled and bleeding from the vessels in the uterus is arrested.

THE STAGES OF LABOUR

First stage: dilatation. Early in labour the uterine contractions come on at intervals of 20 to 30 minutes and last about 40 seconds. They are then accompanied by slight pain, which usually is felt in the small of the back. The term labour pains is often used as a synonym for uterine contractions.

As labour progresses these uterine contractions become more intense and also progressively increase in frequency until, at the end of the first stage, when dilatation is complete, they recur about every three minutes and are quite severe. With each contraction a twofold effect is produced to facilitate the opening up of the cervix. Because the uterus, or womb, is a muscular sac containing a bag of waters (the sac containing the amniotic fluid) that more or less surrounds the child, contraction of the musculature of its walls should diminish its cavity and compress its contents. Because its contents are quite incompressible, however, they are forced in the direction of least resistance, which is in the direction of the internal os, or upper opening of the neck of the womb, and are driven, like a wedge, farther and farther into this opening. In addition to forcing the uterine contents in the direction of the cervix, shortening of the muscle fibres that are attached to the neck of the womb tends to pull these tissues upward and away from the opening and, thus, add to its enlargement. By this combined action each contraction of the uterus not only forces the bag of waters and fetus downward against the dilating neck of the womb but also pulls the resisting walls of the latter upward over the advancing bag of waters, presenting (farthest advanced) part of the child.

Duration of first stage

In spite of this seemingly efficacious mechanism, the duration of the first stage of labour is rather prolonged, especially in women who are in labour for the first time. In them the average time required for the completion of the stage of dilatation is between 13 and 14 hours, while, in women who have previously given birth to children, the average is eight to nine hours. Not only does a previous labour tend to shorten this stage but this tendency often increases with succeeding pregnancies, with the result that a woman who has given birth to three or four children may have a first stage of one hour or less in her next labour.

The first stage of labour is often somewhat prolonged, also, in women who become pregnant for the first time after they have passed the age of 35, because the cervix dilates less readily. A similar delay is to be anticipated in cases where the cervix is extensively scarred as a result of previous labours, amputation, deep cauterization, or any other operation on the cervix. Even a woman who has borne several children and whose cervix, accordingly, should dilate readily may have a prolonged first stage if the uterine contractions are weak and infrequent or if the child lies in a faulty position and, as a direct consequence, cannot be forced into the mother's pelvis.

On the other hand the early rupturing of the bag of waters often increases the strength and frequency of the labour pains and thereby shortens the stage of dilatation; occasionally, premature loss of the waters leads to molding of the uterus about the child and thereby delays dilatation by preventing the child's normal descent into the pelvis. Just as an abnormal position of the child and molding of the uterus may prevent the normal descent of the child, an abnormally large child or an abnormally small pelvis may interfere with the descent of the child and prolong the first stage of labour.

Second stage: expulsion. About the time that the cervix becomes fully dilated, the bag of waters breaks, and the force of the involuntary uterine contractions is augmented by voluntary bearing-down efforts of the mother. With each labour pain she takes a deep breath and then contracts her abdominal muscles. The increased intra-abdominal pressure thus produced may equal or exceed the force of the uterine contractions. When properly used, accordingly, these bearing down efforts may double the effectiveness of the labour pains. As the child descends into and passes through the birth passages, the sensation of pain is often increased. This condition is especially true in the terminal phase of the stage of expulsion, when the child's head distends and dilates the maternal soft parts as it is being born.

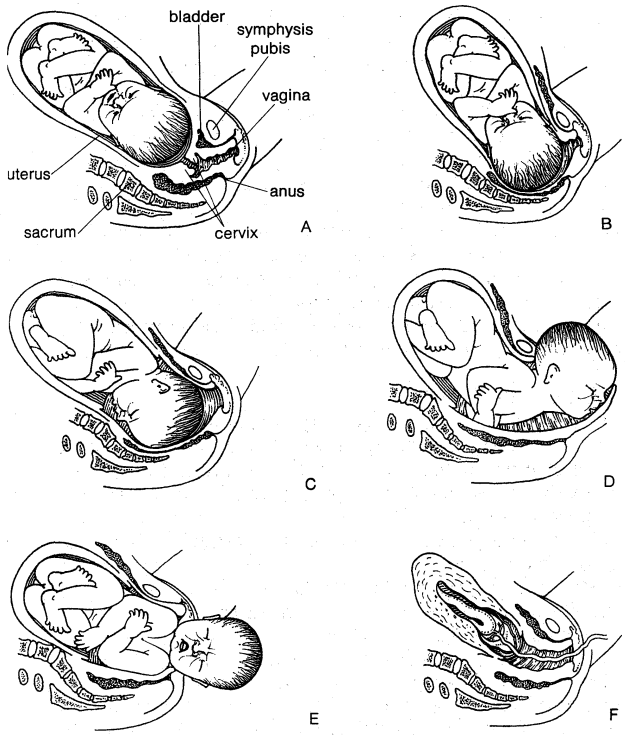
Transverse position of the head. The manner in which the child passes through the birth canal in the second stage of labour depends upon the position in which it is lying and the type of the mother's pelvis. The sequence of events described in the following numbered paragraphs is that which frequently occurs when the mother's pelvis is of the usual type and the child is lying with the top of its head lowermost and transversely placed and the back of its head (occiput) directed toward the left side of the mother (A in the Figure). The top of the head, accordingly, is leading and its long axis lies transversely.

Flexion and internal rotation of head

1. The force derived from the uterine contractions and the bearing-down efforts exerts pressure on the child's buttocks and is transmitted along the vertebral column to drive the head into and through the pelvis. Because of the eccentric attachment of the spine to the base of the skull, the back of the head is made to advance more rapidly than the brow with the result that the head becomes flexed (*i.e.*, the neck is bent) until the chin comes to lie against the breastbone (B in the Figure). As a consequence of this flexion mechanism, the top of the head (occiput) becomes the leading pole and the ovoid head circumference that entered the birth canal is succeeded by a smaller, almost circular circumference, the long diameter of which is about two centimetres (three-fourths inch) shorter than that of the earlier circumference.

2. As the head descends more deeply into the birth canal, it meets the resistance of the bony pelvis and of the slinglike pelvic floor, or diaphragm, which slopes downward, forward, and inward. When the back of the head, the leading part of the child, is forced against this sloping wall on the left side, it naturally is shunted forward and to the right as it advances, just as a ferryboat is shunted into its wharf by meeting the sloping resistance of its ways (C in the Figure). This internal rotation of the head brings its longest diameter into relation with the longest diameter of the pelvic outlet and thus greatly assists in the adaptation of the advancing head to the configuration of the cavity through which it is to pass.

3. Further descent of the head directly downward in the direction in which it has been travelling is opposed by the lower portion of the mother's bony pelvis, behind, and



Sequential changes in the position of the child during labour. (A) Onset of labour. (B) Flexion. (C) Internal rotation of the head. (D) Extension. (E) External rotation of the head. (F) Uterus immediately after birth; cord has been cut and placenta is separating from uterine wall.
Drawing by Ramon Goas after Alfred C. Beck

the resisting soft parts that are interposed between it and the opening of the vagina (C in the Figure). Less resistance, on the other hand, is offered by the soft and dilatable walls of the lower birth canal, which is directed forward and upward. The back of the child's head accordingly advances along the lower birth canal, distending its walls and dilating its cavity while the head progresses. Soon the back of the child's neck becomes impinged against the bones of the pelvis, in front, and the chin is forced farther and farther away from the breastbone. Thus, as extension (bending of the head backward) takes the place of flexion, the occiput, brow, eye sockets, nose, mouth, and chin pass successively through the external opening of the lower birth canal and are born (D in the Figure).

Extension of head and then resumption of original position

4. The neck, which was twisted during internal rotation of the head, untwists as soon as the head is born. Almost immediately after its birth, therefore, the occiput is turned toward the left and backward.

As the child's lower shoulder advances, it meets the sloping resistance of the pelvic floor on the right side and is shunted forward and to the left toward the middle of the pelvis in front. This position brings the long diameter of the shoulder circumference into relation with the anteroposterior, or long diameter, of the pelvic cavity. Because of this internal rotation of the shoulders, the occiput undergoes further external rotation backward and to the left so that the child's face comes to look directly at the inner aspect of the mother's right thigh (E in the Figure).

Birth of the shoulders. Soon after the shoulders rotate, the one in front appears in the vulvovaginal orifice and remains in this position while the other shoulder is swept forward by a lateral bending of the trunk through the same upward and forward curve that was followed by the head as it was being born. After this shoulder is delivered, the shoulder in front and the rest of the child's body are expelled almost immediately and without any special mechanism.

An average of about one hour and 45 minutes is required for the completion of the second stage of labour in women who give birth for the first time. In subsequent labours the average duration of the stage of expulsion is somewhat shorter.

Posterior position of the occiput. The child may lie so that the back of its head is directed backward and toward either the right or left side. The leading pole is then in the right or left posterior quadrant of the mother's pelvis, and the condition is referred to as a posterior position of the occiput. In such cases the back of the child's head usually rotates to the front of the pelvis and labour proceeds as in transverse positions. Because of the longer rotation required, labour may be somewhat more prolonged than in transverse positions.

Face presentation. When the child's head becomes bent back (extended) so that it enters and passes through the pelvis face first, the condition is known as a face presentation. The chin is then the leading pole and follows the same course that is followed by the back of the head in occipital presentations. If the chin lies to the front as it enters the pelvis, labour often is easy and of short duration. Should it be directed backward, on the other hand, considerable difficulty may be encountered, and the head may have to be flexed or rotated artificially.

Breech presentation. Passage of the lower extremities or the buttocks through the pelvis first, called breech presentation, is encountered about once in every 30 labours. Because the head in such cases is the last part of the child to be delivered and because this part of the delivery is the most difficult, the umbilical cord (navel string) may be compressed while the aftercoming head is being born, with the result that the child may be asphyxiated. Asphyxia or injurie , to the child that result from the attendant's effort to hasten the delivery in order to prevent the child's asphyxiation are responsible for the loss of three times as many breech babies as head-on babies.

The infant mortality rate in countries with well-developed systems of medical care varies from 2 percent to 10 percent according to the size of the child and skill of the attendant. Because very small, premature infants are particularly susceptible to the dangers of breech delivery, the mortality among them is very high when they are born breech first.

Transverse presentation or cross birth. In this relatively rare situation the long axis of the child tends to lie across, or transverse to, the long axis of the mother. Unless the child is very small or has been dead for some time and has become greatly softened, delivery through the natural passages is impossible in such cases. For this reason the child must be turned by the attendant or delivered by the surgical procedure called cesarean section (through the mother's abdominal wall) if it is alive.

Slight risks of pregnancy

That the above-mentioned complications are infrequent and can be cared for easily is shown by the excellent statistics of many maternity services, in which the maternal death rate is less than one per 1,000 and would be still lower if the deaths caused by complicating systemic diseases were excluded. The infant mortality rate is also usually low, ranging between 1.5 percent and 3 percent. It likewise would be much lower if the premature and poorly developed infants were excluded. In other words, the risk to a healthy mother who carries her child to maturity is less than one per 1,000, and the risk to her mature child is about 0.5 percent.

Third stage: placental stage. With the expulsion of the child, the cavity of the uterus is greatly diminished (F in the Figure). As a consequence the site of placental attachment becomes markedly reduced in size, with the result that the placenta (afterbirth) is separated in many places from the membrane lining the uterus. Within a few minutes subsequent uterine contractions complete the separation and force the afterbirth into the vagina, from which it is expelled by a bearing-down effort. The third stage of labour, accordingly, is of short duration, seldom lasting longer than 15 minutes. Occasionally, however, the separation may be delayed and accompanied by bleeding, in which case artificial removal of the placenta is necessary. The attendant always remains with the mother until the possibility of hemorrhage has been eliminated by firm retraction of the uterus.

(A.C.Bk.)

RELIEF OF PAIN IN LABOUR

Not much was done to relieve the suffering of childbirth before chloroform was first widely used in the 1850s for this purpose. Because of its toxicity, even this drug was employed sparingly and ultimately came into disuse. Early in the 20th century, a combination of the drugs morphine and scopolamine was given, according to the "twilight sleep" technique, to lessen the pain that accompanies the stage of cervical dilatation. While this technique was found to be less safe than had been hoped, it gave a new and strong impetus to the search for a satisfactory method of pain relief, with the result that the old methods were improved and many new ones were suggested. No completely satisfactory method was found, and, at the beginning of the 1970s, it still remained impossible to secure total relief from pain in all women without adding to the maternal and fetal risk. On the other hand, the judicious employment of one or more of the drugs described briefly in the following paragraphs eliminates much of the pain of childbirth with safety (see also ANESTHETIC).

Twilight sleep

Morphine and scopolamine. By the use of injections of morphine and scopolamine, a condition of seminarcosis, twilight sleep, is induced. When the method is successful, the mother awakens after her child is born and has no recollection of having felt any pain during her labour. Such a result is attributable either to the actual relief of pain or to her inability to remember such pain as did occur because of the scopolamine-induced loss of memory. If given too soon, these drugs may stop the labour completely. If given too late, their effect on the child's respiratory centre may cause its death from asphyxia. This lack of safety has led to more or less abandonment of the twilight sleep routine.

Barbiturates. The barbiturate drugs, usually combined with scopolamine, are frequently employed for the relief of pain during labour. As in the case of twilight sleep, their successful use is followed by a loss of memory, and the mother either has no pain or forgets it completely. To obtain a satisfactory result, unpleasant side effects sometimes are encountered. The commonest of these is excitement of varying degrees. Restraint, accordingly, may be necessary, and the constant presence of a nurse or other attendant is required. The patient's restlessness or inability to use her bearing-down efforts properly often makes it necessary to deliver the child with instruments and with the mother under anesthesia. Although some infants delivered in this way do not breathe so readily as do those born to mothers who have had no medication for the relief of pain, the effect of the barbiturates on the child is much less marked than that produced by the twilight sleep routine.

Meperidine. The drug meperidine, combined with scopolamine, has been employed to relieve pain and to induce forgetfulness (amnesia) during labour. Evidence accumulated after the mid-20th century seemed to indicate that this combination would ultimately prove to be the safest and best method of pain relief during the first stage of labour. The dilatation of the cervix apparently is hastened by use of meperidine. If the combination of meperidine and scopolamine is supplemented by local anesthesia during the stage of expulsion, the greater part of the pain of childbirth may be eliminated without risk to mother or child.

Nitrous oxide. Nitrous oxide, or laughing gas, is one of the most popular of the analgesic (pain-relieving) agents. It is given during the latter part of the first and throughout the second stage of labour and is administered only for the duration of the uterine contractions. While the child is being delivered, the addition of ether is usually necessary. Unless adequate amounts of oxygen are added to the nitrous oxide mixture that the mother breathes, the child may suffer from lack of oxygen or die from asphyxia.

Chloroform, ether, ethylene, and cyclopropane. Chloroform, ether, ethylene, and cyclopropane are often given while the child is being delivered. Chloroform is rarely used because of its toxicity. Ether is much safer than chloroform but is not pleasant to take and is irritating to the respiratory passages. Ethylene is not widely used in obstetrics, because it is disagreeable to take and is highly explosive. For cyclopropane, an excellent gaseous anesthetic agent well suited to the needs of obstetric practice but highly explosive, a special apparatus is required for administration, and adequate measures must be employed to prevent the occurrence of static sparks. Trichloroethylene, a volatile anesthetic agent, is suitable for self-administration during the latter part of labour. The patient inhales it through a special inhaler that she places over her nose and mouth at the onset of each uterine contraction. Loss of consciousness occurs before the contraction ends; the inhaler thus drops away from her nose and the administration of the gas is discontinued. Trichloroethylene anesthesia is used extensively in Great Britain.

Self-administration of trichloroethylene

OPERATIVE OBSTETRICS

Most women deliver a baby spontaneously. Complications, however, that were present before labour or that develop during labour may threaten the life of the mother or of the baby and may require intervention by the attending physician.

When a complication develops, the obstetrician's choice of treatment depends upon the problem. He may use the obstetrical forceps, an instrument that is so constructed that it can be applied to the fetal head for the purpose of extracting the fetus by traction or of rotating the fetal head in order to correct an unsatisfactory or undeliverable position. Forceps delivery is indicated when the mother's expulsive forces are unable to effect a spontaneous delivery because of ineffective uterine contractions, abnormalities in the mechanism of labour, or lesser degrees of disproportion between the size of the baby's head and the mother's pelvis. Labour is terminated by forceps when the mother or the fetus is in distress that can be relieved by prompt delivery. It is often used for delivery of the aftercoming head. Forceps delivery may save the mother from the stress of the second stage of labour in some cases of heart disease, tuberculosis, and other debilitating diseases. Infection occurring during delivery, certain cases of placenta praevia, and premature separation of the placenta may be indications for forceps delivery. (Placenta previa, described in the article PREGNANCY, is development of the placenta so that it covers the internal opening of the cervix and must precede the infant during delivery.) Conditions must be suitable, of course, for forceps delivery before the instrument is employed.

Manual rotation may be used instead of the forceps when the fetal head is in an abnormal position that makes delivery difficult or impossible. In carrying out the procedure, the obstetrician inserts his hand into the birth canal and turns the fetal head to a more favourable position.

Bleeding from the uterus during or after the third stage of labour may be controlled by manual removal of the placenta and packing of the uterus. In the former method the operator inserts his hand into the uterine cavity and separates the adherent placenta from the uterine wall. Modern obstetricians rarely pack the uterus with gauze to control bleeding, because it fails to stop bleeding, and it carries with it an increased risk of infection.

Cesarean section

When conditions are such that the delivery of a child through the vagina would be a hazard to the mother or to the fetus, it may be necessary or preferable to resort to cesarean section, a procedure in which the fetus is delivered through a surgical opening in the uterus made after the uterus has been exposed through an opening in the abdominal wall.

Cesarean section is so much better for the mother and the fetus when there is bleeding from a placenta previa or prematurely separated placenta that it is the preferred method of delivery in many cases of uterine hemorrhage during the later months of pregnancy and labour. Difficult or prolonged labour is a frequent indication for abdominal delivery. Uterine inertia, once one of the most common causes for prolonged labour, makes cesarean section necessary less frequently than it did a decade ago because of the widespread use of very weak solutions of oxytocin to stimulate uterine contractions in such cases.

(Oxytocin is a hormone produced by the nerve cells of the hypothalamus, a centre at the top of the brainstem; the hormone is stored in the posterior lobe of the pituitary.)

Prolonged or difficult labour resulting from either a small maternal pelvis, a large fetus, or a malposition or malpresentation of the fetus is one of the most frequent reasons that some women have their first, or primary, cesarean section.

As a general rule after a woman has undergone one abdominal delivery, she is delivered thereafter in the same fashion. The principle of "once a cesarean, always a cesarean" came about because of the danger of rupture of the uterine scar, with loss of the baby's and often the mother's life if she is permitted to deliver vaginally after a previous cesarean section.

The woman who suffers from obstructed labour is delivered abdominally if labour cannot be terminated vaginally without danger to her or the baby. Threatened rupture of the uterus is treated by immediate cesarean section. Ordinarily the patient in labour who has been neglected and for this reason is actually or potentially infected is delivered by some type of cesarean section.

A number of conditions that would not be considered imperative reasons for abdominal delivery may, at times, be factors influencing the choice of cesarean section. Breech presentation by itself is not a reason for cesarean section; a large baby, however, that is presenting by the breech in a woman past her middle 30s who is bearing a child for the first time could quite possibly be considered an indication for cesarean section because of the increased hazard to a large baby in a breech delivery and the possibility that it might be the patient's only opportunity to bear a child. Absolute or relative indications for cesarean section include a number of uncommon conditions, such as multiple pregnancy; prolonged pregnancy; a malignant tumour of one of the pelvic organs; a previous operation, the results of which interfere with normal function of the uterus; an anomaly of the genitalia; a benign pelvic tumour that obstructs the birth canal; a paralytic neuromuscular disorder, such as poliomyelitis, that prevents the mother from exercising her voluntary expulsive forces; or death of fetuses before or during earlier childbirths.

Cesarean section is often necessary when the mother is diabetic or has toxemia of pregnancy, a complex series of metabolic disorders that involve or may involve excessive vomiting, convulsions, and coma. The procedure is also resorted to in some instances when a woman has her first child late in her child-bearing years. Cesarean section may seem necessary if the mother has not only chronic high blood pressure and disease of the blood vessels but also pre-eclampsia, a condition in which there is a sudden rise in blood pressure, fluid accumulates in the tissues, and protein in the blood reaches abnormal heights. Premature expulsion of the umbilical cord may call for cesarean section if the child is alive and immediate vaginal delivery is not possible.

In rare instances the obstetrician may make incisions in the incompletely dilated uterine cervix. The incisions permit vaginal delivery in those cases in which labour is unduly prolonged and the incomplete dilation of the cervix is the only thing that is delaying delivery.

Procedures now rarely used

Three procedures formerly employed but now seldom used are (1) internal podalic version, an operation in which the obstetrician inserts his hand into the uterus and turns the fetus around so that it delivers by the breech (see above); (2) symphysiotomy; and (3) pubiotomy. Symphysiotomy is a surgical incision through the symphysis pubis, the joint between the two pubic bones of the pelvis at the front midline of the body; and pubiotomy is cutting through a pubic bone. Symphysiotomy and pubiotomy were formerly used when the fetus was too large for the mother's pelvis but have now been largely abandoned, at least in the United States, in favour of cesarean section.

Another procedure, formerly necessary in emergencies but now never obligatory, is the destruction of a living, normal infant, because its delivery intact would be incompatible with preservation of the mother's life. Today, modern techniques and antibiotics make possible the delivery of any infant intact, alive or dead, by methods that are likely to be safer for the mother than a difficult, destructive operation.

ACCIDENTS DURING LABOUR

Lacerations. *Lacerations of the perineum, vulva, and vagina.* Vaginal lacerations usually make themselves known by profuse bleeding after delivery of the baby. Not all extensive lacerations cause bleeding, however, and a large tear in the vaginal wall may not be discovered until the obstetrician inspects the vagina after the placenta is delivered. There is no difficulty in diagnosing lacerations near the external opening of the birth canal, because they are easily seen by the obstetrician. Even minor lacerations are repaired, because, if they are not, granulation tissue may form in the wounds and delay healing. Deep lacerations require anatomical reconstruction of the torn tissues. Extensive tears of the perineum (the tissues between the genital organs and the anus) can often be avoided by performing an episiotomy—an incision in the vulvar orifice, the external genital opening—before delivery of the infant's head. Also, attention on the obstetrician's part to the mechanism of labour, manual assistance in delivery of the head and shoulders, avoidance of too rapid delivery, delivery between pains, and the proper use of the forceps are some of the many measures that help to avoid injuries not only to the perineum but to all of the genital tissues.

Cervical lacerations. The cervix, the lower end of the uterus that projects into the vagina, is usually inspected after the placenta has been delivered. Superficial tears look somewhat like a frayed edge on the cuff-like cervix. Deeper lacerations usually cause serious bleeding immediately before or after delivery of the placenta.

Small cervical lacerations are not repaired; they heal spontaneously. Deeper tears are sutured. The management of extensive tears into the body of the uterus or the broad ligaments that support the uterus depends on the extent of the injury and its location; an abdominal operation is sometimes required to control the bleeding and to repair the uterus. Occasionally hysterectomy—removal of the uterus—is necessary.

Other accidents. *Rupture of the uterus.* Rupture of the uterus may occur spontaneously; it may be caused by trauma, or it may occur when a cesarean-section scar gives way. The classical signs of impending spontaneous rupture are gradually increasing, constant, severe pain in the lower part of the abdomen, restlessness, a rising temperature, an increasing pulse rate, and a tense, tender uterus that does not relax between strong contractions. When rupture occurs, the patient complains, usually, of extreme pain and then a sensation of something tearing or giving way. Uterine contractions stop. There is extensive internal bleeding. The baby's body can be felt in the mother's abdomen beside the contracted uterus.

Signs of impending rupture

Prompt delivery, almost always by cesarean section, is the treatment of impending rupture. The patient is anesthetized to stop uterine contractions as soon as the diagnosis is made.

An immediate abdominal operation follows the diagnosis of uterine rupture. Bleeding from the torn uterine walls must be stopped as promptly as possible. The fetus is removed. A hysterectomy is usually performed, because the ragged uterine scar is likely to rupture again if the patient has another term pregnancy, and bleeding from the torn uterus is difficult to control. Such patients often require generous quantities of transfused blood. Antibiotics are given, because infection is or may be present.

Injuries to the pelvic supporting tissues. Injuries to the pelvic supporting tissues may not be evident during labour. Months or years later the diagnosis will be made, because the patient complains of something bulging from the vagina, involuntary loss of urine while coughing or laughing, a sensation of things falling down, dragging pelvic discomfort, and difficulty in emptying the lower bowel. The bulging mass formed by a cystourethrocele or rectocele, when seen at vaginal examination, confirms it.

(A cystourethrocele is a sagging of the urethra and bladder out of position; a rectocele is a protrusion of the rectum into the vagina.) At times the result of the injuries to the pelvic supporting ligaments and muscles may be so severe that the uterus lies completely outside the vagina, and the vagina is turned inside out, forming a huge ball-shaped mass lying between the patient's thighs.

Treatment depends on the severity of the symptoms. Many women live out their lives without much distress from small cystourethroceles and rectoceles. Larger lesions are repaired surgically.

Inversion of the uterus. Another accident that may occur during labour is inversion of the uterus. The uterus turns inside out and upside down so that its inner surface lies outside and against the wall of the vagina. Inversion causes sudden shock. There may be severe bleeding. The diagnosis is made by noting the pear-shaped mass, covered by a shaggy, dark-red, bleeding surface, filling the vagina or hanging outside. The placenta may be still attached to it.

Restoration of a uterus to its normal position is accomplished after the patient's shock and hemorrhage are treated, and she is anesthetized. The obstetrician inserts his hand into the patient's vagina and lifts up the uterus. The tension applied to the uterine ligaments by this procedure usually reinverts the uterus.

Embolisms. Embolisms may occur. (An embolism is a blockage of a blood vessel, as by a blood clot or bubble of air.) Amniotic fluid embolism causes sudden, severe respiratory distress, signs of shock, cyanosis (blueing of the skin), heart collapse, and circulatory failure. If the diagnosis is made promptly, oxygen, blood transfusion, and the injection of fibrinogen, a clotting factor, into a vein may be lifesaving.

Effects of amniotic fluid embolism

Air embolism causes the patient to become suddenly short of breath and cyanotic. She may have heart pain and show signs of shock. The heart beats irregularly, and swishing sounds, caused by the presence of air mixed with blood in the heart, can often be heard. Death follows quickly unless the diagnosis is made at once. Treatment consists of drawing the air from the heart with a needle and syringe.

Fibrinogenopenia. Fibrinogenopenia, a shortage of fibrinogen, which causes defective blood clotting and serious bleeding, may result from a number of circumstances, including premature separation of the placenta and a hard, forceful labour. The diagnosis is confirmed by testing the mother's blood. It lacks fibrinogen if it does not clot firmly in ten minutes or if the clot that does form dissolves when incubated at 37° C (98.6° F) for one hour. Treatment includes management of the condition responsible for the lack of fibrinogen, replacement of blood lost through bleeding, and the administration of fibrinogen intravenously.

Placenta accreta. Abnormal adherence of the placenta to the uterus, a condition called placenta accreta, is suspected when the placenta cannot be expelled. If the adherence is only partial, there is usually brisk bleeding from the remainder of the placental site. Formerly, placenta accreta was treated by immediate removal of the uterus. In the late 1960s it was suggested that the placenta should be left in place to disintegrate, in which case the patient is protected from infection with antibiotics.

Abruptio placentae

Genital bleeding during labour. Genital bleeding during labour is a symptom of a number of obstetrical complications in addition to those discussed above. The diagnosis of premature separation of the placenta (abruptio placentae) is made when the patient complains of sudden abdominal pain and when there is uterine tenderness and vaginal bleeding. The mother may go into shock, and the fetus may die. There may be signs of hidden bleeding and concealed blood within the uterus. This condition is differentiated from placenta previa by the fact that the placenta is not in the lower uterine segment.

In cases of suspected placenta previa, the placenta can be located with considerable accuracy, as by X-ray. The definitive diagnosis of placenta previa is made, however, when vaginal examination, which is not performed until everything is in readiness for immediate vaginal or abdominal delivery, reveals part of the placenta in the lower uterus.

In every case of serious blood loss, bleeding must be controlled and lost blood replaced in amounts sufficient to protect the mother. Most patients with placenta previa and premature detachment of the placenta are treated by cesarean section. Cesarean section is not performed for premature detachment of the placenta (abruptio placentae) until a shortage of fibrinogen is corrected. Removal of the uterus may be necessary in rare cases of abruptio placentae when the uterine wall has been so infiltrated by blood that it cannot contract.

Severe bleeding not related to the pregnancy, such as that resulting from rupture of one of the pelvic veins or rupture of the spleen, may occur during labour. When it occurs, it must be treated regardless of the pregnancy.

Accidents to the umbilical cord. An accident to the umbilical cord is suspected when there is marked irregularity in the fetal heart rate and particularly when the irregularity is accentuated by uterine contractions. A prolapsed cord—that is, a cord lying below the head—can be felt through the membranes on vaginal examination. After the membranes have ruptured, the cord can be felt and seen in the vagina. It may hang out of the vulva. The patient is delivered by cesarean section if the infant is alive and if the head can be prevented from pressing on the cord while preparations are made for the operation. The baby is delivered vaginally if the cervix is completely dilated and if conditions are favourable for prompt vaginal delivery. Attempts to replace the cord in the uterus are seldom successful. Vaginal delivery is allowed to continue if the infant is dead.

True knots in the cord and rupture of the cord with bleeding are seldom diagnosed until after delivery. They are usually associated with sudden and, at the time, inexplicable fetal death. (J.W.Hu.)

NATURAL CHILDBIRTH

In the 1940s Grantly Dick-Read, a British obstetrician, developed a technique of delivery called natural childbirth that minimized the surgical and anesthetic aspects of delivery and concentrated upon the mother's conscious effort to give birth to her child. Although opposed by many physicians who felt that it denied the progress of modern medicine and needlessly primitivized the process of birth, the method was gradually accepted and by the late 1950s was practiced by a sizable percentage of women, especially in the United States and England.

Arguments for natural childbirth

Natural childbirth, as formulated by Dick-Read, stems from his premise that childbirth need not be accompanied by excessive pain and that labour pains are the result of unnatural physical tension caused by fear. Dick-Read maintains that fear can be counteracted by understanding and an ability to relax, and his method prescribes for the expectant mother a lengthy course of instruction in the mechanics of labour and birth as well as exercises to strengthen the musculature and to encourage proper breathing. During her labour the mother is aided by trained personnel so that she does not forget her training, and small quantities of anesthetic are made available to her when needed. No claims are made that natural childbirth is totally painless; it rather enables the mother's physical response to transcend discomfort.

Natural childbirth is not considered advisable for women in poor health or for those who have physical defects or psychological problems. For the healthy, determined young mother, it presents the advantage of allowing her to participate actively, rather than passively, in labour and to experience the actual moment of birth. A disadvantage is the amount of time required for the prenatal instruction course, which is impractical for many mothers. (Ed.)

PUERPERIUM OR PERIOD OF INVOLUTION

Within six to eight weeks after childbirth, most of the structures of the maternal organism that underwent change during pregnancy return more or less to their prepregnancy state. The enlarged uterus, which at the end of gestation weighed about 1,000 grams (35 ounces),

shrinks to a weight of about 60 grams (two ounces). Along with this process of uterine involution, the lining membrane of the uterus is almost completely shed and replaced by a new lining, which is then (six to eight weeks after delivery) ready for the reception of another fertilized ovum (egg). The greatly dilated neck of the womb and lower birth passage likewise undergo marked and rapid involution, but they seldom return exactly to their prepregnancy condition. As a rule examination of a woman who has given birth accordingly reveals evidence of this. The markedly stretched abdominal wall also undergoes considerable involution, particularly if abdominal exercises are carried out. Although the intradermal tears (striae gravidarum) become smaller and silvery white, they do not completely disappear but remain as evidence of the marked and rapid stretching of the skin that took place during pregnancy. The breasts, unlike most of the other organs, continue to increase in size. By the second or third day after childbirth, they become so distended that they are painful. After the milk comes in on the third day, this distension recedes, but the breast enlargement persists as long as lactation continues. If the child is not allowed to nurse on its mother's breasts, lactation ceases within a short time, and mammary involution follows (see also LACTATION, HUMAN).

BIBLIOGRAPHY. Readers wishing to pursue the subject further are directed to the following specialized texts: N.J. EASTMAN and L.M. HELLMAN, *Williams Obstetrics,* 13th ed. (1966); J.P. GREENHILL, *Obstetrics,* 13th ed. (1965); J.W. HUFFMAN, *Gynecology and Obstetrics* (1962).

(J.W.Hu.)

Pascal, Blaise

Blaise Pascal, a mathematician, physicist, religious philosopher, and master of French prose, laid the foundation for the modern theory of probabilities, formulated what came to be known as Pascal's law of pressure, and propagated a religious doctrine that taught the experience of God through the heart rather than through reason. The establishment of his principle of intuitionism had an impact on such later philosophers as Jean-Jacques Rousseau and Henri Bergson and also on the Existentialists.

By courtesy of the Bibliotheque Nationale, Paris

Pascal, red crayon drawing by Jean Domat, c. 1649. In the Bibliothèque Nationale, Paris.

Pascal's life to the Port-Royal years. He was born on June 19, 1623, at Clermont-Ferrand, France, where his father, Étienne Pascal, was presiding judge of the tax court. Pascal's mother died in 1626, and in 1631 the family moved to Paris. Étienne, who was respected as a mathematician, devoted himself henceforth to the education of his children. While his sister Jacqueline (born in 1625) figured as an infant prodigy in literary circles, Blaise proved himself no less precocious in mathematics. In 1640 he wrote an essay on conic sections, *Essai pour les coniques,* based on his study of the now classical work of Girard Desargues on synthetic projective geometry. The young man's work, which was highly successful in the world of mathematics, aroused the envy of no less a personage than the great French Rationalist and mathematician René Descartes. Between 1642 and 1644, Pascal conceived and constructed a calculating device to help his father—who in 1639 had been appointed intendant (local administrator) at Rouen—in his tax computations. The machine was regarded by Pascal's contemporaries as his main claim to fame, and with reason, for in a sense it was the first digital calculator.

Until 1646 the Pascal family held strictly Roman Catholic principles, though they often substituted *l'honnêteté* ("polite respectability") for inward religion. An illness of his father, however, brought Blaise into contact with a more profound expression of religion, for he met two disciples of the abbé de Saint-Cyran, who, as director of the convent of Port-Royal, had brought the austere moral and theological conceptions of Jansenism into the life and thought of the convent. Jansenism was a 17th-century form of Augustinianism in the Roman Cathòlic Church. It repudiated free will, accepted predestination, and taught that divine grace, rather than good works, was the key to salvation. The convent at Port-Royal had become the centre for the dissemination of the doctrine. Pascal himself was the first to feel the necessity of entirely turning away from the world to God, and he won his family over to the spiritual life in 1646. His letters indicate that for several years he was his family's spiritual adviser, but the conflict within himself—between the world and ascetic life—was not yet resolved. Absorbed again in his scientific interests, he tested the theories of Galileo and Evangelista Torricelli (an Italian physicist who discovered the principle of the barometer). To do so, he reproduced and amplified experiments on atmospheric pressure by constructing mercury barometers and measuring air pressure, both in Paris and on the top of the Puy de Dôme, a mountain overlooking Clermont-Ferrand. These tests paved the way for further studies in hydrodynamics and hydrostatics. While experimenting, Pascal invented the syringe, refined Torricelli's barometer, and created the hydraulic press, an instrument based upon the principle that became known as Pascal's law: pressure applied to a confined liquid is transmitted undiminished through the liquid in all directions regardless of the area to which the pressure is applied. His publications on the problem of the vacuum (1647–48) added to his reputation. When he fell ill from overwork—he was of a weak constitution—his doctors advised him to seek distractions; but what has been described as Pascal's "worldly period" (1651–54) was, in fact, primarily a period of intense scientific work, during which he composed treatises on the equilibrium of liquid solutions, on the weight and density of air, and on the arithmetic triangle: *Traité de l'équilibre des liqueurs et de la pesanteur de la masse de l'air* (Eng. trans., *The Physical Treatises of Pascal,* 1937) and also his *Traité du triangle arithmétique.* In the last treatise, a fragment of the *De Alea Geometriae,* he laid the foundations for the calculus of probabilities. By the end of 1653, however, he had begun to feel religious scruples; and the "night of fire," an intense, perhaps mystical "conversion" that he experienced on November 23, 1654, he believed to be the beginning of a new life. He entered Port-Royal in January 1655, and though he never became one of the solitaries, he thereafter wrote only at their request and never again published in his own name. The two works for which he is chiefly known, *Les Provinciales* and the *Pensées,* date from the years of his life spent at Port-Royal.

Contributions to physics and mathematics

Association with the Convent of Port-Royal

"Les Provinciales." Written in defense of Antoine Arnauld, an opponent of the Jesuits and a defender of Jansenism who was on trial before the faculty of theology in Paris for his controversial religious works, Pascal's 18 *Lettres écrites par Louis de Montalte à un provincial* deal with divine grace and the ethical code of the Jesuits. They are better known as *Les Provinciales.* They included a blow against the relaxed morality that the Jesuits were said to teach and that was the weak point in their controversy with Port-Royal; Pascal quotes freely Jesuit dialogues and discrediting quotations from their own works, sometimes in a spirit of derision, sometimes with indignation. In the two last letters, dealing with the question of

grace, Pascal proposed a conciliatory position that was later to make it possible for Port-Royal to subscribe to the "Peace of the Church," a temporary cessation of the conflict over Jansenism, in 1668.

The *Provinciales* were an immediate success, and their popularity has remained undiminished. This they owe primarily to their form, in which for the first time bombast and tedious rhetoric are replaced by variety, brevity, tautness, and precision of style; as Nicolas Boileau, the founder of French literary criticism, recognized, they marked the beginning of modern French prose. Something of their popularity, moreover, in fashionable, Protestant, or skeptical circles, must be attributed to the violence of their attack on the Jesuits. In England they have been most widely read when Roman Catholicism has seemed a threat to the Church of England. Yet they have also helped Catholicism to rid itself of laxity; and, in 1678, Pope Innocent XI himself condemned half of the propositions that Pascal had denounced earlier. Thus, the *Provinciales* played a decisive part in promoting a return to inner religion and helped to secure the eventual triumph of the ideas set forth in Antoine Arnauld's treatise *De la fréquente communion* (1643), in which he protested against the idea that the profligate could atone for continued sin by frequent communion without repentance, a thesis that thereafter remained almost unchallengeable until the French church felt the repercussion of the revocation of the Edict of Nantes (which had granted religious freedom to French Protestants) in 1685. Whereas the Jesuits seemed to represent a Counter-Reformation predominantly concerned with orthodoxy and obedience to ecclesiastical authority, the *Provinciales* advocated a more spiritual approach, emphasizing the soul's union with the Mystical Body of Christ through charity.

Morality and spirituality

Further, by rejecting any double standard of morality and the distinction between counsel and precept, Pascal aligned himself with those who believe the ideal of evangelical perfection to be inseparable from the Christian life. Although there was nothing original in these opinions Pascal nevertheless stamped them with the passionate conviction of a man in love with the absolute, of a man who saw no salvation apart from a heartfelt desire for the truth, together with a love of God that works continually toward destroying all self-love. For Pascal, morality cannot be separated from spirituality. Moreover, his own spiritual development can be traced in the *Provinciales*. The religious sense in them becomes progressively refined after the first letters, in which the tone of ridicule is smart rather than charitable.

"Pensées." Pascal finally decided to write his work of Christian apologetics, *Apologie de la religion chrétienne*, as a consequence of his meditations on miracles and other proofs of Christianity. The work remained unfinished at his death. Between the summers of 1657 and 1658, he put together most of the notes and fragments that editors have published under the inappropriate title *Pensées* ("Thoughts"; Eng. trans., *Pensées*, 1962). In the *Apologie*, Pascal shows the man without grace to be an incomprehensible mixture of greatness and abjectness, incapable of truth or of reaching the supreme good to which his nature nevertheless aspires. A religion that accounts for these contradictions, which he believed philosophy and worldliness fail to do, is for that very reason "to be venerated and loved." The indifference of the skeptic, Pascal wrote, is to be overcome by means of the "wager": if God does not exist, the skeptic loses nothing by not believing in him; but if he does exist, the skeptic gains eternal life by believing in him. Pascal insists that men must be brought to God through Jesus Christ alone, because a creature could never know the infinite if Jesus had not descended to assume the proportions of man's fallen state.

The second part of the work applies the Augustinian theory of allegorical interpretation to the biblical types (*figuratifs*); reviews the rabbinical texts, the persistence of true religion, the work of Moses and the proofs concerning Jesus Christ's God-like role; and, finally, gives a picture of the primitive church and the fulfillment of the prophecies. The *Apologie* (*Pensées*) is a treatise on spirituality. Pascal was not interested in making converts if they were not going to be saints.

Pascal's apologetics

Pascal's apologetic, though it has stood the test of time, is primarily addressed to individuals of his own acquaintance. To convert his libertine friends, he looked for arguments in their favourite authors: in Michel de Montaigne, in the Skeptic Pierre Charron, in the Epicurean Pierre Gassendi, and in Thomas Hobbes, an English political philosopher. For Pascal, Skepticism was but a stage. Modernist theologians in particular have tried to make use of his main contention, that "man is infinitely more than man," in isolation from his other contention, that man's wretchedness is explicable only as the effect of a Fall, about which a man can learn what he needs to know from history. In so doing, they sacrifice the second part of the *Apologie* to the first, keeping the philosophy while losing the exegesis. For Pascal as for St. Paul, Jesus Christ is the second Adam, inconceivable without the first.

Finally, too, Pascal expressly admitted that his psychological analyses were not by themselves sufficient to exclude a "philosophy of the absurd"; to do so, it is necessary to have recourse to the convergence of these analyses with the "lines of fact" concerning revelation, this convergence being too extraordinary not to appear as the work of providence to an anguished seeker after truth (*qui cherche en gémissant*).

He was next again involved in scientific work. First, the "Messieurs de Port-Royal" themselves asked for his help in composing the *Élements de géométrie*; and second, it was suggested that he should publish what he had discovered about cycloid curves, a subject on which the greatest mathematicians of the time had been working. Once more fame aroused in him feelings of self-esteem; but from February 1659, illness brought him back to his former frame of mind, and he composed the "prayer for conversion" that the English clergymen Charles and John Wesley, who founded the Methodist Church, were later to regard so highly. Scarcely capable of regular work, he henceforth gave himself over to helping the poor and to the ascetic and devotional life. He took part intermittently, however, in the disputes to which the "Formulary"—a document condemning five propositions of Jansenism that, at the demand of the church authorities, had to be signed before a person could receive the sacraments—gave rise. Finally a difference of opinion with the theologians of Port-Royal led him to withdraw from controversy, though he did not sever his relations with them.

Pascal died on August 19, 1662, after suffering terrible pain, probably from carcinomatous meningitis following a malignant ulcer of the stomach. He was assisted by a non-Jansenist parish priest who found him to be "most submissive to the Sovereign Pontiff and to the Church."

Assessment. At once a physicist, a mathematician, an eloquent publicist in the *Provinciales*, and an inspired artist in the *Apologie* and in his private notes, Pascal was embarrassed by the very abundance of his talents. It has been suggested that it was his too concrete turn of mind that prevented his discovering the infinitesimal calculus; and in some of the *Provinciales* the mysterious relations of human beings with God are treated as if they were a geometrical problem. But these considerations are far outweighed by the profit that he drew from the multiplicity of his gifts; his religious writings are rigorous because of his scientific training; and his love of the concrete emerges no less from the stream of quotations in the *Provinciales* than from his determination to reject the vigorous method of attack that he had used so effectively in his *Apologie*.

MAJOR WORKS

MATHEMATICS, LOGIC, AND THE FOUNDATIONS OF SCIENCE: *Essai pour les coniques* (1640); *Lettre sur le sujet de la machine inventée par le sieur B.P. pour faire toutes sortes d'opération d'arithmétique* (1645); *De l'autorité en matière de philosophie*, the first editor's title of Pascal's preface to a projected *Traité du vuide*, consisting of a general statement of the principle of scientific research (written 1647, printed 1779); *Traité du triangle arithmétique avec quelques autres petits traités sur la même matière* (written 1654, printed 1665); *De l'esprit géométrique—de l'art de persuader* (written *c.* 1658, first printed in two parts 1728 and 1776);

Histoire de la roulette, appellée autrement trochoïde ou cycloïde (1658); *Lettres de A. Dettonville, contenant quelques-unes de ses inventions de géométrie* (1658–59).

PHYSICS: *Expériences nouvelles touchant le vuide* (1647); *Récit de la grande expérience de l'équilibre des liqueurs* (1648); *Traites de l'équilibre des liqueurs et de la pesanteur de la masse de l'air* (written 1651, printed 1663; *The Physical Treatises of Pascal: The Equilibrium of Liquids and the Weight of the Mass of the Air*, 1937).

RELIGIOUS PHILOSOPHY AND CONTROVERSY: *Abrégé de la vie de Jésus-Christ* (written 1654 or 1655, printed 1846); *Lettre escritte à un Provincial par un de ses amis sûr le sujet des disputes présentes de la Sorbonne* (January 1656, followed by 17 more pamphlets on the same themes down to March 1657), all 18 being subsequently republished together as *Les Provinciales* (1657; augmented edition, with supplementary pamphlets, 1659); *Projet de mandement contre l'Apologie pour les casuistes* (written 1658, printed 1779); *Ecrits sur la grâce* (four treatises drafted between 1656 and 1658 and printed from 1779 onward); *Pensées de M. Pascal sur la religion et sur quelques autres sujets* (1670; the first Eng. trans. was *Monsieur Pascal's Thoughts, Meditations, and Prayers*, 1688; numerous new translations and versions have appeared since then); *Prière pur demander à Dieu le bon usages des maladies* (written 1659, printed 1666 and 1670).

OTHER WORKS: *Trois discours sur la condition des grands* (comprised 1660, printed 1670). Collected editions of Pascal's works also include several private letters of mathematical or spiritual interest.

COLLECTED EDITIONS: Considerable editions of Pascal's *Oeuvres complètes* were undertaken by the Abbé Charles Bossut, 5 vol. (1779); and by Léon Brunschvicg, Pierre Boutroux, and Félix Gazier, 14 vol. (1908-25). The handiest edition is by LOUIS LAFUMA, *l'Intégrale* (1960). Collections in English include *The Miscellaneous Writings of Pascal* (1849); *Thoughts, Letters, Minor Works* (1910); and *Provincial Letters, Pensées, Scientific Treatises*, Encyclopædia Britannica "Great Books of the Western World" (1952).

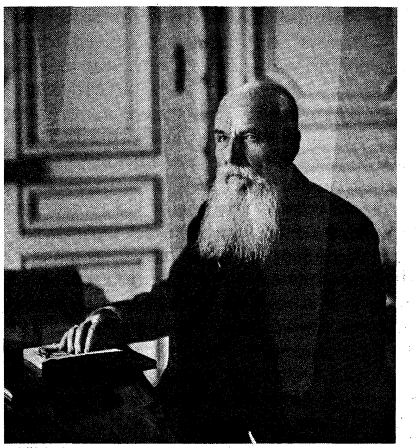
BIBLIOGRAPHY. The standard edition of Pascal's works is that by LEON BRUNSCHVICG, PIERRE BOUTROUX, and FELIX GAZIER, 14 vol. (1908–25); in part replaced by the edition of JEAN MESNARD, 2 vol. (1964–72); and by *Pensées sur la religion*, ed. by LOUIS LAFUMA, 3rd ed., 3 vol. (1960); also editions by H.F. STEWART of *Pensées* (1950; with Eng. trans.) and of *Provinciales* (1920, reissued 1951). Biographical and critical studies include: JEAN MESNARD, *Pascal, l'homme et l'oeuvre* (1951; Eng. trans., *Pascal: His Life and Works*, 1952), an excellent biography, with bibliography; *Pascal* (1965; Eng. trans., 1969); and *Pascal et les Roannez*, 2 vol. (1965); LUCIEN JERPHAGNON, *Le Caractère de Pascal* (1962); PHILIPPE SELLIER, *Pascal et saint Augustin* (1970); EMILE CAILLIET, *The Clue to Pascal* (1943, reprinted 1970); and *Blaise Pascal, l'homme et l'oeuvre*, "Cahiers de Royaumont" (1956), a collection of articles and discussions from the proceedings of a congress on Pascal held at Royaumont in 1954.

(J.Or./L.J.)

Pašić, Nikola

A statesman of remarkable foresight and tenacity, Nikola Pašić was one of the founders of the Kingdom of the Serbs, Croats, and Slovenes (later Yugoslavia). The leader of Serbia's Radical Party from 1881, he played a major role in politics under both the Obrenović and the Karageorgević dynasties. The Balkan Wars (1912–13) and World War I (1914–18) brought his ideas of Pan-Serb nationalism to fruition; but his insistence on Serbian hegemony in the multinational state that he brought into being sowed seeds of domestic dissension that were greatly appeased by the federalization of Yugoslavia, long after his death.

Early career. Nikola Pašić was born on December 19 (old style; December 31, new style), 1845, into a commercial family of modest means at Zaječar. He studied engineering in Belgrade, then graduated from the Zürich Polytechnikum, where his interest in contemporary liberalism and democratic institutions was stimulated by the Russian anarchist Mikhail Bakunin. Returning to Serbia (1873), he joined the Socialist group led by Svetozar Marković and, as editor of the newspaper *Oslobodjenje* ("Liberation"), became an important exponent of Marković's views. But participation in the wars against Turkey (1876–78) led Pašić to revise his political thinking, and, having concluded that King Milan Obrenović's oligarchy was depriving Serbia both of progressive leadership and of national perspective, Pašić decided to enter politics actively. Elected to parliament in 1878, he worked, as leader of the opposition, against the authoritarian monarchy in an endeavour to establish a parliamentary democracy. He also helped to found the Radical Party (1881).



Pašić.
H. Roger-Viollet

When a popular rising instigated against Milan's government by the Radicals in Zaječar (1883) led to further repression and to the severe punishment of many Radical leaders, Pašić was forced to flee through Austria to Bulgaria. Condemned to death in his absence, he plotted with other émigré leaders to unite Serbia and Bulgaria in a war of liberation. After Milan's abdication in favour of his son Alexander (1889), Pašić returned to Serbia with both his political convictions and his determination much strengthened. He was elected president of the Skupština (Parliament) and, on two occasions, mayor of Belgrade. Pašić was appointed premier for the first time from February 1891 to August 1892 and as foreign minister accompanied the new king, Alexander Obrenović, on a state visit to Russia (1892), where he established firm personal and political ties with the tsarist regime. He became Serbian minister to St. Petersburg (Leningrad) in 1893 but resigned in protest at former king Milan's recall to Serbia (1894).

The attempt on Milan's life in 1899 resulted in trumped-up charges of regicide being brought against the members of the Radical Party. Pašić, who was among those sentenced to death and who was unaware of the forceful Russian interventions being made on his behalf, won himself an amnesty by making a humiliating confession in which he implicated many of his former colleagues. He then left the country voluntarily to return only when Milan had finally withdrawn.

The Karageorgević restoration

When the Karageorgević dynasty, in the person of King Peter I, was restored by the bloody coup d'etat of 1903, Pašić, as leader of the Radical Party, concentrated his efforts on establishing the party both as the backbone of the new regime and as the moving force in Serbian politics. The methods by which he strove to preserve his leadership often seemed ruthless and opportunistic, for he never hesitated to adapt his policies or to sacrifice his personal relationships to the political necessity of the moment. (This led, not unnaturally, to acrimonious conflicts with the party in opposition and to schisms within his own.) From December 1904 to May 1905 he served as premier and as minister for foreign affairs—displaying great skill by counteracting Austria-Hungary's attempts to impose a tariff war on Serbia. He held both posts again from May 1906 to June 1908 but was occupying only a minor governmental position in October 1908, when Austria-Hungary's annexation of the Turkish provinces of Bosnia and Hercegovina provoked a major international crisis. He was reappointed premier in October 1909, only to be replaced in 1911 by Milan Milovanović, his greatest political rival. Though he cooperated with Milovanović in concluding a pact with Bulgaria—from which was eventually to develop the Balkan League,

whose aim was war against Turkey—younger politicians and many military leaders continually conspired to remove him from his position as party leader, and in 1912 his imminent dismissal was avoided only by Milovanović's sudden death. Thenceforth reinstated as premier and minister for foreign affairs, Pašić led Serbia through two victorious wars: the first against Turkey (1912) and the second against Bulgaria (1913).

Despite his increased prestige, a further attempt was made to oust Pašić from office when the opposition parties joined the military leaders in an attempt to overthrow the ruthless Radical-controlled administration in Macedonia, and it was only as a result of direct Russian intervention that King Peter was prevailed upon, albeit grudgingly, to grant Pašić a mandate for new elections. The accession as regent of Prince Alexander (King Peter's younger son) on June 24, 1914, gave Pašić some support, and his position was further confirmed when the threat of war with Austria-Hungary prevented the elections from being held.

Leadership during World War I. After the murder of Archduke Francis Ferdinand by a Serb nationalist at Sarajevo on June 28, 1914, Pašić was most compliant in dealing with the formidable terms of Austria-Hungary's ultimatum to Serbia but was nevertheless unable to avert the declaration of war on June 28. World War I initially silenced Serbian political discord: parliament, which had already been dissolved, reassembled at Niš, and in November 1914 a coalition government under Pašić's premiership was formed. In this situation Pašić shrewdly played on heightened nationalistic sentiments in order to transform Serbia's war into his own dream of liberating all the Serbs—and even other South Slavs—in Austria-Hungary. He strongly opposed the secret Treaty of London, whereby Russia, France, and Great Britain promised much South Slav territory to Italy.

Withdrawal to Corfu

The Austro-German conquest of Serbia forced Pašić's government and the army to withdraw from Serbia to Corfu (winter 1915). When in 1917 the government executed leaders of the Black Hand, a clandestine military organization, on charges of plotting the assassination of the regent and of seeking actively to participate in politics, a lasting schism in Serbian politics was created; but, though the coalition government broke up, Pašić continued to govern with a homogeneous Radical Cabinet. When his position was still further weakened by the fall of Russia's tsarist regime (1917), he was obliged temporarily to abandon his strict Pan-Serbian attitude and to negotiate on equal terms with Ante Trumbić's Yugoslav Committee, a body of South Slav exiles from Austria-Hungary with its seats in London and Paris. The resulting Corfu Declaration (July 1917) laid down the broad lines for a postwar unified Yugoslav state.

Postwar career. As World War I neared its end, Pašić began to claim that Serbia, as the dominant political and military force among the South Slavs, had the exclusive right to speak on the Allied side on their behalf. As a result, the Triple Entente governments withheld from the Yugoslavs the recognition that they accorded to the much less favourably situated Czechoslovaks in August 1918. On November 9, 1918, at Geneva, Pašić, under pressure from the Serbian opposition and from the Allied governments, joined delegates from the Yugoslav Committee, from the National Council recently formed in Zagreb, and from the Serbian opposition in signing a declaration that provisionally envisaged a Yugoslavia in which the Serbian government (Belgrade) should share power with the representatives of Austria-Hungary's former South Slav subjects (Zagreb). The Serbian government, which Pašić himself had secretly dissuaded, rejected the declaration. As a result, when Austria-Hungary collapsed, the Allies were unable to agree on a solution for the relations of the South Slavs with Serbia, while Italy reasserted its territorial claims under the Treaty of London. Despite the danger, Pašić persevered in his obstructive tactics toward the Yugoslav Committee and to the National Council in Zagreb.

Nevertheless, an uneasy compromise was finally achieved when Serbia and the South Slav provinces were united on December 1, 1918, as the Kingdom of the Serbs, Croats, and Slovenes. Though he was denied the premiership of the kingdom, Pašić went with Trumbić and Vesnić as one of the new state's delegates to the Peace Conference at Versailles (1919).

Pašić failed fully to comprehend the fateful difference between Serbia's homogeneity and the complexity of the new kingdom, which comprised several nations, each with its own distinct historical development and cultural identity. Ignoring requests for individual recognition from Croats, Slovenes, Macedonians, and Bosnian Muslims, he continued to regard them simply as Serbs—albeit Serbs of three religions and several names. When, therefore, he was reappointed premier in 1921 he immediately pushed through parliament a unitary constitution that, under the guise of establishing a homogeneous nation, actually confirmed the existing Serbian hegemony and, by abolishing historic and autonomous provinces, established a strongly centralized regime under a powerful monarchy. He eliminated the Democrats from the government (winter 1921) and formed an entirely Radical Cabinet. He failed to secure a majority in the elections of March 1923 but stayed safely in office, thanks to blunders by the opposition. Though from July to October 1924 he had to give way to a coalition government under Ljubomir Davidović, by adroit interparty manoeuvring he was able immediately afterward to return to power much stronger than before. His relations with King Alexander and with the Anticentralist Croats and Slovenes nevertheless became increasingly strained. In February 1925 Pašić was forced to dissolve parliament, but by adopting drastic measures—among them the imprisonment of Stjepan Radić and other Croatian Peasant Party leaders—he secured a small working majority. A temporary political collaboration with Radić later the same year failed to produce a stable government; and, when Radić publicly criticized the still-increasing tendency toward centralization and unification, Pašić had to resign in March 1926. He died on December 10, 1926, at Belgrade, three weeks before his 81st birthday.

The Serbian hegemony

BIBLIOGRAPHY. CARLO SFORZA, *Pachitch et l'union des Yougoslaves* (1938; Eng. trans., *Fifty Years of War and Diplomacy in the Balkans: Pashich and the Union of the Yugoslavs,* 1940), the only known biography of Pašić in English, written by a prominent Italian diplomat and author who was Pašić's personal friend; H. WICKHAM STEED, *Through Thirty Years, 1892–1922* (1924), memoirs of an outstanding British political observer and writer who cooperated with R.E. Seton-Watson to promote the Yugoslav cause during World War I; PAVLE D. OSTOVIC, *The Truth About Yugoslavia* (1952), an informative work on Yugoslav politics by an observer who served as the secretary of the Yugoslav Committee from 1914 to 1918.

(A.S.Pa.)

Passage Rites

Passage rites are ceremonial events, existing in all historically known societies, that mark the passage from one social or religious status to another. This article describes these rites among various societies throughout the world, giving greatest attention to the most common types of rites, and discusses their purposes from the viewpoints of the people observing the rites, and their social, cultural, and psychological significance as seen by scholars seeking to gain an understanding of human behaviour.

NATURE AND SIGNIFICANCE

Many of the most important and common rites of passage are connected with the biological crises of life—birth, maturity, reproduction, and death—all of which bring changes in social status and, therefore, in the social relations of the people concerned. Other rites of passage celebrate changes that are wholly cultural, such as initiation into societies composed of people with special interests—for example, fraternities. Rites of passage are universal, and presumptive evidence from archaeology in the form of burial finds strongly suggests that they go back to very early times. The worldwide distribution of these rites long ago attracted the attention of scholars, but the first substantial interpretation of them as a class of phenomena

was presented in 1909 by the French anthropologist and folklorist Arnold van Gennep (1873–1957), who coined the name rites of passage. Van Gennep saw the rites as means by which individuals are eased, without social disruption, through the difficulties of transition from one social role to another. On the basis of an extensive survey of preliterate and literate societies, van Gennep held that the rites consist of three distinguishable, consecutive elements, called in French *séparation, marge,* and *agrégation,* which may be translated as separation, transition, and reincorporation, or as preliminal, liminal, and postliminal stages (before, at, and past the threshold). The person (or persons) on whom the rites centre is first symbolically severed from his old status, then undergoes adjustment to the new status during the period of transition, and is finally reincorporated in society in his new social status. Although the most commonly observed rites relate to crises in the life cycle, van Gennep saw the significance of the ceremonies as being social or cultural, celebrating important events that are primarily sociocultural or man-made rather than biological. The British anthropologist A.M. Hocart (1884–1939) held that the passage from one status to another was the result rather than the cause of these ceremonies; thus the rites both induced and allayed personal and social stress rather than merely allaying it. Basing his views on circumstances in a few ancient civilizations, Hocart thought that all rites of passage were based on the model of ritual of investiture of kings, in which symbolic killing and rebirth of the new ruler, and sometimes actual killing of the old, were required. Later scholarship has shown that symbolic death and rebirth into the new status are common forms of symbolism in rites of passage of various kinds and that the symbolic killing and rebirth of rulers is therefore not appropriately viewed as the prototype of all rites.

Arnold van Gennep's theory

FUNCTIONS

Modern scholars in the social sciences characteristically accept the views of van Gennep about the social and psychological significance of rites of passage; that is, passage rites are seen to have positive value for the individual in relieving stress at times when great rearrangements in his life occur, such as are brought by coming of age, entering marriage, becoming a parent, or at the death of a close relative, and in providing instruction in and approval of his new roles. The rites are seen also to be socially supporting in various ways. Such support includes roles of the rites in preventing social disruption by relieving the psychological stress of the individuals concerned; providing clear instruction to all members of societies to continue life in normal fashion with new social alignments, the affirmation they provide of social and moral values expressed and thus sanctioned as part of the ceremonies; and the social unity they foster by joint acts and joint expression of social values. During most of man's history, rites of passage have generally been religious events; that is, they have been conducted in a religious framework and regarded as religious acts and hence possessed special authority. From the viewpoint of modern social science, however, their nature is generally seen as being fundamentally secular. Mankind gives social attention to all events regarded as being socially important. Until recent times, religion was intimately connected with most aspects of life, and events of such social importance as the changes in society that the rites celebrate were most frequently incorporated in the system of religious belief and act. The tendency of recent decades toward secularization of rites of passage strongly suggests that the primary significance of most rites is social or secular rather than religious. In the modern, scientifically minded nations of the world many rites of passage, such as rites of initiation into fraternal and honorary societies are wholly secular; others have only small elements of religion, and even marriage may be a wholly secular rite.

Rites as entertainment

One of the primary functions of rites of passage that is often overlooked by interpreters, perhaps because it appears obvious, is the role of the rites in providing entertainment. Passage rites and other religious events have in the past been the primary socially approved means of participating in pleasurable activities, and religion has been a primary vehicle for art, music, song, dance, and other forms of aesthetics.

From its beginning, the study of the significance of rites as a class of phenomena has attempted to account for similarities and differences in the rites among societies of the world. The similarities are striking and doubtless reflect the close similarity in ways of human thought. Modern attempts to account for similarities and differences have generally given little attention to and reached no consensus concerning the nature of the innate psychological factors involved in the genesis of the rites. Attempts to understand rites of passage have instead generally been sociocultural interpretations that view the rites as part of an integrated sociocultural system, the man-made part of human life. Religion and rites of passage are thus seen as elements in this system that affect and are affected by other elements in the system such as means of gaining a livelihood and the manner in which society is aligned in groups. Most modern analysts accordingly have interpreted both differences and similarities in rites of passage on the basis of their sociocultural context. The inventive and symbolic capabilities of mankind are treated as a constant factor, and analytic attention is given to differences and similarities in the sociocultural contexts in which rites are found. In attempting to understand why marriage is an extremely elaborate rite in one society and very simple in another society, for example, scholars have looked to the social order and to the manner of gaining a livelihood to judge the relative importance of the enduring unions of spouses. Following the view that culture, including the social order, composes a coherent, inclusive system, modern scholarship has, in short, most commonly interpreted rites of passage in terms of their functional significance in the social system.

The significance or stated goals of rites of passage as these exist in the minds of the actors are regarded as quite inadequate for gaining an understanding of the functional significance of the rites. Very often, rites of passage are said to have goals such as dispatching the spirit of a dead person to another world, protecting the newborn, the new adult, and the newlywed from evil influences. Often the explicit goals of the rites have been forgotten and their continuation is a matter of following tradition, so that means have become goals. Although scholars have noted the explicit goals of these rites, they have characteristically given greatest emphasis to inferring functional significances that are not obvious to the actors in the rites. In so doing they have broadened their investigations from observations of the symbolism of rites to include prominently all of the behaviour during the rites and their social contexts, learning the social identities of the performers and their relationships to other performers and the entire society.

CLASSIFICATIONS OF RITES

No scheme of classification of passage rites has met with general acceptance, although many names have been given to distinguishable types of rites and to elements of rites. The name purification ceremonies, for example, refers to an element of ritual that is very common in rites of passage and also in other kinds of religious events. In most instances, the manifest goal of purification is to prepare the individual for communication with the supernatural, but purification in rites of passage may also be seen to have the symbolic significance of erasing an old status in preparation for a new one (see also PURIFICATION RITES).

Other names that have been given to passage rites often overlap. Life-cycle ceremonies and crisis rites are usually synonymous terms referring to rites connected with the biological crises of life, but some modern scholars have included among crisis rites ritual observances aimed at curing serious illnesses. Ceremonies of social transformation and of religious transformation (for both see this section, below) overlap and similarly overlap crisis rites. Religious transformations, such as baptism and rites of ordination, always involve social transformations; social transformations such as at coming-of-age and induction

into office may also bring new religious statuses, and life-cycle ceremonies similarly may or may not involve changes in religious statuses. It is nevertheless sometimes useful to distinguish the various rites by these names.

Life-cycle ceremonies. Life-cycle ceremonies are found in all societies, although their relative importance varies. The ritual counterparts of the biological crises of the life cycle are the numerous kinds of rites celebrating childbirth, ranging from baby "showers" to pregnancy rites to rites observed at the actual time of childbirth and, as exemplified by Baptism and the fading Christian rite of Churching of Women, a ceremony of thanksgiving for mothers soon after childbirth. These rites involve the parents as well as the child and in some societies include the couvade, which in its so-called classic form centres ritual attention at childbirth upon the father rather than the mother. At this time the father follows elaborate rules of ritual procedure that may include taking to bed, simulating labour pains, and symbolically enacting the successful birth of a child. In all societies some ritual observances surround childbirth, marriage, and death, although the degree of elaboration of the rites varies greatly even among societies of comparable levels of cultural development. Rites at coming-of-age are the most variable in time in the life span and may be present or absent. In some societies such rites are observed for only one sex, are elaborate for one sex and simple for the other, or are not observed for either sex. Characteristically, rites at coming-of-age are not generally observed in the modern industrial civilizations or, as in the Jewish Bar Mitzwa and the Protestant confirmation of the United States, exist today more or less as vestiges of formerly important religious rites. Among the elaborate civilizations of Asia, rites at coming-of-age have similarly waned in recent times. The elaborate rites observed a century ago in Japan when young men and young women reached social maturity are only rarely observed today and are virtually unknown to the general population. Death is given social attention in all societies, and the observances are generally religious in intent and import. In societies that fear dead bodies the deceased may be abandoned, but they are nevertheless the focus of ritual attention. Most commonly, rites at death are elaborate, and they include clearly all of the stages of separation, transition, and reincorporation first noted by van Gennep (see also DEATH RITES AND CUSTOMS).

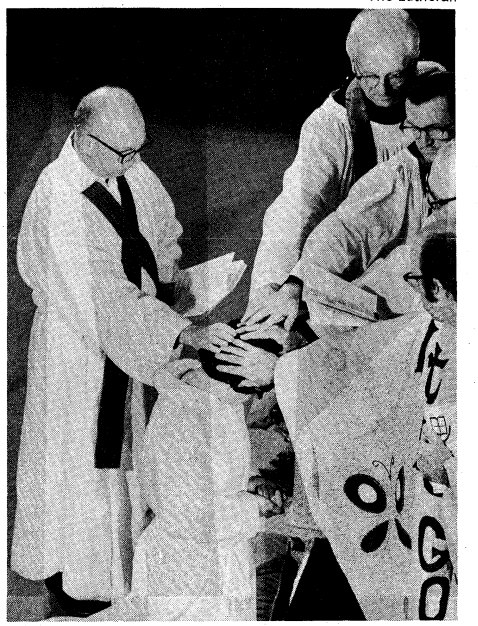
The couvade

Ceremonies of social transformation. Ceremonies of social transformation include all of the life-cycle ceremonies, since these involve social transitions for the subjects of the ritual and also for other persons. When a man or woman dies, for example, he assumes a new social role as a spirit that may be socially important to the living; the bereaved spouse becomes a widow or widower; and the children have an unnamed but changed status as lacking one parent. A vast number of rites of social transformation, such as rites of initiation into common-interest societies, have no direct or primary connection with biological changes. These are abundant in the United States and European nations, usually as secular ceremonies. In primitive societies, rites of this kind mark induction into age-graded societies, principally limited to males, and a variety of common-interest societies such as warrior societies, curing societies (special groups whose purpose is to cure illnesses), and graded men's societies that are hierarchically ranked in prestige. Whether hereditary or achieved by appointment or election, assumption of important office in various kinds of societies is often observed by elaborate ritual. Any other events involving changes in social status tend to become the subjects of institutionalized ritual, which is then a prerequisite for the new status. Common examples are initiation ceremonies of college fraternities, sororities, and honorary societies; adult fraternal societies, and social groups of other kinds centred on common interests. Other social changes of importance that apply to a substantial number of people but do not involve initiation into organized social groups are also given ritual attention. Common among these are graduation exercises, festivities marking retirement from work, and various kinds of award ceremonies.

Ceremonies of religious transformation. Religious-transformation ceremonies signal changes in religious statuses, which may be matters of the greatest importance to the people. Performing ritual such as making sacrifices and offerings may be required in the normal course of life, and these acts may be regarded as conferring a new religious status or state of grace. Sacrifices are a frequent feature of rites of passage, and for important ceremonies such as coronations and funerals of rulers, have sometimes required the sacrifice of many human beings (see also SACRIFICE). Among the laity, entry into a religious society or the assumption of any other new religious role is customarily an event celebrated by rites such as those of baptism and confirmation. Among professional religious personnel, the achievement of any distinct status of specialization is ordinarily observed by rites corresponding to the Christian rites of ordination—the rites through which religious functionaries become entitled to exercise their respective functions. As with other rites of passage, these rites may be simple or complex, and their degree of complexity may generally be easily seen as reflecting the religious and social importance of the newly acquired status. A single element of an elaborate rite in one society, such as circumcision or the dressing of the hair in a distinctive way, may in another society be the central or sole event of rites of either social or religious transformation. These ceremonies may, accordingly, be called rites of circumcision or be identified by the name of the style of hairdress.

Sacrifices

The Lutheran



Ordination of a Lutheran pastor by the laying on of hands.

Other ceremonies. The term rites of passage is occasionally applied to institutionalized rites for curing serious illness and, rarely, to cyclic ceremonies such as harvest festivals. No new social or religious status is ordinarily gained by recovery from illness or participation in harvest rites, however, and these ceremonies have probably been included among the rites of passage because of similarities in their ritual procedures. In some societies, recovery from a very critical illness is regarded as a divine sign that the erstwhile invalid should assume the role of a religious specialist, but rites of ordination are quite separate. Some elements of ceremonies pertaining to changes in the seasons may be seen as incorporating acts of separation and incorporation, symbolically saying goodbye to the old season and welcoming the new, but these are not customarily called rites of passage. Although clearly denoting a change in social status, divorce has rarely been regarded as a rite of passage. Festive observances at this time are perhaps common in some societies, but they are often informal practices of the individual or simple acts

of local custom, such as discarding wedding rings, that are not institutionalized in the entire society. The absence of divorce from the conventional roster of rites of passage illustrates an outstanding characteristic of this class of rites: all celebrate events that are either socially approved or, like death and illness, unavoidable. Rites of passage that signal the assumption of social statuses disapproved by society are both out of keeping with the prevailing interpretation of the rites as being socially supportive and would broaden them to cover such events as trials by jury and commitment to prison for serious crimes.

Symbolic aspects of ceremonies. Whatever their subclassification, elaborate rites of passage are commonly rich in symbolism that prominently includes representations of the states of separation and transition and, especially, insignia of the new status. Most common among these markers of new status are alterations and embellishments of visible or invisible parts of the body, distinctive garments and bodily decorations, and insignias corresponding to symbols of office. All parts of the body that may be altered or embellished without ordinarily causing serious disability have served as the symbols of social statuses and have been elements of rites of passage. Outstanding among these insignias are special styles of hairdress, clothing, and ornaments; the filing, staining, and removal of teeth; the wearing of ornaments in pierced ears, noses, or lips; tattoos and, among dark-skinned people upon whom tattoos would not be visible, their counterpart of scarification, which produces designs in relief; and circumcision or other genital operations (see also RELIGIOUS DRESS AND VESTMENTS).

Insignia and body decorations

Several motifs or themes of symbolism commonly recur among societies widely separated from each other geographically and culturally. One such theme symbolizes death and rebirth into the new status. Initiates may be ceremonially killed and then made symbolically to act like infants who, during the rites, are made to mature into their new statuses. Another common form of symbolism makes use of doors or other portals that signify entry into the new social domain. Ordeals are a rather common feature of coming-of-age ceremonies for both males and females, and they are also used in rites of initiation into men's societies of various kinds. Success in passing the ordeals is customary and signifies mastery of the roles that are to be assumed.

A universal feature of rites of passage is the proscription of certain kinds of ordinary behaviour. Sexual continence is a common rule, as is the prohibition of ordinary work such as farming, hunting, and fishing. Many rites prohibit certain behaviour or prescribe the reverse of ordinary behaviour. Among Indians of the western United States, for example, a taboo against scratching the body with the fingers was common during ritual periods. In other societies, ritual behaviour required that the subjects of ritual sit in a remarkable fashion, wear articles of clothing inside out or backward, or wear the clothing of the opposite sex. These acts all may be seen as dramatizations, by contrast, of the events that they celebrate, thereby making them memorable.

A representative example. Rites of passage marking very important events customarily include all of the three stages described by van Gennep. A representative example is afforded by the traditional rites surrounding childbirth as these were commonly observed in Japan until recent years. Observances began when a woman learned she was pregnant. Partly for stated reasons of promoting health and partly for supernaturalistic reasons, she thenceforth abstained from certain foods and ate others. During the fifth month of pregnancy she donned a special girdle, ordinarily procured from a Buddhist temple and supernaturally blessed. Relatives offered prayers for the well-being of the woman and her child. When birth seemed imminent, she was isolated from all other persons except the women who attended her and remained in isolation for a fixed number of days after parturition. This period was most commonly 33 days, which was divided into stages preceeding from severe restriction of her acts to final complete resumption of all normal activities. She had at first to follow a number of special rules of diet and could not perform normal household tasks. During the period of isolation, the mother was regarded as polluted from the flow of blood during childbirth and therefore dangerous to other people and dangerous or offensive to supernatural beings of the Shintō religious pantheon. She could not make the usual offerings or say prayers before the household shrines to Shintō gods or have any other kind of contact with them. To avoid offending the sun goddess, her clothing and that of her child when laundered could never be hung in direct sunlight to dry but instead were placed in the shadows of the eaves of the house. For the same reason, she covered her head with a cloth when she stepped outside the house near the end of the period of isolation. Water and cloths used in washing the mother after parturition were considered to be polluted and were buried in the ground beneath the floor of the room of confinement. After a fixed number of days passed, the mother was permitted to resume bathing and again perform some but not all of her ordinary work in the house. Other restrictions on behaviour were removed at fixed times, and when the full period passed, the mother and her female aides performed a ceremony of purification by sprinkling salt on the mother and on the floors of the dwelling. The beginning of a new, normal period free from pollution also was symbolized by kindling a new fire in the household cooking stove. Now ready to return to normal life, the mother ate a ceremonial meal with other members of the family and resumed ordinary relationships with supernatural beings and other human members of the community.

Japanese passage rites

PASSAGE RITES IN THE CONTEXT OF THE SOCIAL SYSTEM

Most of the scholarly interpretations of rites of passage of the 20th century have considered their relation to the social system and have seen the functional significance of the rites as a contribution to the maintenance of society as a system of congruent parts. Explicit or implicit in this line of reasoning is the idea of equilibrium found in any scientific theory concerned with systems. For the system to operate effectively, its elements must be mutually supportive or congruous, and the system is then described as being in a state of equilibrium. Social systems embrace a fixed number of people and a fixed number of roles. Changes in either the number of the people or in the proportions of statuses disturb social equilibrium. When a child is born, a new member is added to society; the social behaviour and statuses of its parents change, and these changes also affect other members of society. Other social changes that are the subjects of passage rites similarly disrupt the state of social equilibrium. Rites of passage are seen to foster the development of a new state of equilibrium in adjustment to the social changes upon which the rites focus. By means of the rites, members of society are informed of the new social circumstances and at the same time give social approval to them. Individuals upon whom the rites focus are assured of success in their new roles by the ritual observances and are given psychological reassurance in a number of other ways. They and all other members of society are instructed by the ritual enactment of their new social relations to return to normal behaviour incorporating the added or lost personnel and the added, lost, or changed social statuses. The same general kind of reasoning is applied to various other religious ceremonies. The anthropologists Eliot D. Chapple and Carleton S. Coon interpret all rites of passage and other group rites as "rites of intensification." Calling special attention to the ritual depiction of habitual relationships for the statuses involved, Chapple and Coon state that this behaviour "has the effect of reinforcing or intensifying their habitual relations, and thus serves to maintain their conditioned response . . . In the technical (physiological) sense, the performance of these rites prevents the extinction of habits . . . to which the individual has been trained."

Changes of status and social equilibrium

Closely related to the function of passage rites in restoring social equilibrium, in the anthropologists' interpretation, are a group of additional effects or functions, some of which apply first to the individuals whose statuses

change and, through their behaviour, to the entire social group. Other functional effects apply directly to the entire society. By allaying the anxiety of individuals who are undergoing change, social disruption is avoided. Rites of passage characteristically give assurance of mastery of the new roles and often include instruction in the new roles. In the many societies in which statuses and roles are clearly distinguished by sex, the rites symbolically emphasize these differences, thereby instructing the initiates and aiding them in sexual identification. The anxiety and potential social disruption caused by death and the grief of the bereaved are similarly held in check. Funeral rites customarily point up grief and then firmly instruct the bereaved to resume normal behaviour that is not disruptive to others. The joint performance of rites and the joint expression of moral and other social values that are included among ritual acts may be seen as directly promoting group solidarity through communion with one's fellows and affirmation or reaffirmation of rules and ideals that foster social harmony.

Rites of passage and all other group rites are seen to be socially supporting in still another implicit way. The joint rites are customarily a rehearsal or dramatization, with supernatural sanction, of a part or all of the social order of the society. Relatives have special roles that are congruent with, or enactments of, their positions in normal social life, and the entire social hierarchy may be on display during the rites through the assignment of ritual roles. Thus statuses of kinship, caste, social equality, and hierarchy are all seen to be reinforced by dramatic presentation of them.

Accepting this group of interpretations of the social significance of rites of passage, anthropologists have also attempted to understand variations in the degree of elaboration of rites of passage among societies of the world. A fundamental assumption is the commonplace idea that the greater the importance of a social change the greater the ritual attention will be. The birth, marriage, and death of a ruler obviously are more important to the entire society than these events in the life of a commoner. The importance of such events is not always obvious, however, and their relative importance is often difficult to see when different societies are compared. Rites of marriage, for example, may be very simple or very elaborate in different societies of the same economic base and comparable levels of cultural development. Recourse to consideration of features of the social order has allowed a reasonable explanation of the differences. Marriage rites in matrilineal societies, for example, which are organized into subgroups primarily upon a principle of descent through female lines only, tend to be simple, and divorce in these societies is also simple. Marriage rites in patrilineal societies (in which descent is through male lines), however, tend to be elaborate, and divorce initiated by females is difficult.

Marriage rites in matrilineal and patrilineal societies

In matrilineal societies, the social core is composed of groups of male and female relatives united by female lines, which are economically distinct from other groups and self-sufficient. Where the matrilineal principle of organization is strong, the role of the husband and father, who belongs to a matrilineal group different from that of his wife, is not that of economic provider for his wife and children. Instead, he is the economic mainstay for his sister and her children, and his contact with his wife may be limited to spending nights with her. The brothers or other male relatives of a mother not only provide economically for her children but also assume what is elsewhere the role of the father in socializing children. Enduring unions of marriage are not vital to such matrilineal societies. If marriages end in divorce, the matrilineal ordering of society assures approved social identification, economic support, and affective ties for the children and their mother and also assures continuance of the society as long as males are available as procreators. In patrilineal societies, however, the role of the mother, who is the outsider in the group, is vital for the birth and rearing of the children, and she and her children are dependent upon her husband for economic support. Strong sanctions are placed upon marriages in these societies to help ensure lasting unions. Marriage ceremonies are correspondingly elaborate, often involving the transfer of property, which among some African societies is called marriage insurance for the reason that it must be returned if the marriage falls asunder.

In societies such as those of the United States and European nations, where the important unit of kinship is ordinarily limited to the nuclear family of parents and children and where important social affiliation does not depend upon descent through one sex of progenitors, enduring unions of marriage are also vitally important. Rites of passage at marriage traditionally have been required by law as well as by the church, and many other sanctions on lasting marriages are imposed by laws concerning divorce, communal property, and the care of children. The bride and groom who have undergone the whole series of traditional rites of passage from engagement parties to the religious ceremony may reasonably be seen as more firmly married than couples united by a simple civil ceremony (see also SOCIAL DIFFERENTIATION AND STRATIFICATION; CASTE SYSTEMS; KINSHIP).

PSYCHOLOGICAL ASPECTS OF PASSAGE RITES

Less scholarly attention has been given to psychological than to social or cultural aspects of rites of passage, in large part because the scholars concerned with such rites in world societies have been principally anthropologists, who lean toward sociocultural interpretations. As the foregoing discussion of passage rites in social context illustrates, psychological aspects of rites nevertheless enter strongly if often implicitly into anthropological interpretations as fundamental matters in social solidarity and social disorder. Emotional ties to kin and other members of society, personal identification with social groups and religious statuses, and commitment to religious ideology and other values are reinforced and sometimes created by rites of passage. In a realistic sense, the rites serve as blueprints for social relations and religious behaviour that make clear the acceptable ways to act and at the same time point up and reinforce affective relations with other people and with the supernatural. Familial rites of ancestor worship, for example, are not only reinforcements of familial solidarity but also have psychological value in reinforcing emotional ties among relatives.

Psychological interpretations of passage rites have given greatest emphasis to their value in allaying personal anxiety. A recurrent feature of the rites are acts of magic that assure that the outcome of the endeavour will be successful. In the words of the anthropologist Bronisław Malinowski these acts serve symbolically and psychologically "to bridge over the dangerous gaps in every important pursuit or critical situation" that exist because of man's lack of control of the universe. By such magical means as miniature boats floated in streams or carried away by the tide, the dead are shown symbolically to go successfully to the other world, and childbirth and successful maturation are similarly depicted magically. The subjects of rites of passage frequently act out their future roles to the approval of all others. Numerous acts of magic that are not essential to changes in social status may be incorporated in rites of passage and may be seen to give psychological assurance relating to the future life of the individual. Traditional Japanese practices at childbirth, for example, required that when a girl was born, the placenta be buried in the ground outside the entrance to the dwelling to insure that the girl, when mature, marry in normal fashion and leave the family. When a boy was born, the placenta was buried inside the house to ensure that he remain at home when mature. The ordeal that a young man or young woman must often undergo during rites of coming-of-age may similarly be seen to provide psychological assurance of success in the new status. Ordeals of this kind are characteristically uncomfortable or frightening, but they are events that any human being ordinarily can endure.

Psychotherapeutic value of rites

The psychotherapeutic value of passage rites surrounding events in which stress may be acute, such as childbirth, death, and serious illness, is clearly apparent and essentially follows the principles of modern secular psy-

chotherapy. The subject is made the centre of concentrated attention by many people, is given reassuring evidence of their regard for him, and, by means of magic and the intervention of supernatural beings, is assured of a successful outcome. These events are carried out on a high emotional pitch, which gives them added force. When anxiety is induced by religious beliefs themselves, such as by ideas that if ritual acts are not performed calamitous results will follow, the rites of passage may be said both to create and to allay anxiety.

Where particular social statuses have special honour and prestige, the mere existence of these statuses offers opportunities for gaining psychological satisfaction, and the requirements for gaining these statuses serve to guide behaviour in socially approved channels that offer psychological satisfaction.

Other interpretations of psychological aspects of passage rites have relied upon ideas derived from or inspired by the psychoanalyst Sigmund Freud (1856–1939). These have sometimes concerned the symbolism involved in the rites and, in anthropological interpretations, have dealt with both Freudian ideas of symbols and the social order. The psychologist Bruno Bettelheim has interpreted cicatrization (inducement of scars) of males in rites at coming-of-age as symbolic wounds indicating subconscious male envy of the vagina, the counterpart of Freud's idea of penis envy. A psychologically oriented anthropologist J.M. Whiting, and others have combined sociological and psychoanalytic theories in attempting to explain why male initiation ceremonies are conducted in some societies and not in others. Harsh rites, sometimes including genital operations, are held to be correlated with societies in which infant males have long and intimate contact with their mothers, and husbands are prohibited from sexual intercourse with their wives for a period of two years or more. The long and exclusive relationship between mother and son is assumed to lead to strong emotional dependence upon the mother by the son, which becomes potentially disruptive at the time the son reaches puberty. The harsh rites are seen to break the bond of dependency and avoid potential social disruption that might otherwise result from discord between son and father at this time.

PASSAGE RITES IN THE CONTEXT OF THE RELIGIOUS SYSTEM

Certain passage rites represent first and foremost transformations in the religious statuses or circumstances of the people concerned. As already noted, rites of passage are customary upon the assumption of a new status as a religious professional. During most of man's history, however, rites of passage have carried among their implications a change to a new religious state for the ordinary members of society as well as for the professional religious person. Among the culturally advanced societies of the world with orders of priests, ideas of the significance of symbolism in passage rites may be elaborate and sophisticated, representing the rites as different states of grace or, as in Hinduism, cyclic states involving death and rebirth. In many societies, one is not fully or properly a human being until he has undergone the rites of passage appropriate for his age and sex. In some societies, fully human status is not reached until the rite of Baptism has been performed, and children who die before that time may be interred with special rites in places separate from those of the dead who have been baptized. When passage rites are religious ceremonies, as has generally been the circumstance until modern times, some state of sacrament or divine blessing, vaguely or clearly defined, is entailed. At the time of death, rites of passage placing the deceased in the realm of the supernatural customarily have been required. Symbolism in many rites of passage denotes communion with the supernatural. In common with many other kinds of religious events, then, passage rites relate the individual and the society to the sacred world, conferring benefit upon him thereby (see also SACRED OR HOLY).

Rites of passage frequently have ethical import of value for the maintenance of social equilibrium. Where ethical or moral codes and religious beliefs are intimately connected or identified as one and the same, as in Christianity, Judaism and Islām, the role of religious beliefs and acts may be seen to have strong value as social sanctions since the moral injunctions apply to human relations as well as to man's relations with the supernatural world. All societies have moral or ethical codes, rules of what is appropriate and inappropriate in human relations, and these are enforced by various means. Rites of passage, as noted above, commonly incorporate statements or dramatizations of moral values, and rites at coming-of-age often give moral instruction in highly explicit terms. No necessary or inherent connection exists, however, between morality and religious beliefs. Any serious breach of proper moral conduct results in the imposition of a network of sanctions, many of them secular. In some societies, religious beliefs have little bearing on morality in relations with one's fellow men, although violations of rules applying to relations with supernatural beings and supernatural forces may be regarded as bringing inevitable punishment or misfortune through the supernatural agency. Whenever morality is a part of religious precepts, the direct sanctioning force of passage rites stressing moral rules may be powerful and important to the maintenance of society. In other societies, the ethical import of passage rites and other features of religion may operate less directly. An example is provided by societies that revere but do not deify ancestors. Any breach of morality reflects unfavourably upon the ancestors, who may undertake no action of censure but nevertheless serve as a sanctioning force that is reinforced by death rites.

Morality and religious beliefs

PRIMARY PASSAGE RITES

In simple, primitive societies dependent for subsistence upon hunting and gathering, in which social groups are small and specialization in labour is limited to distinctions by sex and age, no social statuses may exist except those of child, adult, male, female, and disembodied spirit. Among primitive societies somewhat more advanced technologically and culturally, however, specialized groups based upon common interests appear, and these customarily require rites of induction or initiation. In culturally sophisticated societies, with elaborate divisions of labour, social statuses of leadership and specialized occupation are multiple. If all societies of the world, preliterate and literate, are considered, the most commonly recurrent rites of passage are those connected with the normal but critical events in the human life span—birth, attainment of physical maturity, mating and reproduction, and death.

Birth rites. Rites surrounding the birth of a child are often a complex of distinct rituals that prescribe different behaviour on the part of the mother, the father, other relatives, nonfamilial members of the society, and with respect to the newborn. Observances may begin when pregnancy is first noted and may continue until the time of delivery, when the full rite of passage is observed, and for a variable period of time afterward. In many simple societies and in European societies of the past, the expectant mother is isolated from other members of society at this time for the stated reason that the blood that flows during childbirth has inherently harmful qualities. Where this belief is strong, the classic couvade may be practiced. Regions of the world in which this practice was formerly common include the Amazon Basin of aboriginal South America, Corsica, Spain, among the Basques of France and Spain, and among various societies of Asia. Old ethnological writings have created the impression that ritual attention is limited entirely to the father. Later investigations have made it appear doubtful that the mother in any society is free from ritual requirements. In many societies, rites that have been called the couvade are observed by both parents. The anthropologist Alfred L. Kroeber (1876–1960) reported that among most of the many tribes of aboriginal California, rites at childbirth were much alike for both mother and father. To prevent harm to their child and to other people during the ritual period, the parents observed food taboos; ate in seclusion; avoided contact with other people; did as little work as possible; and refrained from various other acts of ordinary

Isolation of expectant mothers

behaviour that included cooking, touching tools, and eating salt, meat, and fish. Women often were under injunctions to scratch themselves only with a stick or a bone for fear that their nails at this time would leave permanent scars on their bodies.

Practices of sympathetic and contagious magic relating to birth and the later well-being of both child and mother are abundant and diverse. Among Indians of aboriginal British Columbia, the mother inserted a smooth beachstone, an eel, or other slippery object under her garment at the neckline, permitting it to slide to the ground to symbolize and insure quick and successful childbirth. In societies of Southeast Asia and Indonesia, religious specialists dressed as women simulated successful delivery. Rites directed toward the newborn similarly symbolize or ensure health and well-being and, after some days, weeks, or months have passed, often include Baptism or other ritual acts that introduce the child to supernatural beings. Both child and mother are often regarded as being defenseless at this time, and many ritual acts have the purpose of protecting them from harmful supernatural beings and forces. In Southeast Asia and Indonesia, a practice called mother roasting, which requires that the mother be placed for some days over or near a fire, appears once to have had the goal of protecting the mother from such evil influences. This practice survives today in altered form in the rural Philippines, where it is regarded as having therapeutic value.

Mother roasting

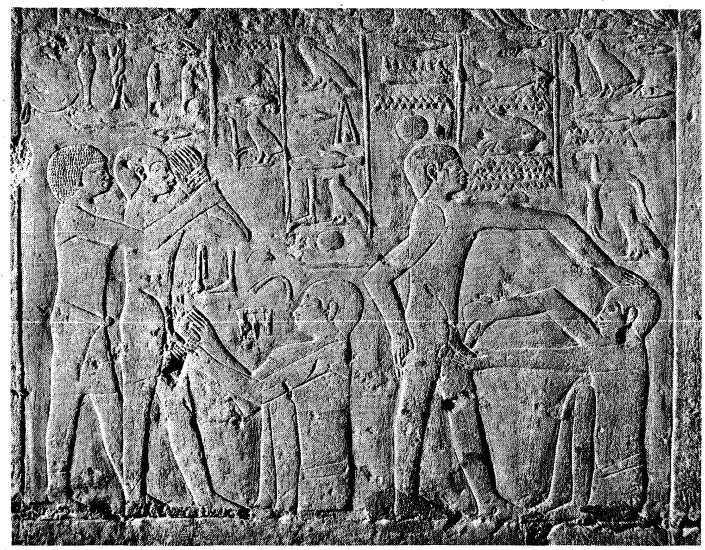
Native explanations of the ritual procedures at childbirth reflect beliefs of a mystic affinity between parent and child, and many of the prescriptions have the manifest goals of preventing harm to the infant until it is able to fend for itself. Among South American Indians practicing the classic couvade, this belief of affinity between father and child relates to the soul, which is not fully transmitted to the child until the end of the ritual period.

In addition to the social (communal) and psychological significances of birth rites already noted, scholars have offered interpretations of these ceremonies as reinforcing familial ties. The classic couvade has been seen by Malinowski as a sympathetic symbolic stressing of the relationship between the husband and the wife and her kin, which is instituted when the child is born. In addition to serving as a means of allaying husbandly anxiety over the welfare of the wife, the practices of the couvade establish social paternity, which, in turn, promotes familial and societal solidarity.

Initiation rites. The most prevalent of rites of initiation among societies of the world are those observed at coming-of-age. These have frequently been called puberty rites, but, as van Gennep argued long ago, this name is inappropriate. Puberty among females is often defined as the time of onset of menses (the menstrual flow), but no such clearly identifiable point exists in the sexual maturation of males. Moreover, the age at which rites of attaining maturity are observed vary greatly from society to society, going far beyond the normal range of years at which sexual maturity is attained. The definition of maturation is thus seen to be largely social or cultural rather than solely biological.

The full range of stages of passage rites is often followed in rituals at coming-of-age. Ordeals or other tests of manhood and womanhood are also common. Some of these practices in preliterate societies seem incomprehensible or absurd until their nature as evidence of qualification for the new social statuses is understood. Among the Bemba tribe of Africa, for example, girls were required to catch water insects with their mouths and to kill a tethered chicken by sitting on its head. Circumcision or other genital operations are also a fairly common feature of rites celebrating the attainment of maturity. Although most commonly applying to males, genital operations are performed on females in a few societies. It seems quite clear that circumcision and other alterations of the sexual organs have not until modern times been regarded as therapeutic surgery. These operations may have psychological significance following Freudian lines of interpretation, but it seems clear that they are also significant as insignia of social status. Where circumcision is the practice for male initiates, the uncircumcised male is not a full-fledged adult. It may be remembered that at this time other parts of the body are also modified, by incision, piercing, filing, tattooing, and by other kinds of practices that are not painful. Circumcision may in fact have no direct relation to the attainment of sexual maturity. In native Samoa, boys were circumcised at any age from 3 to 20.

Ordeals and other practices



Egyptian circumcision relief, tomb of Ankhmahor, Saqqārah, 6th dynasty (2345–2181 BC).
Henri Stierlin—Ziolo

An outstanding feature of rites at coming-of-age, which is generally less prominent or absent from other rites of passage, is their emphasis upon instruction in behaviour appropriate to the status of adults. Instruction in dress, speech, deportment, and morality may be given over a period of months. Very commonly, instruction is first given at this time in matters of religion that have heretofore been kept secret, and initiates may at this time be expected or required to commune with the supernatural, sometimes by means of revelatory trances induced by fasting, violent physical exertion, or the consumption of plant substances that produce hallucinations or otherwise alter the sensibilities.

Separation of male initiates from their mothers and all other females is also common, and ritual events may dramatize the transition from a world of women and children to one that is ideally male. Symbolism of these rites dramatizes the separation in such ways as by requiring the young men temporarily to wear the clothing of women and by rigid exclusion of all females from participation in the rites.

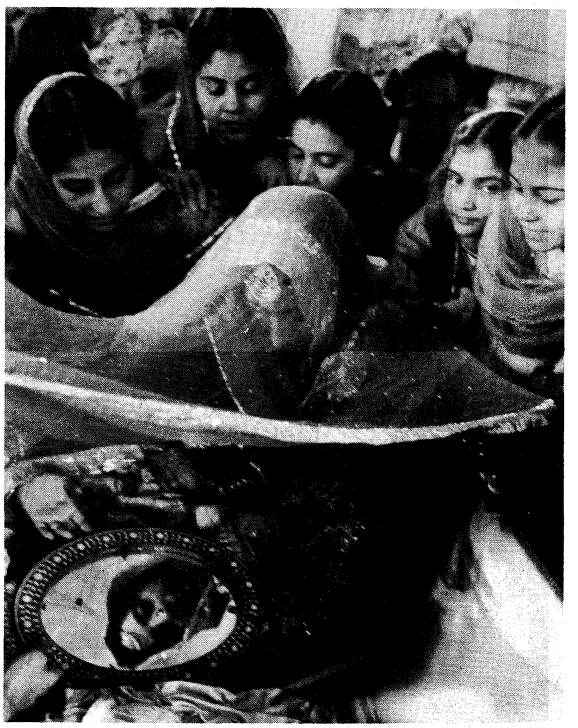
Among the technologically and scientifically advanced societies of the world, initiation rites have become increasingly secular. The great religions of the world all included rites at coming-of-age, but for much of the modern population of these nations, the rites are either not observed or are simpler vestiges of the old religious ceremonies. The most common rites of initiation are predominantly or wholly secular ceremonies conducted to celebrate such events as entry into a common-interest association or graduation from school. Rites of initiation such as into age-graded groups or common-interest societies follow essentially the same pattern as those at coming-of-age, and their simplicity or elaboration may be correlated with the importance of the new statuses.

As seen by social analysts, the significance of initiation rites of all kinds is the same as that of other rites of passage. Some emphasis is given to their didactic value and to their significance in sex-role identification. One question that has not been answered is why rites at coming-of-age are so poorly developed today among the technologically advanced societies of the world. Many factors, including changed views of the nature of the universe and changed social conditions, appear to have contributed to the decline of rites of passage. The super-

Rites at coming-of-age in modern times

naturalism traditionally present in the rites is no longer acceptable to many people, and in the United States and parts of Europe the association of adult status with sexual maturity as expressed in the term puberty rites has been unwelcome, a matter that should be excluded from notice rather than celebrated. Probably far more important in discouraging the rites has been the extreme variation in these nations in the age of social maturity. In the United States, the ages differ at which one may legally drive a car, enter marriage, own and control property, buy alcoholic drinks and tobacco, enter military service, and vote; and in some of these matters the ages differ from state to state. The demands of modern civilization have, moreover, lengthened the age of social maturity, the time at which one is an economically productive member of society, and, dependent upon the number of years of formal education, have produced great variation in the age of full social maturity. The social and psychological value of rites of coming-of-age in making the transition to adulthood seem substantial, but it seems certain that modern cultural circumstances are incompatible with the conduct of such rites.

Frank Horvat—Black Star



Hindu wedding custom; the bride regards her reflection in a mirror.

Marriage rites. It is assumed by anthropologists that marriage is one of the earliest social institutions invented by man, and, as already noted, rites of marriage are observed in every historically known society. These rites vary from extremes of elaboration to utmost simplicity, and they may be secular events or religious ceremonies. Subclasses of rites of marriage, named and unnamed, exist in many societies, beginning with ceremonies of betrothal that require complex formalities of transfer and exchange of goods, which are often regarded as compensation to the bride's kin group for their loss of the bride. Ceremonies of dramatic, sham "capture" of the bride by the groom and his relatives and friends have been common in both preliterate and literate societies. Marriage in these societies is seen by social analysts as a cooperative liaison between two different groups of kin, between which some feelings of hostility exist. Ceremonies of token capture are conducted even when betrothal and all other arrangements for marriage have long been completed to the expressed satisfaction of both sides, and the sham captures are interpreted as socially sanctioned channels or the expression and relief of feelings of hostility between the two kin groups. In some historically known societies of Africa such sham battles between kin of brides and grooms may occur, with full societal approval, for years after a marriage during any kind of religious rite.

Sham capture of brides

Like rites at coming-of-age, ceremonies at marriage have often included clearly visible insignia of the new social status, in such forms as wedding rings, distinctive hair dress and garments, and tattoos, ornaments, or other embellishments that are regarded also as being decorative. Traditionally, preliminary rites have often provided instruction in the wifely role. Such instruction might be informal or conducted as a part of ritual. Rites of marriage proper also often give instruction through mimicry, dancing, and other symbolic acts that dramatically depict the woman's proper role in society, expressing her economic and social obligations and privileges with reference to her children, husband, other relatives, and still other members of society. Tests of maturity and rites with the purpose of promoting fertility have also commonly been included.

In addition to sharing the functional significances of other passage rites, marriage ceremonies may be seen especially to stress social bonds between husband and wife and their kin groups. In most societies and during most of human history, romantic love has not been the means by which spouses are selected. Convention, often strongly sanctioned, has limited marriage to only certain classes of people. Mutual attraction between the spouses has been a matter of little or no importance. The importance of marriage with respect to spouses, children, other kin, and the orderly maintenance of society is readily inferrable. Rites of marriage place a sanction on unions of marriage that may be very powerful and thus serve as both a means of conducting an orderly and satisfying human life and also as sanctions for the orderly maintenance of society. A general correlation may be seen between the degree of elaboration of marriage rites and the social importance of enduring marriages in the society in question. Where, as in some of the large, industrial nations of the world, marriage rites are simple and sometimes secular, a host of other sanctions operate similarly to foster lasting unions.

Death rites. All human societies have beliefs in souls or spirits and an afterlife, and all conduct rituals when people die. The earliest archaeological evidence of rites of passage comes in the form of deliberate burials of Neanderthal man, interments that included ornament shells and stone implements. Like other passage rites, ceremonies at death vary greatly but have much in common. Attention centres upon the deceased and upon the bereaved. The body of the deceased must be disposed of. This is most commonly done by interment, although other practices are also followed, such as cremation, abandonment, natural decomposition and later burying of the bones, and mummification. Special symbolic attention is given to allow or induce the soul of the deceased to enter the spirit world, and in societies where the souls of the dead are feared as being potentially malevolent, special ritual measures are taken to safeguard the living. The dead remain indefinitely as members of society, and where reverence for or worship of ancestors is a religious practice, the social affinity with deceased is close.

The bereaved

Pointed ritual attention is given to the bereaved, who at this time must observe many special rules of behaviour. In many societies, funerals are joyous rather than solemn affairs, but where lighthearted festivities are conducted they customarily appear as a final stage of the rite of passage. The bereaved are everywhere in some degree isolated from other members of society and prohibited from many normal activities. These and other features of the rites point up the transition the bereaved is undergoing, dramatize grief at the climax of the ceremonies, and by final ceremonies symbolically instruct the bereaved to set excessive grief behind and return to normal social life in the new role. Climaxes of grief are dramatically portrayed by such events as gashing the body, wailing and weeping, and the use of hired mourners, all conducted on signal for fixed intervals. A joyous celebration may fol-

low, involving song, dance, feasting, and other pleasurable activities, in which the bereaved prominently participates and is thereby reincorporated into the normal social system.

More than other passage rites, ceremonies at death are closely identified with religion, and many of the acts of these rites have explicit religious goals. Prominent among these is the assurance of a continuation of existence in an afterlife, in which the deceased in some measure continue to be members of living societies (see also DEATH RITES AND CUSTOMS).

BIBLIOGRAPHY. ARNOLD VAN GENNEP, *Les rites de passage* (1909; Eng. trans., 1960), the pioneering study and standard work on passage rites; D.M. SCHNEIDER and K. GOUGH (eds.), *Matrilineal Kinship* (1961), a discussion of the relevant features of social organization, especially matrilineal kinship, but also, in comparison, patrilineal kinship; BRUNO BETTELHEIM, *Symbolic Wounds* (1954), a Freudian inspired work interpreting the significance of ritual acts of circumcision and other genital operations; E.D. CHAPPLE and C.S. COON, *Principles of Anthropology* (1942), interesting and useful information on social interaction, societal equilibrium and disruption, and the role of rites of passage in restoring equilibrium; J.G. FRAZER, *The Golden Bough*, 3rd ed., 12 vol. (1911–20), a classic work that contains much descriptive information, often scattered, on rites of passage and many other features of religion; A.M. HOCART, *Social Origins* (1954), an interesting interpretive work although somewhat dated; FRANK W. YOUNG, *Initiation Ceremonies* (1965), concerns rites at coming-of-age, interpreting their significance in relation to the roles in society of males and females and the manner in which society is organized into social groups.

(E.N.)

Passeriformes

The order Passeriformes, the passerines, or perching birds, is the dominant avian group on Earth today. Considered the most highly evolved of all birds, passerines have undergone an explosive evolutionary radiation in relatively recent geological time and now occur in abundance on all continents except Antarctica and on most oceanic islands. Their rapid evolution and adaptation to virtually all terrestrial environments have resulted in a large number of species, some 5,100, compared to only about 3,500 species for all other birds.

The order Passeriformes is divided by most taxonomists into four suborders: Eurylaimi, Tyranni, Menurae, and Passeres. The first three, containing about 1,100 species, are considered more primitive and are often grouped informally as the "suboscines" for convenient comparison with the very large fourth suborder, the oscines, or songbirds (about 4,000 species).

GENERAL FEATURES

Size range and structural diversity. Passerines are small to medium-sized land birds, ranging from about 7.5 to about 117 centimetres (three to 46 inches) in overall length. Among the tiniest species are the New World flycatchers (Tyrannidae), New Zealand wrens (Xenicidae), titmice (Paridae), flowerpeckers (Dicaeidae), tanagers (Thraupidae), and waxbills (Estrildidae). The heaviest are the lyrebirds (Menuridae) of Australia and the ravens (*Corvus*). The longest species, the ribbon-tailed bird of paradise (*Astrapia mayeri*), is actually not so large in body bulk but has extremely long tail feathers. Most passerine species fall within the range of about 12.5 to 20 centimetres (five to eight inches) in length and from 15 to 30 grams (0.5 to one ounce) in weight. A house sparrow (*Passer domesticus*), for example, is 12 to 15 centimetres (five to six inches) long and weighs about 26 grams (0.9 ounce); a cardinal (*Cardinalis cardinalis*) 20 to 23 centimetres (eight to nine inches) and approximately 44 grams (1.5 ounces).

Passerines have evolved a great diversity of feeding adaptations. The majority are insectivorous, at least at certain times of their lives. Members of the order have evolved many ways for finding insect food: swallows (Hirundinidae) are aerial feeders; New World flycatchers "hawk" insects by flying out from a perch; vireos (Vireonidae) glean insects from small twigs and foliage; woodcreepers (Dendrocolaptidae), nuthatches (Sittidae), and creepers (Certhiidae) search for insects in crevices in tree bark; and many other species pick and scratch on the ground and in leaf litter. More specialized passerines eat aquatic insects (dippers: Cinclidae), fish (some New World flycatchers: Tyrannidae), fruit (cotingas: Cotingidae; and many others), leaves (plantcutters: Phytotomidae), nectar (sunbirds: Nectariniidae), small land vertebrates (shrikes: Laniidae), and seeds (finches and many others). For these different food habits, various structural specializations have developed, especially in the bill and feet (see below *Form and function*).

Importance to man. *Aesthetic and economic importance.* Since prehistoric times people have enjoyed watching and listening to songbirds. The almost infinite variety of colours, patterns, behavioral traits, songs, and calls found in these birds appeals to man's aesthetic tastes. As objects of beauty and interest passerines have been incorporated into human culture, folklore, poetry, music, sculpture, and painting. Songbirds have also been used as symbols, such as the European goldfinch (*Carduelis carduelis*), which represented the Passion of Christ in Renaissance art, or the raven (*Corvus corax*), which sometimes has signified a messenger for the devil, an evil omen.

Songbirds as religious symbols

Passerines are widely kept as cage birds. The origins of this practice are lost in antiquity, but it is known that by the 5th century BC the Greeks kept a variety of songbirds, including finches, nightingales and other thrushes, magpies (*Pica*), and starlings (*Sturnidae*). Canaries (*Serinus canaria*) were brought to Europe from their native Canary Islands in the 16th century and have since been developed by domestication and breeding into many varieties. Other passerines now widely kept as pets are the cardueline and estrildine finches and the starlings (particularly Asian mynahs, *Gracula*). The magnitude of the cage-bird "fancy" is indicated by importation statistics on wild and semidomestic birds: in one recent year alone, over 420,000 passerines (excluding canaries) were legally imported into the United States as cage birds, a number far exceeding that of parrots, the only other bird group whose members are commonly kept as pets. Many countries, including the United States and Great Britain, prohibit the capture and sale of nearly all native songbirds.

Songbirds are also economically important in other ways. Although seldom considered food in economically advanced areas, they are nonetheless important dietary items in many rural or heavily populated countries. China, Japan, and other Oriental countries, for instance, have highly developed techniques for catching small birds, and in cities such as Hong Kong and Tokyo passerines are commonly sold in food markets. In Italy, France, and Belgium the capture of migratory songbirds for the pot or for cage birds is still extensive. Laws against such activities are difficult to enact or enforce in areas in which the habit has become part of the culture; hence, for example, more than 10,000,000 passerines were still being caught annually in Italy in the late 1960s.

Fortunately, killing songbirds for their feathers is no longer as prevalent as it once was. Until the early 20th century, however, there were few protective laws, and the wearing of birds (especially on women's hats) was common. In 1886 a young ornithologist reported that he had counted feathers from no fewer than 40 bird species, including 22 kinds of passerines, on hats seen on two afternoon walks in a fashionable part of New York City.

Other cultures have used songbird feathers for personal adornment, but usually for men rather than women. This practice often came about not only for the beauty of the feathers themselves but also because the feathers were used as symbols of such bird qualities as speed and aggressiveness. Most noteworthy are the feathers of male birds of paradise (Paradiseidae) used as headdresses by tribesmen of New Guinea. An estimated 80,000 adult birds are still being killed annually for this purpose. Other ancient uses of passerine feathers have now largely been terminated, either because the birds are extinct (in the case of Hawaiian feather cloaks) or because more suitable modern substitutes have been found (Melanesian feather money).

black-and-yellow broadbill
Eurylaimus ochromalus
Eurylaimidae

straight-billed woodcreeper
Xiphorhynchus picus
Dendrocolaptidae

pale-legged ovenbird (hornero) with nest
Furnarius leucopus
Furnariidae

pearled treerunner
Margarornis squamiger
Furnariidae

black-crested antshrike
Sakesphorus canadensis
Formicariidae

rufous antpitta
Grallaria rufula
Formicariidae

purple-breasted cotinga
Cotinga cotinga
Cotingidae

saffron-crested tyrant-manakin
Neopelma chrysocephalum
Pipridae

blue-backed manakin
Chiroxiphia pareola
Pipridae

yellow-billed tit-tyrant
Anairetes flavirostris
Tyrannidae

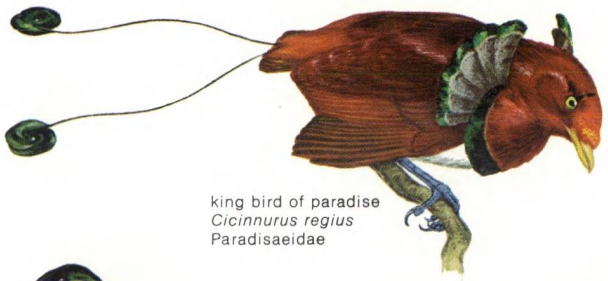
d'Orbigny's chat-tyrant
Ochthoeca oenanthoides
Tyrannidae

rusty-backed antwren
Formicivora rufa
Formicariidae

suiriri flycatcher
Suiriri suiriri
Tyrannidae

rusty-margined flycatcher
Myiozetetes cayanensis
Tyrannidae

chucao tapaculo
Scelorchilus rubecula
Rhinocryptidae



king bird of paradise
Cicinnurus regius
Paradisaeidae

Albert lyrebird (1 metre long)
Menura alberti
Menuridae

golden swallow
Kalochelidon euchrysea
Hirundinidae

bronzed drongo
Dicrurus aeneus
Dicruridae

western black-headed oriole
Oriolus brachyrhynchus
Oriolidae

rufous treecreeper
Climacteris rufa
Climacteridae

green jay
Cyanocorax yncas
Corvidae

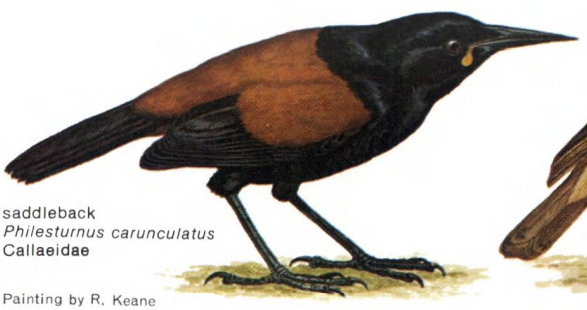
black-backed butcherbird
Cracticus mentalis
Cracticidae

magnificent bird of paradise
Diphyllodes magnificus
Paradisaeidae

Nepal treecreeper
Certhia nipalensis
Certhiidae

red-headed tit
Aegithalos concinnus
Paridae

orange-winged sitella
Neositta chrysoptera
Sittidae



saddleback
Philesturnus carunculatus
Callaeidae

flappet lark
Mirafra rufocinnamomea
Alaudidae

lowland rail-babbler
Ptilorrhoa caerulescens
Timaliidae

Painting by R. Keane

varied triller
Lalage leucomela
Campephagidae

brown-eared bulbul
Hypsipetes flavala
Pycnonotidae

Philippine fairy bluebird
Irena cyanogaster
Irenidae

black-capped mockingthrush
Donacobius atricapillus
Mimidae

golden-breasted starling
Cosmopsarus regius
Sturnidae

black-bellied wren
Thryothorus fasciatoventris
Troglodytidae

Ruppell's robin chat
Cossypha semirufa
Turdidae

rufous-bellied niltava
Muscicapa sundara
Muscicapidae

golden-faced pachycare
Pachycare flarogrisea
Pachycephalidae

long-tailed silky flycatcher
Ptilogonys caudatus
Ptilogonatidae

white-breasted wood swallow
Artamus leucorhynchus
Artamidae

black-faced flycatcher
Monarcha frater
Muscicapidae

northern brubru
Nilaus afer
Laniidae

Fulleborn's longclaw
Macronyx fulleborni
Motacillidae

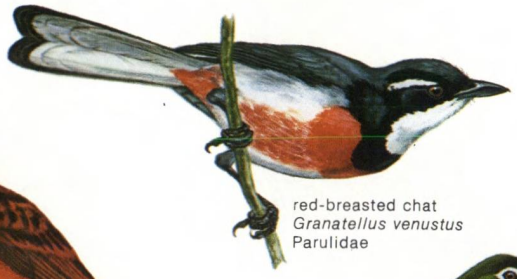
four-coloured bush shrike
Telophorus quadricolor
Laniidae

cardinal honeyeater
Myzomela cardinalis
Meliphagidae

golden-winged sunbird
Nectarina reichenowi
Nectariniidae

yellow oriole
Icterus nigrogularis
Icteridae

tit berrypecker
Oreocharis arfaki
Dicaeidae



red-breasted chat
Granatellus venustus
Parulidae

scarlet finch
Haematospiza sipahi
Carduelidae

palila
Loxioides bailleui
Drepanididae

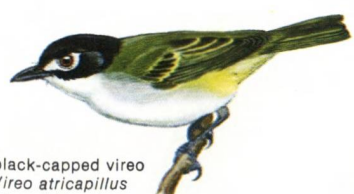
black-fronted white-eye
Zosterops atrifrons
Zosteropidae

golden-crowned tanager
Iridosornis rufivertex
Thraupidae

paradise whydah
Vidua paradisaea
Ploceidae

violet-eared waxbill
Uraeginthus granatina
Estrildidae

black-chested sparrow
Aimophila humeralis
Fringillidae



black-capped vireo
Vireo atricapillus
Vireonidae

slaty-capped shrike-vireo
Smaragdolanius leucotis
Vireolaniidae

green honeycreeper
Chlorophanes spiza
Thraupidae

black-winged red bishop
Euplectes hordeaceus
Ploceidae

Painting by R. Keane

Songbirds as agricultural pests

Some passerines, on the other hand, are serious economic pests. In areas in which one-crop agriculture is extensive, certain bird species have undergone population explosions because of almost unlimited food availability; in turn, their crop depredations can be serious. One example of this is in Africa, where immense flocks of a small weaver, the red-billed quelea, or Sudan dioch (*Quelea quelea*), numbering as many as 20,000,000 birds in one flock, do millions of dollars worth of damage to various small grain crops each year. Other serious pests are the Java sparrow (*Padda oryzivora*) in Asian rice fields and mixed flocks of New World blackbirds (*Icteridae*) and European starlings (*Sturnus vulgaris*) in grainfields in the United States. The same starling and the house sparrow, both introduced to the United States from Europe, have become urban pests by fouling buildings with excrement and blocking rain gutters and ventilators with their nests. Starlings occasionally also have been implicated in accidents; in 1960 a flock at the airport in Boston was sucked into a jet's engines and the resultant crash killed 61 people.

Ecological importance. The greatest importance of passerines is ecological. As the dominant form of birdlife in virtually all terrestrial environments, the perching birds are a major component of the world's ecosystems. They consume great quantities and varieties of food—grains, fruits, insects and other invertebrates, small amphibians and reptiles, and even small mammals—and in turn serve as food for other animals; they act as hosts for parasites and are occasionally parasitic themselves; they both propagate and distribute plants by pollinating flowers and carrying viable seeds to new locations; and they have the mobility (through migration) to utilize habitats that are available only at certain times of the year. A few aspects of the ecological impact of passerines are known, but, until the science of ecology has advanced, the true magnitude of their importance cannot be evaluated with precision.

Drawing by R. Keane

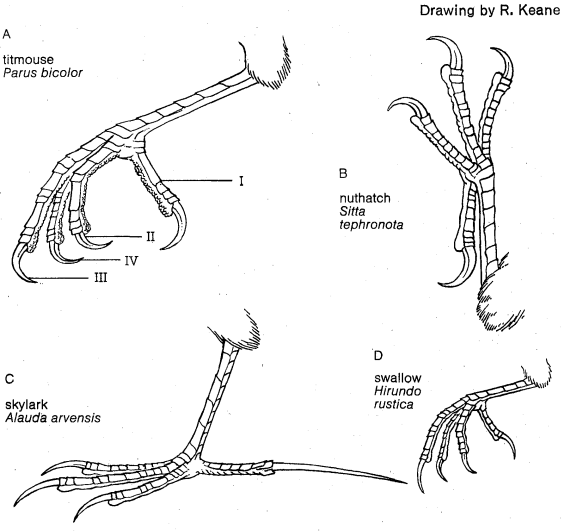


Figure 1: Modifications of the foot of perching birds, for (A) perching and clambering; (B) bark climbing; (C) ground walking; (D) perching only (weak foot). Right feet are shown.

NATURAL HISTORY

Types of territories

Reproduction. *Territoriality and courtship.* The breeding behaviour of passerines is diverse. Most species are solitary nesters, a single monogamous pair of birds maintaining a territory that is large enough to support all their activities during the breeding season: courtship, mating, nesting, and food gathering. Others have similar territories, but the birds forage outside the defended area for most of their food (*e.g.*, the North American red-winged blackbird, *Agelaius phoeniceus*). Still others are colonial nesters, defending only the nest site and a small area immediately adjacent to it. Some species build individual nests close together in a colony (oropendolas, Icteridae; some swallows; the house sparrow), and others construct massive communal nests in which the breeding pair defends only its own nest cavity (palm chat, *Dulus*; several weaverbirds, Ploceidae). In a few species polygynous (polygamous) males establish special display territories (leks) for courtship and mating in which no nesting takes place. In these courtship arenas the males, usually brilliantly coloured, attract females through song and posturing and sometimes by dancing, manipulation of objects, and other elaborate displays. The best known arena-displaying males are the cocks of the rock (*Rupicola*), manakins (Pipridae), birds of paradise, and bower-birds (Ptilonorhynchidae). After mating in or near the lek, a female leaves to build a nest and raise the young without assistance from the male. Still other species build no nest at all, but are brood parasites (some cowbirds, Icteridae; whydahs, Ploceidae); the female lays her eggs in the nests of other (usually smaller) species, and the young are raised entirely by the foster parents.

Nesting. Nest sites are varied: they include holes in the ground, trees, banks, and rock crevices; on ledges; on the surface of the ground; within the larger nests of other species (including nonpasserines) or near wasp nests (presumably for the protection the wasps afford); and in a wide variety of vegetation—grasses, shrubs, and trees.

Passerine nests are usually elaborately constructed and may contain many different kinds of materials: mud, grasses, hair and feathers, strips of bark, plant fibres and downs, rootlets, twigs and sticks, leaves, string, spider webs, cast snake skins, lichens, and many other substances. Most species build open nests, usually cup-shaped. Others form domed or ball-shaped closed nests, with an entrance at the side (occasionally at the top or bottom). One of the most famous closed nests is that of the South American ovenbirds of the genus *Furnarius* (Furnariidae), whose name derives from its thick-walled mud "oven" nest, often built on top of a fence post or some other exposed site. The North American ovenbird, *Seiurus aurocapillus* (a wood warbler, Parulidae), also builds a domed oven-shaped nest, but of plant materials on the forest floor. Some species, especially members of the Icteridae, make soft hanging nests that range to two feet or more in length. The thorn birds (*Phacellodomus*), as well as many other Furnariidae, build huge nests of twigs suspended from the ends of tree branches; these nests, which may be more than two metres (nearly seven feet) long and contain many compartments, are used by only a single nesting pair, sometimes with nonbreeding helpers (probably the young of the previous season). These nests are often appropriated by troupials (*Icterus icterus*), which evict the owners, even destroying the eggs and young in the process. A few other species also take over nests for their own use, notably the piratic flycatcher (*Legatus leucophaius*, a tyrannid) and the bay-winged cowbird (*Molothrus badius*).

Woven nests

Nests of many passerines are constructed with amazing skill. The tailorbirds of Asia (*Orthotomus*) are noted for nests built in a pocket that the birds make by sewing together the edges of one or more leaves, using plant fibres or other materials. Some species, especially the weavers, are able to tie knots with strips of grass or palm leaves and thus weave an exceptionally tight and compact nest. Others build equally firm nests by felting the materials together. In contrast, a few passerines build flimsy nests (some Cotingidae), apparently as an adaptation toward lessened visibility to predators, for such nests are attended minimally by the parents, seemingly to draw as little attention to the site as possible. Other birds excavate their nests in soft earthen banks, use old woodpecker holes, or find natural crevices in trees or rocks. The type of nest built by the members of a single family may be varied (extremely so in the Furnariidae) or consistent: all woodcreepers nest in holes; all vireos weave a cup between the arms of a forked branch.

Incubation and parental care. Passerines lay clutches of one to 14 eggs, clutch size being unrelated to the size of the bird. The largest species, the two lyrebirds (*Menura*), lay a single egg; some of the smaller titmice (*Parus*) have been recorded with the biggest clutches. In most passerines the female incubates the eggs alone,

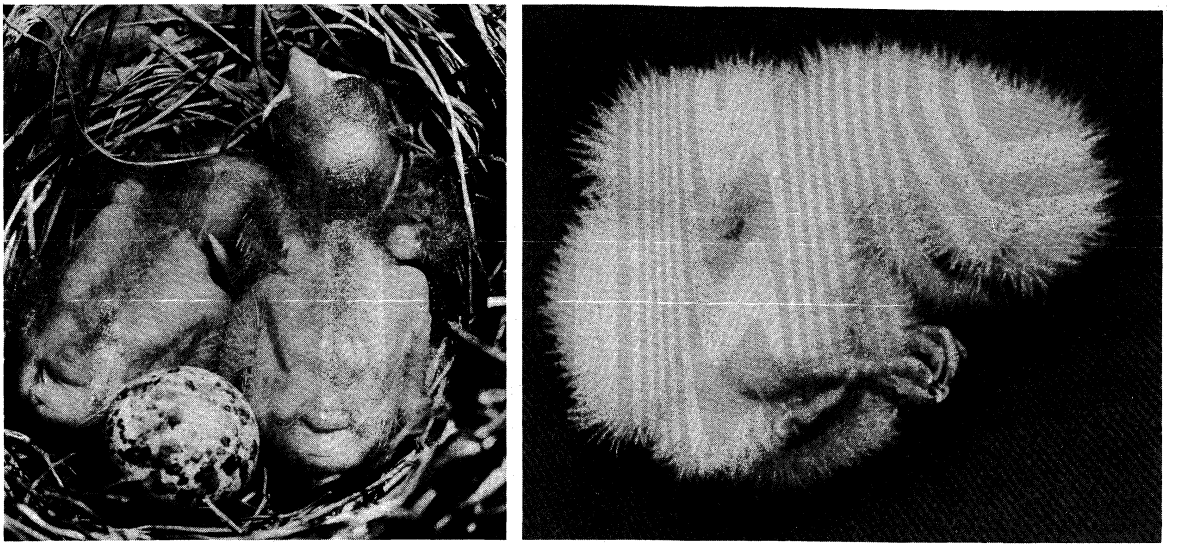


Figure 2: *Two main types of young.*
(Left) Naked young of the North American cardinal (*Cardinalis cardinalis*). (Right) Downy young of the South American bare-throated bellbird (*Procnias nudicollis*).
(Left) Root Resources—EB Inc., (right) David W. Snow

but in some groups—such as the antbirds (Formicariidae), certain grosbeaks (*Pheucticus*), and others—the male shares equally in incubation. Males of most species help to feed the young. Some passerines have only one nest per breeding season, but others may have two or more, especially if one nest is destroyed before the young fledge. The incubation period generally varies from 11 to 21 days depending on the species but is well over a month in lyrebirds. The hatchlings are typically blind, sparsely covered with down, and helpless; some species hatch completely naked, and a very few are densely covered with down at hatching (some cotingas, antbirds of the genus *Formicarius*, and some Campephagidae). The young remain in the nest for eight to 30 or 35 days (about 42 in the lyrebirds) but most commonly from ten to 15 days. After they fledge, they require some days or weeks to become fully independent of their parents.

Variety of vocalizations among passerines

Sound production. An outstanding aspect of passerine behaviour is the ability to sing. Song is best developed in the oscines, which have a highly complex vocal organ or syrinx, but even the more primitive suboscines are capable of a variety of vocal sounds. The woodcreepers (Dendrocolaptidae), ovenbirds (Furnariidae), and antbirds (Formicariidae) sing relatively simple songs, consisting of a few notes or whistles, often repeated rapidly in a trill, roll, or rattle. Manakins (Pipridae) also utter simple trills or whistles: in addition, some species are capable of a loud nonvocal snapping sound, which is produced by specialized wing feathers. The cotingas sing a wider variety of songs, from quiet musical notes to the incredibly loud and far-carrying "gongs" of the bellbirds (*Procnias*). The New World flycatchers are well-known for their range of distinctive call notes, and many species sing well and melodiously. In some groups (notably the *Empidonax* complex), the plumages of closely related species are so similar that the birds can be distinguished in the field only by their calls and songs. Both the lyrebirds and scrub-birds (Atrichornithidae) have syringes more like those of oscines and are known for their loud and complicated songs. They are also accomplished mimics; lyrebirds mimic the songs of almost all birds within their hearing, as well as many mechanical sounds. Many species of oscines have complicated and beautiful songs, notable examples being the nightingales (*Luscinia*) and some other thrushes, larks (Alaudidae), mimic thrushes (Mimidae), and wrens (Troglodytidae). The possession of the complex oscine syrinx does not guarantee a complex song, however, and many "songbirds," such as waxwings (*Bombycilla*) and swallows, utter simpler sounds than do many suboscines.

Only the male of most passerine species sings a true song, although the female can produce a variety of call notes and other sounds. In some species in which the female sings, she seldom does so during the breeding season unless it is a duet with her mate. Such duetting or antiphonal singing of paired birds is so well developed in certain species that it is difficult to determine that the song is coming from two individuals. In the African black-and-red shrike (*Laniarius barbarus erythrogaster*) the reaction time between the male's start of song and the female's response has been timed at 0.135 second.

Interactions with ants. *Anting.* A characteristic but poorly understood behaviour pattern of passerines is the practice of anting. This peculiar ritual has two forms: active anting, in which a bird picks up worker ants in its bill and wipes them on its feathers in a stereotyped manner, and passive anting, in which the bird squats or lies down in a group of ants and assumes an exposing stance so that the ants will crawl up into its feathers. Birds may also apply ants to their plumage while passively anting, but species that use the active stance (the majority of recorded passerines) apparently never use the passive stance. Birds show definite discrimination in the type of ants used, avoiding stinging species and selecting those that exude or spray formic acid or other defense fluids (ants of the subfamilies Formicinae and Dolichoderinae of the family Formicidae). A great deal of controversy has existed over the function of anting. Some authorities have theorized that it is a form of self-stimulation, but British ornithologist K.E.L. Simmons has argued convincingly that it is a type of feather maintenance. Formic acid and other ant fluids are known to be insecticidal; dressing the feathers with ants would thus kill or deter avian parasites, such as lice and mites. Additional components of ant fluids include essential oils, which could be used by birds to supplement the oils from their own uropygial (preen) gland. After a bout of anting, birds often continue feather-maintenance activities by bathing, oiling (from the uropygial gland), and preening. Recent studies have shown anting to be most prevalent during molt, when the bird's skin is irritated by the growth of new feathers. Anting clearly is innate behaviour, and its remarkable uniformity in at least 30 passerine families, both oscine and suboscine, implies that it has real importance to the bird. Some individuals have been seen to ant with such things as cigarette butts, orange peels, mothballs, and smoke, apparently reacting to the pungent fumes of these objects in the same way as they do to the strong odours of ants. A few nonpasserines have also been observed going through motions that are similar to anting, but as yet true anting is known only in the Passeriformes.

Active versus passive anting

Ant-following. Another specialized form of behaviour, also associated with ants, is the practice known as ant-following. In the New World tropics, nomadic army

ants move in huge troops, swarming over the forest floor in columns as wide as ten metres (more than 30 feet) or more. Because the ants devour all the small animal life in their path, a moving column of them is edged by fleeing insects, spiders, millipedes, isopods, small frogs, and lizards. The ant columns are accompanied by troops of birds that seize the fugitives. Ant-following birds apparently do not eat the ants but only the insects and other small animals trying to escape. A number of passerine species, notably several antbirds, are believed to be entirely dependent on army ants for finding food. Many other birds also follow ants when they come upon them; these include woodcreepers, manakins, New World flycatchers, tanagers, wrens, and occasional ovenbirds. Even some nonpasserines may join a troop of ant-followers—motmots (Momotidae), tinamous (Tinamidae), and hawks—although the hawks may be more attracted by the ant-following birds than by the insects. The same ant-dependent species have also been known to follow large animals, including man, that stir up insects with their feet.

Popperfoto, London

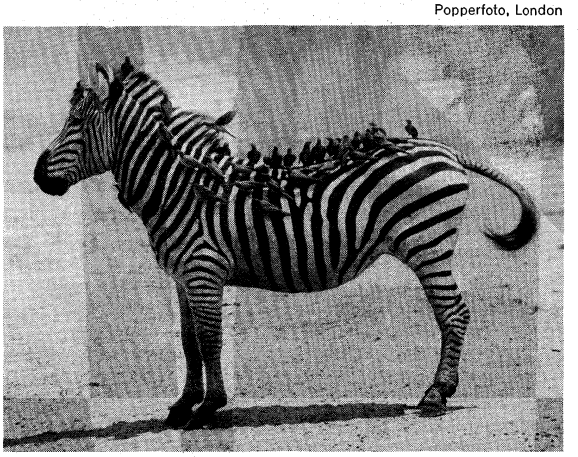


Figure 3: A zebra covered by a large number of oxpeckers.

Association of passerines with large animals

A few passerines, although not ant followers, will escort large quadrupeds, such as cattle, buffalo, and deer, to catch the insects that fly up around them and to feed on the ticks and flies parasitizing the animals themselves; especially noted for this behaviour are the cattle tyrant (*Machetornis rixosa*, Tyrannidae) tickbirds or oxpeckers (*Buphagus*, Sturnidae), and several cowbirds. In Australia yellow robins (*Eopsaltria*, Muscicapidae) will follow the much larger lyrebirds as they scratch and feed along the ground.

FORM AND FUNCTION

External features. *Feet and legs.* The single feature that distinguishes passerines from all similar birds is their "perching" foot. In this foot type, all four toes are well developed and free from one another; in some families (wrens and most suboscines) the front toes may be partially fused at the base, but the distal portions (extremities) are functionally free. The hindtoe (hallux) is joined on the same level with the front toes and opposes them, so that the foot can grip a perch. The only exception to this passerine foot type is found in the well-named *Paradoxornis paradoxus*, or three-toed parrotbill (Panuridae), in which the outer toe is reduced to a short clawless stump, fused to the middle toe; other species of *Paradoxornis* have normal feet.

Modifications of the foot

Although all passerines can perch, not all do so habitually. A number of species (some tapaculos, Rhinocryptidae; larks; pipits, Motacillidae) are largely terrestrial and have feet modified for walking and running; the terrestrial foot is differently proportioned from the typical perching one, often with longer toes and longer, straighter claws (particularly on the hallux), probably as an aid in maintaining balance when running. The dippers, or water ouzels (*Cinclus*), are semi-aquatic, but although they successfully swim on the water surface and walk under water searching for food on stream bottoms, they have retained the typical passerine foot. The single slight difference in the *Cinclus* foot is that the claw of the middle toe sometimes has a thin horny flap (of unknown function) on its inner border. Some other passerines, notably swallows, live a largely aerial life and have small and weak feet. The typical arboreal songbird has a well-developed foot, with the middle front toe longer than the others. Birds such as woodcreepers and nuthatches that often cling to vertical surfaces have strong, curved, sharp claws. Those that spend much of their time walking and scratching on the ground (although not limited to terrestrial activity) tend to have heavy, straighter, and rather blunt claws. Most passerines, however, have moderately curved sharp claws that are suited to grip a variety of rounded or rough surfaces.

The lower leg of passerines, the tarsometatarsus (usually called simply the tarsus), is normally covered by a horny sheath (podotheca). Exceptions include some swallows, which have feathered tarsi. Although the various different patterns of scale size and distribution of the normal unfeathered podotheca have been used by some taxonomists to differentiate families or groups of families, study has revealed so much variability in the tarsal patterns of certain families that it is no longer considered a reliable family character; it may still be useful as a generic or specific character. In most oscines the posterior (plantar) surface of the tarsus is bilaminate—*i.e.*, covered by two long plates, or laminae.

Bill. The bills of passerines are extraordinarily diverse in size, shape, and proportions. This diversity was long thought to be indicative of the birds' relationships and so was used as a prime taxonomic character. It is now believed, however, that bills are evolutionarily plastic, reacting with relative ease to selective pressures, particularly to changes in feeding habits. Thus on a broad scale a passerine's bill shape reveals less about its family affinities than it does about its food preferences, and, although bill shape may be an aid to determining a bird's relationships, it must be considered in the light of other features and of the degree of variation found in the family. Two often-cited examples of the adaptiveness of bills are the geospizine, or Darwin's finches (Fringillidae) of the Galápagos Islands, and the Hawaiian honeycreepers, Drepanididae (see EVOLUTION). Each is a closely interrelated group of birds with different kinds of bills in the several species and genera. Those of drepanidids range from heavy, seed-cracking, grosbeak-like bills through thin, pointed insectivorous types to the long, decurved (curved downward) bills of nectar feeders. These Hawaiian birds are now thought to be closely related members of a single family of nine genera. On the basis largely of bill shape, they were once classified into four different families and 18 genera.

Most birds, including passerines, show little sexual dimorphism (difference between sexes) in bills except for minor differences in size (reflecting general body size differences) and sometimes in colour. The most outstanding exception is the extinct huia (*Heteralocha acutirostris*, Callaeidae), originally classified as two different species. The male of this New Zealand bird had a strong, chiselling bill, whereas the female had a long, decurved, pliable bill. Reportedly the two sexes fed cooperatively, the male digging in decaying wood and the female probing in crevices to extract grubs. The species unfortunately was prized by the Maoris, who used the white-tipped tail feathers in ceremonial headdresses, as well as by Europeans, and, after most of its habitat had been destroyed, the huia was hunted to extinction about the end of the 19th century.

Specializations of the bill

Passerine bills may be broadly classified into eight morphological and functional types:

1. Insectivorous: a generalized type found in many passerines, ranging from relatively straight and pointed (as in the wood warblers, Parulidae), through bills with a slight or pronounced hook (some New World flycatchers), to those that are short, with a wide gape and usually surrounded by rictal bristles (stiff hairlike feathers)—*e.g.*, in aerial feeders, such as swallows. Most insectivorous

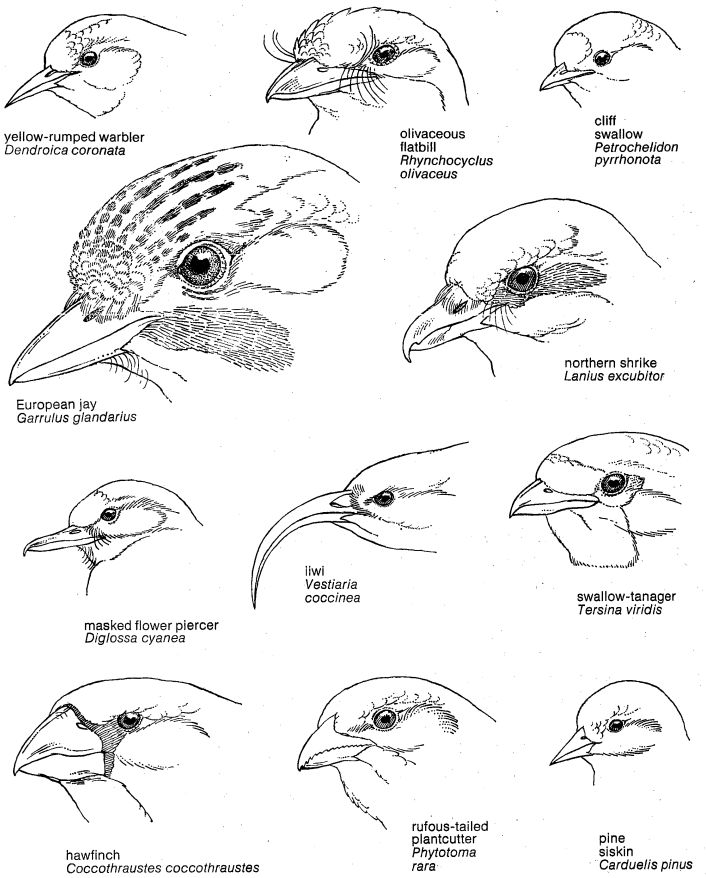


Figure 4: Types of bills found among passerine birds.
Drawing by R. Keane

bills are relatively light in build, but this depends on the type of insect usually taken by the species and also on how generalized a feeder it is.

2. Omnivorous: unspecialized in shape and function but usually strongly built, as in crows and jays (Corvidae).

3. Toothed: strongly hooked at the tip and with a "tooth" (notch) on either tomium (cutting edge) of the upper mandible; adapted to tearing up large, relatively soft prey. This is the typical bill of shrikes (Laniidae) but is also found in some unrelated birds, such as the Australian bellmagpies (Cracticidae) and some tanagers.

4. Tearing: a relatively light bill with a strong hook at the tip, for tearing open objects, such as flowers, to obtain the insects and nectar within. Found in flower piercers (*Diglossa*, Thraupidae).

5. Probing: relatively narrow and often downcurved; slender in species that probe flowers for tiny insects and nectar (sunbirds; some Hawaiian honeycreepers) but more heavily constructed in those that probe in wood or under tree bark (creepers, *Certhia*; some woodcreepers).

6. Frugivorous: variable but usually rather wide; ranges from lightly built with a wide gape for swallowing whole fruits (found in some cotingas, and in the swallow-tanager, *Tersina*) to more heavily built for tearing apart tougher fruits (some tanagers).

7. Serrated: conical, with a finely serrated edge, adapted for feeding on leaves, buds, shoots, and fruit. Found only in the plant cutters (Phytotomidae).

8. Conical: Adapted for seed eating. Ranges from exceedingly stout and blunt (*e.g.*, the hawfinch, *Coccothraustes*, which can crack remarkably hard objects, such as cherry pits) to relatively small and pointed (siskins, *Carduelis*). Some forms specialized for particular kinds of seed extraction (*e.g.*, crossbills, *Loxia*, which feed on pine seeds).

This classification indicates morphological and functional types of bills, but it does not imply that a species with a particular type of bill will feed only on the food for which it is best adapted. Although some birds are extremely specialized in their feeding habits, most are opportunistic feeders, seizing upon whatever food is readily available and can be "handled" with the bill. Hence many basically granivorous or frugivorous birds catch insects, especially when feeding nestlings, and many insectivorous species exploit seasonally available plant food. Myrtle warblers (*Dendroica coronata*) and tree swallows (*Iridoprocne bicolor*), for example, feed on bayberries in fall and winter, and eastern kingbirds (*Tyrannus tyrannus*) and other New World flycatchers will eat a variety of fruits and berries in season.

The mandibles of passerines, like those of all other birds, are composed of bone covered with a horny sheath, the ramphotheca. The ramphotheca is worn down by normal use and, in most birds, is capable of growing to replace the lost material. In individuals with damaged bills or those (such as cage birds) that do not have the opportunity to wear down the constantly growing ramphotheca, the bills overgrow at the tip.

Plumage and pterylosis. The colours, patterns, and textures of passerine feathers are considered important taxonomic characters, especially in determining genera, species, and subspecies. Plumage is also occasionally used in a very broad way to indicate evolutionary levels. Spots, streaks, and dull colours are generally considered more primitive than bold or complicated patterns and bright colours, but there are many exceptions to this rule.

Passerines often are sexually dimorphic in their plumage, with adult males wearing brighter colours and more striking patterns than do females. In some families, notably tanagers (Thraupidae), wood warblers (Parulidae), and New World orioles (Icteridae), the temperate zone species show more sexual dimorphism than do tropical members of the same families. In addition, many species (especially those in temperate climates) are seasonally dimorphic, with a bright plumage during the breeding season and a dull one in winter. Juvenile plumages of both sexes tend to be cryptically coloured (*i.e.*, adapted for concealment), as is that of the adult female.

Colour varieties in passerines

Virtually any colour may be found in one passerine or another, and the order offers a wide array of specialized feather types, such as the waxlike tips on the flight feathers of waxwings (Bombycillidae); the tufts of stiff feathers in some honeyeaters (Meliphagidae); iridescent "spangles" in some manakins, sunbirds, and tanagers; and the almost unbelievable array of "wires," iridescent gorgets, velvety ruffs, racquet tails, and filamentous plumes of the birds of paradise.

Another taxonomically important character is the number and distribution of feathers (pterylosis) on the bodies of passerines. From external appearance all birds seem to be more or less evenly covered by feathers; in actual fact, however, most birds have their feathers growing from relatively narrow tracts (pterylae) in the skin. From the pterylae the feathers fan out and cover the remainder of the bird's body. In passerines, the feathers are arranged in eight distinguishable tracts, with apteria (relatively bare skin) between them. Variations in tract width and length and especially differences in feather number and distribution are often useful in determining relationships. Of particular interest are the occurrence of apteria within tracts and the configuration of the ventral tract. Also used in classification are the numbers of flight feathers. The remiges (flight feathers on the wings) of most passerines consist of ten primaries on the "hand" (manus) and nine secondaries on the forearm (ulna). In all perching birds the tenth (outermost) primary is reduced to some degree, and in many families only nine may be found. The number of secondaries is more variable, with some species having as many as 14 (the satin bowerbird, *Ptilonorhynchus violaceus*). Tail feathers (rectrices) are also variable; most passerines have 12, but the number ranges from six to 16.

Of importance in some species is the relative length of the primaries. This "wing formula" is often useful in distinguishing between species of such difficult groups as the New World flycatchers and the Old World warblers (Sylviidae).

Internal features. *Syrinx.* In a group of birds as vocal as the passerines, it is natural that the structure of the vocal apparatus should have evolutionary significance. Differing from the mammalian larynx in both location and structure, the syrinx consists of a resonating chamber at the lower end of the windpipe (trachea), with associated membranes, cartilages, and muscles. These modifications involve elements of the bronchi (the two tubes connecting the trachea with the lungs) as well as those of the trachea. Since the mid-19th century the basic subdivisions of the order Passeriformes have been based primarily on the structure of the syrinx. Some of the features of the syrinx long thought to be indicators of relationships have recently been shown to have little taxonomic validity. Other taxonomic characters now must be studied and evaluated before the classification can be adjusted.

Syringeal muscles are classified into two groups: extrinsic muscles, which connect the syrinx with other parts of the anatomy, and intrinsic muscles, which extend from one part of the syrinx to another. The number, shape, and attachments of the intrinsic muscles are likely to remain important in passerine classification. Those birds in which the muscles are inserted on the middle of the bronchial semi-rings (C-shaped cartilages that strengthen the bronchi) are sometimes called mesomyodian (most members of the suborder Tyranni) and those with the insertion on the ends of the semi-rings are acromyodian (Menurae, Passeres). The Eurylaimi and a few others have no intrinsic muscles. Further distinction is made in the number of pairs of intrinsic muscles, most importantly in the Passeres, which have four.

Vocal-organ specializations

The passerine syrinx exists in four basic types:

1. Unspecialized: relatively little modification of the tracheobronchial region; few, if any, cartilaginous specializations, and no intrinsic muscles; found in broadbills (Eurylaimidae), pittas (Pittidae), New Zealand wrens, asities (Philepittidae), plantcutters, most cotingas, and a few manakins and tyrant flycatchers.
2. Tracheophone: most of the specializations limited to the tracheal region; intrinsic muscles number none to two pairs; pessulus (a bony bar lying at the junction of the bronchi) absent; found in all members of the Furnarioidea (South American ovenbirds, woodcreepers, antbirds, and tapaculos).
3. Intermediate tracheobronchial: various modifications of cartilages and membranes; one or two pairs of intrinsic muscles; pessulus present or absent; found in the sharpbill (*Oxyruncus*) and most manakins and tyrant flycatchers.
4. Oscine (acromyodean): complex musculature involving four pairs of intrinsic muscles (but three pairs in lyrebirds and scrub-birds); some cartilaginous specializations; pessulus present (except in larks).

Despite the rather extensive knowledge of the morphology of the syrinx and its importance in passerine classification, very little is known about how it functions to produce sounds. Several theories have been proposed, but none has yet been proved wholly satisfactory.

Skeleton. Of the many variations in passerine skeletal structure, only a few that are important in classification are mentioned here.

In the skull the bony palate, composed of a number of small bones, is termed aegithognathous; also found in swifts (Apodiformes), this palatal type is characterized by the shape and type of fusion of the small bones of the palate. Within this basic type the many minor variations in shape, size, and position of the component bones are useful in delimiting closely related groups of birds, especially suboscines.

Elsewhere on the head, variations in the hyoid apparatus, a complex of small bones that supports the tongue, have been used in passerine classification.

In the sternum (breastbone) the shape of the anteriormost spine (spina sternalis) and the number of notches in the posterior border are of great interest. The spina sternalis, short and forked in most passerines, is long and simple in the Eurylaimidae (one exception), Philepittidae, and a few Cotingidae. All oscines and most suboscines have a single pair of posterior sternal notches; only the tapaculos and certain of the terrestrial antbirds (*Conopophaga, Pittasoma, Hylopezus, Myrmothera*) have two pairs. The sternum of lyrebirds differs from all others in the order in being exceptionally thick, long, and narrow; it may have no posterior notches at all, or it may have a single shallow pair.

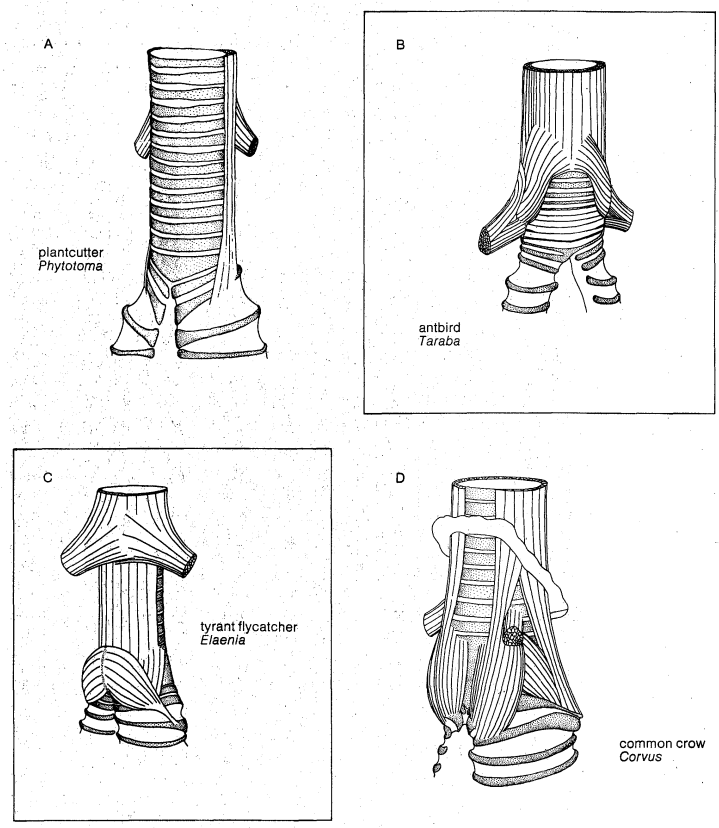


Figure 5: *Four major syrinx types of passerine birds.* Figures show lower trachea, upper bronchi, and attached muscles, seen from the ventral side. (A) unspecialized; (B) tracheophone; (C) intermediate tracheobronchial; (D) oscine.
From P.L. Ames in (A,B,C) *Bulletin of the Peabody Museum of Natural History*, no. 37 (1971), Yale University; (D) J.C. George and A.J. Berger, *Avian Myology* (1966), Academic Press

Musculature. A number of different muscle systems have been important in passerine classification. Important examples, in addition to those of the syrinx, are the muscle complexes controlling the tongue, the jaws, the wings and pectoral girdle, and the legs and pelvic girdle. One character that has been used for over a century is the condition of the deep plantar tendons. These narrow straps extend from the bellies of the two deep flexor muscles on the leg, down the back of the tarsometatarsus, and attach to the toes. They act to close the toes (hence to grasp a perch). In the Eurylaimidae these tendons are connected by a short band (vinculum), but in all other passerines they are entirely separate. This difference has been used by some to divide the passerines into two major groups: the Desmodactyli (vinculum present) and the Eleutherodactyli (vinculum absent).

EVOLUTION AND PALEONTOLOGY

Origins of perching birds

Passerines, like other small land birds with fragile bones and arboreal habits, are rarely fossilized; hence the fossil record gives few clues to their evolution. Their origins, both in time and ancestral type, can only be guessed. Some authorities believe that passerines arose during the Early Cretaceous Era (about 120,000,000 years ago); others believe that it was much later, not until the Late Cretaceous or even Paleocene Epoch (perhaps 65,000,000 years ago).

The earliest known passerine fossils are from the upper Eocene Epoch (about 40,000,000 years ago). One of these is an extinct genus assigned to the modern family of tapaculos (Rhinocryptidae), another is thought to be related to the titmice (Paridae), and still another to the starlings (Sturnidae). *Palaeospiza*, a primitive oscine

classified as a separate family (Palaeospizidae) between the larks and the swallows, is known only from the Colorado Oligocene Epoch (about 35,000,000 years ago). As drying conditions during the Miocene Epoch (26,000,000 to 7,000,000 years ago) reduced the forests and encouraged the spread of grasslands, bird families exploited the newly appearing drier and less forested habitats and radiated into them. There for the first time appear remains of crows (Corvidae), thrushes (Turdidae), wagtails (Motacillidae), shrikes (Laniidae), and wood warblers (Parulidae).

During the Pliocene Epoch (from 7,000,000 to 2,500,000 years ago) the warm, dry conditions continued, and most paleornithologists now believe that all of the living passerine families were in existence by its close. They also believe that most modern species of birds arose during the Pleistocene Epoch (about 2,500,000 to 10,000 years ago), a period of almost constant change during which there were four major advances and retreats of the glaciers. Most of the passerines in the fossil record are from the Pleistocene or Recent epochs and represent either living species or close relatives. Evolution since the retreat of the last ice sheet (about 11,000 years ago) has been mainly at the subspecies level.

With the lack of a significant fossil record, nothing is known of the type or types of birds from which passerines arose. Studies of the anatomy of modern forms have led to a general agreement that the perching birds developed from more than one ancestral type (polyphyletic origin). But until more evidence is unearthed, nothing further can be said.

CLASSIFICATION

Distinguishing taxonomic features. The classification of passerines is probably the single most vexing problem in avian taxonomy today. Few of the subgroupings (especially oscine families) have strikingly different characteristics, the fossil record is inadequate, and the order has undergone an explosively radiative evolution that has resulted in a large number of anatomically similar species. Classification is further complicated by the persistence of intermediate and relict forms and a high degree of convergence (similarity due to different lines being subject to the same types of natural selection) among distantly related groups. It is therefore necessary to examine each passerine species in careful detail (including its anatomy) to be certain of its affinities. With some 5,100 species in the order, this is a monumental task, one that is by no means complete. Traditionally, passerine families have been defined on the basis of a careful (usually anatomical) study of only a few individuals; the great majority of the birds have been assigned to established families solely because they externally resemble the studied species in plumage, bill, leg, and foot characters, as well as in what is known of their life history, behaviour, geographical distribution, etc. It is unfortunate that in the early years of avian taxonomy there was a strong tendency to try to include each newly discovered species, regardless of its geographical range, in some previously defined palearctic family.

Re-evaluation of early taxonomic characters

In the mid-20th century, taxonomists began re-examining the generally accepted 19th-century groupings of birds into families. In these investigations some of the features on which earlier classifications were based have been found to be convergent or too variable to be useful in certain groups (*e.g.*, bill shape, tarsal scutellation). Consequently, passerine taxonomists have been left with a rather sparse body of morphological information upon which to base a classification. In recent years ornithologists have made a concerted effort both to augment some of the century-old work on passerine anatomy and to explore new avenues of morphology, behaviour, reproductive patterns, biochemistry, and zoogeography to help define and relate the many families of perching birds. Much of this work is still in progress and has not yet been incorporated into classification systems. Among traditional and newly studied taxonomic features are external characters—such as rictal bristles and other specialized feathers, colours and patterns of the fleshy parts of the mouth, morphology of the bill and nostrils, colour patterns of adults and young; internal anatomical characters—such as the number of cervical (neck) vertebrae, the condition of the deep plantar tendons, anterior and posterior spines and processes of the sternum, syringeal muscles, palatal and other bones of the skull, feather tracts, jaw and tongue musculature, hyoid (tongue) apparatus, aortic-arch system, pneumatic fossa (cavity) of the humerus, and types of spermatozoa; biochemical analysis of substances—such as egg white, eye lens, plasma proteins, and hemoglobins; and an array of behavioral traits—such as reproductive behaviour, nest building, methods of scratching, etc. Thus, although the current system is by no means satisfactory, improvement is in the offing, and it is hoped that in the near future modern studies will lead to a more firmly drawn passerine classification.

Annotated classification. The classification and sequence of families given here is essentially that proposed in 1971 by American ornithologist Oliver L. Austin, Jr., based on a 1960 classification by Alexander Wetmore, with certain changes (see *Critical appraisal*). Some of the traditional diagnostic characters have been omitted because they have proved unreliable or because preliminary restudy indicates that the old generalizations do not hold true.

ORDER PASSERIFORMES (perching birds, or passerines)

Land birds with a characteristic "perching" foot; 4 toes (never webbed) joined at the same level, with the 1st toe (hallux) directed backward and never reversible. Oil gland unfeathered. Wing eutaxic (no gap between the 4th and 5th secondaries), usually with 9 or 10 primaries. The young altricial—*i.e.*, hatched almost or completely naked of feathers (a few exceptions), helpless, and requiring a considerable period of parental care. Groups marked with a dagger (†) are extinct and known only from fossils.

Suborder Eurylaimi

Syrinx simple, tracheobronchial, lacking intrinsic muscles and cartilaginous specialization; pessulus present. Sternum with anterior spina sternalis long and simple (exceptions) and with one pair of notches in posterior border. Hallux weak but about as long as the other toes. Clavicles well developed. Usually 15 cervical vertebrae (all other passerines have 14). Deep plantar tendons of different type than all other passerines.

Family Eurylaimidae (broadbills). Generally brightly coloured, chunky birds, with large heads, short necks; 12.5 to 28 cm (5 to 11 in.) in length. Bill broad and flattened, covered with a crest in some; gape wide. Front toes partially joined; 10 or 11 primaries. Other characters as in subordinal definition above; 14 species, primarily in forests or cloud forests. Indo-Malaya, Africa, Philippines.

Suborder Tyranni (majority of suboscines)

Syrinx usually more complex; muscles variable; pessulus present or absent. Sternum with short spina sternalis, forked (exceptions noted below); posterior border with one or two pairs of notches. Hallux strong. Clavicles well developed.

Superfamily Furnarioidea

Syrinx complex, tracheal; pessulus absent. Pterylosis (distribution of feathers) variable. Sternum with one or two pairs of notches.

Family Dendrocolaptidae (woodcreepers). Characteristically slender arboreal birds, olive brown to rufous, usually streaked or barred; 14.5 to 37 cm (5.5 to 15 in.). Strong probing bills, laterally compressed, short and straight to long and downcurved; nares holorhinal (see Furnariidae). Short legs, powerful feet; front toes partially joined at base. Tail stiff with spiny tips. Posterior border of sternum with one pair of notches. Syrinx with cartilaginous specializations and two pairs of intrinsic muscles. Pterylosis with distinctive ventral tract type. Forty-eight species, forest or brushland. Mexico through South America.

Family Furnariidae (ovenbirds). A large and extraordinarily diverse group, an excellent example of adaptive radiation. In South America, especially in the southern Brazil–Argentina–Chile region, this family has evolved to fill a broad range of ecological niches; various furnariids look and behave like wrens, thrushes, dippers, creepers, larks, wood warblers, titmice, nuthatches, and even the nonpasserine woodpeckers (Picidae) and sandpipers (Scolopacidae). If such diversity can be summarized, ovenbirds are generally small, dull brown birds, darker above and paler below; 12 to 28 cm (about 5 to 11 in.). One distinctive character is the shape of the

external nares (nostrils), which are schizorhinal (slitlike) at their posterior border rather than holorhinal (rounded), as in other passerines. Front toes partially joined at base. Sternum with one pair of notches (some exceptions). Syrinx with two pairs of intrinsic muscles. Pterylosis distinctive within suboscines but of same type as many oscines. Approximately 215 species, wide variety of habitats; Mexico through South America.

Family Formicariidae (antbirds). A large and diverse family, with loose-webbed plumage, generally in browns, grays, black, and white. Overall size range 9.5 to 37 cm (less than 4 to nearly 15 in.), including several large terrestrial species with long legs but very short tails (hence a deceptively short "body length"). Bill strong, hooked to variable degrees. Front toes slightly joined at base. Sternum with one or two pairs of notches, the two pairs found in most but not all of the long-legged terrestrial species. Syrinx of two types: the first with one pair of ventrally originating intrinsic muscles (largely the arboreal species), the second with no intrinsic muscles. Pterylosis variable; dorsal tract with reduced or absent posterior element (primarily the arboreal species); several different dorsal types among the terrestrial species. Wings generally short and rounded, flight weak; about 231 species, largely in dense forests and brushland. Neotropical; Mexico to northern Argentina.

Family Rhinocryptidae (tapaculos). A small, little-known group. Plumage dull, generally in grays, browns, and black; 9 to 25.5 cm (less than 4 to more than 10 in.). Bill distinguished by a movable flap (operculum) over the nostrils. Legs, feet, and claws strong. Wings short and rounded, flight feeble. Sternum with two pairs of notches. Syrinx with one pair dorsally originating intrinsic muscles (some exceptions). Pterylosis distinctive in ventral tract (exceptions). Twenty-seven species, grasslands, scrublands, and dense forest undergrowth. Central and South America, especially Chile and Patagonia.

Superfamily Tyrannoidea

Syrinx tracheobronchial; pessulus present or absent. Pterylosis: dorsal saddle with full-length apterium (exceptions). Sternum with one pair of posterior notches.

Family Cotingidae (cotingas). An extremely diverse, probably composite family; ranges from plain-coloured flycatcher-like to bright, extravagantly adorned birds: 9 to 45.5 cm (about 4 to 18 in.). Some with extraordinary plumes and lappets (the crow-sized umbrella bird, *Cephalopterus*); others renowned for their loud voices, fleshy wattles, and white plumage (white being unusual in land birds). The family as traditionally constituted can no longer be characterized anatomically: the syrinx is of two types; the tarsal envelope is of several different kinds; the joining of the front toes is variable; the dominant artery of the thigh is either the sciatic or femoral; the spina sternalis is forked except in some bell birds (*Procnias*). Several species with powder downs, a specialized type of feather otherwise known in passerines only in the oscine wood-swallows (Artamidae); 91 species, forests. United States–Mexico border to Bolivia, Peru, and Argentina.

Family Pipridae (manakins). Most piprids are fairly uniform in external appearance; generally small (8.5 to 16 cm [3.5 to 6.5 in.]), rather stubby, with short wings and tail (a few long-tailed species). Most males black with patches of brilliant colour (red, yellow, blue, etc.); females are generally drab olive green. Several species, such as the broad-billed manakin (*Sapayoa aenigma*), are externally and anatomically unlike the majority of manakins and may not be properly included in this family. Bill short, rather broad at base, notched, with a slight hook at tip. Some species with specialized feathers, particularly in the wings (also true of some cotingas and tyrannids). Third toe partially fused at base to 2nd or 4th. Syrinx highly variable, as family is presently constituted; 59 species, tropical and subtropical forest. Southern Mexico to Paraguay and northern Argentina.

Family Tyrannidae (New World, or tyrant, flycathers). Large family of generally (but not exclusively) arboreal birds with plumage in grays, browns, olive greens, some with black, white, and yellow, occasionally brighter colours. Many with erectile crowns or crests, often more colourful than the rest of the plumage; length 7.5 to 40.5 cm (3 to 16 in.). Bill extremely variable but commonly broad, somewhat flattened, and hooked at tip; nostril rounded without an operculum or narrow with membranous operculum; most species with well-developed rictal bristles (stiff hairlike feathers around the mouth). Feet weak except in the few terrestrial species; front toes variably but never strongly fused. Syrinx characterized by 1 pair of intrinsic muscles (none or 2 in some), generally variable in other features. About 367 species, with a wide range of habitats. Northern Canada and Alaska through North and South America to Tierra del Fuego, also Falkland and Galápagos Islands. (The family common name of New World flycatchers serves to distinguish it from the large oscine family of Old World flycatchers, Muscicapidae.)

Family Oxyruncidae (sharpbill). A single, rare, little-known species, sometimes placed in the Tyrannidae. Length about 17 cm (7 in.); externally very much like flycatchers, with the same basic syringeal structure. Differences include a long, straight, and sharply pointed bill, surrounded at the base with short, fine, stiff feathers (not rictal bristles). Nostril narrow, covered by a broad horny operculum. Feet strong. Locally distributed in humid forest from Costa Rica to southeastern Brazil and Paraguay.

Family Phytotomidae (plantcutters). A small family of finchlike birds, gray or brown streaked with black, with areas of rusty red; about 17 cm (7 in.) long. Bill distinctive, stout and conical with a finely serrated edge. Legs short, feet large and strong. Syrinx simple, similar to that of some Cotingidae. Pterylosis: flank margin of ventral tract oblique, contrasting with truncated margin in other New World Tyrannoidea; 3 species, open brush and cultivated regions (where it is often considered a pest). Peru to Argentina and Chile.

Family Pittidae (pittas). A relatively homogeneous family; stout-bodied, long-legged, largely terrestrial, 15 to 28 cm (6 to 11 in.); with a wide range of colours in the loose-webbed plumage: reds, greens, blues, as well as browns, black, and white. Wings short and rounded but strong; tail very short. Syrinx simple, lacking intrinsic muscles and cartilaginous modifications; pessulus absent in most species. Egg white protein pattern differs from that of all other suboscines. Twenty-three species, Old World tropics in forests and scrublands. Africa through Malaysia, Australia, Solomon Islands.

Family Xenicidae (*Acanthisittidae* of many authors; New Zealand wrens). Small birds, 7.5 to 10 cm (3 to 4 in.), look and act much like true wrens (Troglodytidae). Considered evolutionary relicts, 2 species are strongly arboreal, the other 2 terrestrial. One of the latter, the now-extinct Stephen Island rock wren (*Xenicus lyalli*), was markedly terrestrial and may have been flightless; if so, it would have been the only known passerine to have completely lost the ability to fly. The 3 extant species are very weak flyers, with short wings and very short tails. Legs long and slender; 3rd and 4th toes fused basally, and all toes (especially the hallux) with long claws. Forest and scrub; New Zealand.

Family Philepittidae (asities and false sunbirds). The two species of asities (*Philepitta*) are black or yellowish green, rather pitta-like, with stout bodies and long legs. The two false sunbirds (*Neodrepanis*) are very different externally, looking much like true sunbirds (Nectariniidae), with which they were long classified; small, blue, yellow, and greenish with short tails and long, slender, curved bills. All philepittids have bare skin or wattles around the eyes or both. Currently classified in the Tyranni, but the simple structure of the syrinx and the unforked spina sternalis have led some recent authorities to suggest they are more closely allied to the Eurylaimi. Forests of Madagascar.

Suborder Menurae

Syrinx tracheobronchial, acromyodian, with 2 or 3 pairs of muscles inserted on the ends of the bronchial semi-rings; pessulus present. Sternum with short, forked spina sternalis, 1 pair of notches or none. Hallux strong. Clavicles not well-developed.

Family Menuridae (lyrebirds). Among the largest members of the order in length, 75 to more than 100 cm (30 to 40 in.), with some weighing over 1,100 grams (about 2½ pounds), lyrebirds are aberrant passerines in virtually all features. In size, brownish plumage, and terrestrial habits, they look much like pheasants. The tail of 16 feathers is extraordinary: in males of *Menura superba* the outermost pair of feathers is curved like a lyre, the next 6 pairs are light and filamentous, and the innermost pair is modified into narrow "wires"; the tail of *M. alberti* is shorter and simpler; females of both species have rather long but not extremely modified tails. The alula (free digit on the leading edge of the wing) on the short rounded wings has 6 feathers rather than the usual 3 or 4. Rictal bristles present. Legs and feet strong and heavy. Sternum long, narrow, and thick, unique within passeriines; posterior border may be entire or with 1 pair of shallow notches. Syrinx of the oscine type, but with only 3 pairs of intrinsic muscles; it is capable of a remarkable variety of loud ringing calls and the birds are famous for their vocal mimicry. Two species, dense forests; Tasmania (introduced) and eastern Australia.

Family Atrichornithidae (scrub-birds). Wholly unlike lyrebirds in external appearance, rather like large wrens, long-tailed, brownish, and terrestrial; 16.5 to 23 cm (6.5 to 9 in.).

One species, *Atrichornis clamosus*, long thought to be extinct but recently rediscovered. Both species virtually flightless, with very small wings and clavicles small and separated (not fused into a furcula, "wishbone"), a condition unique within passerines. Sternum unlike *Menura*, with 1 pair of deep posterior notches. Legs and feet very strong. Rictal bristles lacking. Similar to lyrebirds in the quality of their strong voices, nesting habits, and basic syringeal structure. Scrub-birds, however, have only 2 pairs of intrinsic muscles in the syrinx. Two species, scrublands; in western and eastern Australia.

Suborder Passeres (songbirds, or oscines)

Syrinx with 4 pairs of intrinsic muscles. Sternum with spina sternalis short and forked, and posterior border with 1 pair of notches. Hallux variable in strength. Clavicles well-developed. All with same complex of syringeal muscles, with only minor variations.

Family Alaudidae (larks). A distinctive and well-defined group, the only oscines with back of tarsus rounded and scaled instead of sharp and unsegmented. Syrinx also unique among Passeres in lacking the usual bony pessulus. Small ground birds, 12 to 23 cm (about 5 to 9 in.), usually cryptically coloured in browns and grayish buffs, plain or streaked, lighter below (several species black below); sexes similar. Bill usually pointed, slightly downcurved; wings long, pointed; legs rather long, hindtoe usually with long, straight claw. About 75 species inhabit open fields, plains, beaches of Old World, with only 1 species, the horned lark (*Eremophila alpestris*), also in North America south to Colombia. Many species migratory. Food: seeds, insects, and other invertebrates. Sing beautifully, often a soaring flight song.

†*Family Palaeospizidae* (*Palaeospiza*). A single fossil species known only from Oligocene of Colorado.

Family Hirundinidae (swallows, martins). A distinctive family placed low on the oscine family tree because of primitive syringeal characters, particularly the double bronchial rings, unique in Passeres. Small birds, 9.5 to 23 cm (about 4 to 9 in.), with compact plumage, often with metallic sheen, usually lighter below; sexes usually alike or nearly so. Wings long, pointed, primaries reduced to 9. Tail medium to long, truncate to forked. Legs short, feet small, weak. Bill short, flat, but wide gape enhanced by rictal bristles. Walk with difficulty but fly strongly, feed on insects caught in flight. Voice usually a twittering or squealing, sometimes melodious. The river martins of Africa and Thailand differ by large, brightly coloured bills, stronger feet, and syringeal characters. The 79 living species are found worldwide except polar regions and certain oceanic islands; all or nearly all migratory.

Family Dicruridae (drongos). Small to medium-sized birds, 18 to 63.5 cm (7 to 25 in.), the longest being those with exceptionally long tails; usually black with purple or greenish sheen, some crested or with spangled neck and head feathers, iris of eye usually red; sexes alike. Bill stout, arched, slightly hooked and notched; long, strong rictal bristles; legs short, feet stout. Wings long; tail variable, of 10 or 12 feathers, usually forked; some species with racquet tails. Arboreal birds that fly well but seldom long or far. Food mostly insects caught on the wing. Voice varied, melodious; capable mimics. The 20 species range through Africa, southern Asia, Malaysia, northern Australia, east to Solomons, in woodlands, savannas, cultivated regions.

Family Oriolidae (forest orioles, fig-birds). Medium-sized birds, 18 to 30.5 cm (7 to 12 in.); brightly coloured, predominantly in yellows, greens, and black; sexes unlike, female duller, young streaked below. Bill strong, pointed, slightly hooked; long, pointed wings with 10 primaries; medium to long tail of 12 feathers. Strictly arboreal birds that feed on insects, fruit; 28 species in forest and open woodland across Eurasia, in Africa, East Indies, Australia, Philippines. Flight strong, undulating. Voices loud, with harsh calls, but melodious songs.

Family Corvidae (crows, jays, ravens, rooks, choughs, nutcrackers, and magpies). Medium to large passerines, including the heaviest oscines, the ravens (*Corvus*); 17.5 to 70 cm (7 to 28 in.). Ten primaries, the 10th (outer) always much shorter than 9th, but longer than primary coverts. Bill strong, powerful, longer than rest of head or nearly so; nostrils usually covered with bristles. Crows (*Corvus*) and allies, large, black or black and gray or white; wings long, tail shorter than wing. Jays generally smaller, often coloured in blues, greens, yellows; wings rounded; tail sometimes two-thirds of total length. About 102 species, almost cosmopolitan (absent from southernmost South America, Antarctica, some oceanic islands, introduced to New Zealand); varied habitats, prefer woodlands, open brushlands. Voices harsh, loud.

Family Callaeidae (wattlebirds). Medium-sized, 25.5 to 53.5 cm (10 to 21 in.); black, brown, or blue-gray, with fleshy blue or orange wattles at the gape; sexes may differ in size, wattles, and bill shape. Weak fliers but hop strongly. Sternum weak; 10-primaried wings short, rounded; tail long. Bill stout and short to long and curved; legs long, feet stout. Eat fruit, nectar, insects. A variety of musical notes and whistles. Three species limited to primeval forests of New Zealand; huia (*Heteralocha*) extinct; wattlebird (*Callaeas*) and saddleback (*Creadion*) rare.

Family Grallinidae (mudnest builders). Ten-primaried oscines of medium size, 19 to 50 cm (7.5 to 20 in.). Differ from most corvids in lacking nasal bristles. Legs long, strong; wings long and pointed to short and rounded. Black and white or dark gray; sexes alike or unlike. Four species limited to Australia, western New Guinea, in woodlands, marshes, cultivated lands, usually near water. Weak fliers with peculiar jumping gait; all build deep, open-bowl nests of mud, stiffened with hair, feathers, grass. Eat insects, snails, seeds, soft fruit. Melodious whistles, harsh and plaintive notes.

Family Cracticidae (bellmagpies). Ten species of medium to large oscines, 25 to 58 cm (10 to 23 in.). Most black, white, and gray; some have brown phase; sexes alike or unlike. Large heavy bill, slightly to strongly hooked; nostrils bare. Wings long, pointed; legs strong, medium to long. Australia, New Guinea, and nearby islands; gregarious inhabitants of brushy plains, open forests, mangrove shores, cleared lands. Eat insects, small animals, some fruit and seeds. Fine singers with ringing, gonglike calls.

Family Ptilonorhynchidae (bowerbirds). Ten-primaried oscines of medium size, 23 to 37 cm (about 9 to 15 in.); sexes alike in a few, but males usually much brighter coloured, often with nuchal (neck) crest but never with plumes or facial wattles. Stout bill, straight to slightly curved; wings rounded; legs and feet stout, hindtoe shorter than middle toe. Seventeen species, in forests of New Guinea and northern Australia. Solitary, largely terrestrial birds; eat berries, seeds, fruits, insects, small animals. Males of most species build elaborate stages or bowers for display and courtship. Loud ringing calls, good mimics.

Family Paradisaeidae (birds of paradise). Small- to medium-sized, 10-primaried, 14 to 117 cm (about 5.5 to 46 in.), greatest length due to streaming tail feathers; greatly varied colours, most males with spectacular plumes on head, flanks, wings, or tail; some with wattles or bare skin on the head; females plain browns or grays. Bill slender to rather heavy, hooked or sickle-shaped in some. Wings rounded; legs short; feet rather stout. About 40 species, New Guinea, northern and eastern Australia, Moluccas, and adjacent islands. Solitary forest birds, rather weak fliers; males of most species have elaborate courtship displays. Eat fruit, seeds, insects, small animals. Prolonged whistles, loud shrill calls.

Family Paridae (titmice and chickadees). Ten-primaried oscines with short to medium, rounded wings; tail short to long; bill rather stout, pointed, shorter than remainder of head; nonoperculate nostrils concealed by thick feathers. Small birds, 7.5 to 20 cm (3 to 8 in.), with thick plumage usually strongly patterned in grays, yellows, brown, black, or white, never streaked, barred, or spotted; sexes usually alike. About 64 species in forests and brushlands of the Philippines, Malaysia, Eurasia, Africa, and North America to Guatemala. Active, gregarious little birds; feed on insects, seeds. Chattering notes, whistled calls.

Family Certhiidae (creepers). Small, slender, climbing birds, 9.5 to 19 cm (3.5 to 7.5 in.), with curved bills as long or longer than rest of head; operculate nostrils free of bristles or feathers; rictal bristles absent. Legs short, thin, outer toe always shorter than middle toe but much longer than inner toe; claws long, hindclaw as long or longer than hindtoe; wings rounded or pointed; tail feathers long, stiff, with pointed tips. Brownish above, streaked with black or light brown, white below; sexes alike. Six species (with the inclusion of the African spotted creeper, *Salpornis*, a controversial point) in temperate woodlands of Eurasia, Africa, North America, south to Nicaragua. Solitary birds, spiral up tree trunks from base probing bark for insects; two species inhabit rocky cliffs. Soft calls, sweet but weak songs.

Family Sittidae (nuthatches). Small, stocky, climbing birds, 9.5 to 19 cm (3.5 to 7.5 in.) with thin, pointed, usually straight bills; rounded, nonoperculate nostrils partly concealed by feathers; short rictal bristles present; tarsus short; long laterally compressed claws on large toes, hallux equal to outer toe, inner toe reaching only to second joint of middle toe; wings rather long, pointed; tail short, square, soft. Typically gray to blue above, white or brownish below; sexes alike or nearly so. About 22 species; typical nuthatches (*Sitta*) distributed across North America and Eurasia to Malaysia and the Philippines, aberrant forms in Philippines, Australo-

Papuan region. Typically in forests, a few in rocky areas. Forage on tree trunks and large branches for insects, eat some seeds and small fruits. Simple call notes and songs.

Family Climacteridae (Australian treecreepers). Six species of small, creeper-like climbing birds, 12.5 to 17.5 cm (5 to 7 in.); of uncertain ancestry and affinities. Legs short; toes long, claws long, curved, strong, especially that of hallux; tail rounded, soft; bill long, somewhat downcurved. Grayish-brown to black above, streaked below, often a lighter eye stripe and wing bar; sexes similar but with slight recognizable differences. Forests of Australia and New Guinea; rather solitary, sedentary birds; feed by spiralling up tree trunks, some forage for insects on ground. Voice high-pitched whistles.

Family Panuridae (bearded tits, parrotbills). Small, titmouse-like birds, 10 to 17.5 cm (4 to 7 in.), distinguished (with one exception, the bearded tit, *Panurus biarmicus*) from all other oscines by the strongly compressed bill, much shorter than head, markedly curved convexly, both tomia (cutting edges) sinuated; nostrils concealed by feathers; wings short, rounded; tail rather long, graduated. Plumage soft, fluffy, plain brownish above, lighter below; sexes alike or nearly so. About 19 species in brushy grasslands, scrublands, and thickets of temperate Eurasia; active, gregarious, travel in small flocks searching for seeds, insects, berries. Low twittering call notes.

Family Chamaeidae (wrentit). Small bird about 16.5 cm (6.5 in.) long, with thick fluffy plumage; plain brown above, lighter and streaked below. Bill short, rounded, pointed, slightly downcurved; nostrils exposed, operculate; rictal bristles distinct; wings short, rounded; tail long, graduated; sexes alike. The single species ranges in brushlands and forest edges from Oregon to northern Baja Californai; usually in pairs. Eats insects, small berries. Voice: a loud whistle on one pitch. Often classified in the Timaliidae.

Family Timaliidae (babblers). A poorly defined, diverse group of more than 250 species, with uncertain affinities and lineage. Small to medium-large, 9 to more than 40 cm (3.5 to 16 in.). Plumage soft, fluffy; bill highly varied, small and weak to long, straight, downcurved, often hooked. Wings short, rounded, fit close to body; legs rather large, strong. Mostly dull coloured, a few brightly coloured; sexes alike and unlike. In woodlands, scrub, and brushlands of eastern and central Eurasia, Africa, Madagascar, Philippines, Malaysia, and Australia; solitary or gregarious, usually arboreal, some terrestrial; all insectivorous, some also eat fruit, berries. Many species with noisy, harsh calls, musical songs.

Family Campephagidae (cuckoo-shrikes and minivets). About 70 species of small- to medium-sized, forest-living birds, 12.5 to 35.5 cm (5 to 14 in.). Africa, India to Japan, East Indies, Australia. Rather stout, slightly downcurved bill, notched, hooked at tip; nostrils partly concealed by short bristles; wings medium, pointed; long tail rounded or graduated; legs short, feet weak to strong. Feathers of back and rump usually with heavy shafts, thickly matted and loosely attached. Cuckoo-shrikes plain gray, black, or whitish birds, often with barred underparts; minivets brightly coloured in reds, yellows. Sexes often completely dissimilar. Eat insects, berries, small fruits. Noisy birds, with whistles and harsh calls.

Family Pycnonotidae (bulbuls). Medium-sized birds, 14 to 28 cm (5.5 to 11 in.), with soft, fluffy plumage, especially soft on lower back and rump; many species with hairlike, vaneless feathers on nape of neck. Bill usually slender, slightly downcurved; rictal bristles well-developed; feet and legs rather small; wings short; tail medium to long. Usually drably coloured in gray, brown, greenish, some with yellow, white, or red patches on head and under tail; sexes alike. About 119 species in Africa and across southern Asia to Japan, Philippines, Moluccas, Borneo; in woodlands, brushlands, cultivated regions. Eat berries, fruit, some insects. Noisy, some good singers, mimics.

†*Family Palaeoscinidae* (*Palaeoscinis*). A fossil species known only from the Miocene of California.

Family Irenidae (leafbirds, ioras, fairy bluebirds). Small to medium birds, 13 to 25 cm (5 to 10 in.); bill fairly long, slightly downcurved or hooked; legs short; wings rounded; tail square to rounded. Like bulbuls, some have hairlike feathers on nape and long fluffy rump feathers but are more brightly coloured, brown or black with contrasting yellow, green, or blue; sexes usually dissimilar. About 14 species in forests and cultivated lands from India to Philippines, south to Borneo, Java, Sumatra. Eat fruit, berries, buds, some insects. Good singers, musical whistles and flutey notes.

Family Cinclidae (dippers). The only largely aquatic oscines; small, 14 to 19 cm (5.5 to 7.5 in.), birds with plump bodies, short concave wings, short, square or rounded tail. Tarsi and toes long and stout. Bill slender, straight, pointed; nostrils with broad operculum; rictal bristles absent. Dense compact plumage plain gray, brown, or black, some with white patches on throat, breast, head; sexes alike. Four species, found along swift, rocky streams in Eurasia, North Africa, and North and South America, south to Argentina. Food: insects, aquatic larvae, and other organisms gleaned from stream bottoms. Have shrill penetrating whistled calls and chattering songs.

Family Troglodytidae (wrens). Small, 9.5 to 22 cm (3.5 to 8.5 in.), chunky birds, mostly brown, usually barred, spotted, or streaked, with white, black, or browns; sexes alike. Bill slender, medium to long, often downcurved; nostrils with operculum; rictal bristles usually indistinct but sometimes obvious with 1 or 2 well-developed. Stout legs and feet, front toes partly joined at base. Wings short, rounded, well-developed 10th (outermost) primary, at least half as long as 9th; tail short, square to rounded, often carried cocked up. The 59 species range through most of North and South America; 1 also found in Eurasia and North Africa; inhabit brushlands, forests, forest edges, rocky slopes, deserts, grassy marshes. Most forage on or near ground in undergrowth for insects, worms, other invertebrates. Highly developed song, musical, variety of bubbling, flutelike to growling notes, harsh chattering calls; some species duet.

Family Mimidae (mockingbirds, catbirds, thrashers). Slender, medium-sized, 20 to about 30 cm (8 to 12 in.); bill medium to long, nearly straight to strongly downcurved; nostrils always exposed, with overhanging membrane; rictal bristles present, but few and somewhat weak. Legs rather long, feet strong, middle toe joined at base to outer toe but not to inner toe. Wings short, rounded; tail long. Coloured brown, gray, black, bluish, underparts usually pale, often white, spotted, or streaked, some solidly coloured, some with white in wings or tail; sexes alike. Exclusively New World, the 31 species range from southern Canada, the West Indies, Galápagos, to southern Argentina and Chile. Eat fruit, seeds, insects. Famous as fine singers and mimics.

Family Turdidae (thrushes, bluebirds, nightingales, wheatears, robins, chats). A large, almost cosmopolitan group of about 305 species of small- to medium-sized birds, 11.5 to 33 cm (4.5 to 13 in.). Bill rather slender; legs and feet fairly stout, tarsus usually booted (smooth sheath, not divided into scales); 10-primaried wing rounded to pointed; tail medium, truncate or graduated, forked in a few. Colours brown, blue, gray, often blended, or in bold black, white, yellow, or red patterns; sexes alike or unalike; young usually spotted below. Virtually worldwide; absent originally only from New Zealand (now introduced there), some oceanic islands, Antarctica, and parts of the Arctic. Usually arboreal, many terrestrial, in varied habitats—forests, deserts, brushlands, grasslands, cultivated fields. Eat mostly fruit, insects, also seeds, leaves, worms, and mollusks. Most are fine singers.

Family Sylviidae (Old World warblers, kinglets). Typically small, 9 to 26 cm (3.5 to 10 in.), with slender bill; longitudinal nostrils with an operculum; medium rounded wings of 10 primaries; short to medium legs. Plumage usually browns, grays, olive greens with little pattern; some streaked or barred; some brightly, boldly marked; sexes alike or similar, young never spotted below. Nearly 300 species, worldwide except in polar regions, South America, some oceanic islands, mostly in woodlands, some in brushlands, marshes. Food essentially insects; voices pleasant, varied, well-developed song in some.

Family Polioptilidae (gnatcatchers, gnatwrens). Dainty, slender, tiny, 10 to 14 cm (4 to 5.5 in.), with long, thin, pointed bills, operculate nostrils partly exposed, and rictal bristles. Rounded wing with 10th primary much less than half as long as 9th; long rounded tail constantly moving. Blue-gray or brown above, lighter below, most with white in outer tail feathers, some with black markings on head; sexes alike or nearly so. The 11 species are active arboreal birds of open forests and semi-arid deserts from southern Canada to Argentina. Food essentially small insects. Rather weak songs of trills and warbles, simple short call notes.

Family Pachycephalidae (whistlers, shrike-thrushes, thickheads). Small- to medium-sized, stout-bodied, 13 to 28 cm (5 to 11 in.); roundish heads, rather heavy bill sometimes hooked at the tip. Wings rather long, pointed, with very short 10th primary; tail medium long, usually rounded. A few species crested or with wattles or bare patches at base of bill. Most are greenish gray to brown above, lighter below, many with yellow or dull red markings; sexes alike, a few unalike; many juvenile plumages spotted or like that of female. About 42 species in Australo-Papuan region, Malaya, Philippines, Oceania; in forests, brushlands, mangroves, savannas. Diet predominantly insects with some fruit. Melodious flutey calls, pairs often duet.

Family Maluridae (wren-warblers, emu-wrens). Small-bodied birds, 7.5 to 25 cm (3 to 10 in.), that carry the long tail cocked up over the back. Bill small, weak; wings short, rounded; legs and feet medium. Emu-wrens (*Stipiturus*) have rectrices reduced to 6 loose-barbed shafts. Most species are brightly patterned in contrasting browns, reds, shiny blues, black, and white; sexes alike or unalike. About 83 species in Australo-Papuan region, including New Zealand; in forests, scrublands, heaths. Food largely insects and other invertebrates. Many are good singers, mimics.

Family Muscicapidae (Old World flycatchers). A large (about 340 species) family of small insectivores, 7.5 to 22.5 cm (3 to 9 in.); the paradise flycatchers (*Terpsiphone*) with long tails, to 53.5 cm (21 in.). Typically with flat, broad bills, well-developed rictal bristles; short, weak legs and feet. Wings short and rounded to long and pointed; tail short and narrow to long and fanned or with long central plumes. Varied colours, many plain browns, grays, some with bright blues, reds, and black and white; some crested or with facial wattles; sexes alike and unalike, young usually spotted. Eurasia from tree line south through Africa, Australia, and in Pacific Islands to Hawaii; forests, scrublands, cultivated and riverine areas. Voices usually weak, monotonous, but well-developed song in a few.

Family Prunellidae (accentors, hedge sparrows). Small, drab, 12.5 to 17.5 cm (5 to 7 in.); slender, pointed bills, wide at base, culmen (ridge of upper bill) slightly rounded. Tarsus rather short, feet strong. Wings rounded to pointed with very short 10th primary; tail shorter than wing, square or emarginate. Colour browns or grays, usually streaked or spotted; sexes similar. Twelve species across Eurasia and northern Africa; various habitats, parks and gardens, brushlands, barren mountain slopes below snow line. Feed on or near ground on insects, berries, seeds. Thin chattering, twittering songs, metallic calls, often in flight.

Family Motacillidae (pipits, wagtails). Small, slender-bodied ground birds, 12.5 to 23 cm (5 to 9 in.). Pipits similar to larks in appearance but differ in having a bilaminate tarsus and pointed wing with 9 primaries. Wagtails have longer tails, brighter colours. Bill thin, pointed; legs long, slim, with elongated hindtoe and hindclaw (with some exceptions); tail usually edged with white or yellow. Plumage brown, streaked, mottled, sexes similar (pipits); or marked in bold black, white, or yellow patterns, sexes unalike (wagtails). About 53 species, worldwide except polar regions and some oceanic islands; in open grasslands, deserts, shores, cultivated areas. Feed on insects, spiders, mollusks, some vegetable matter. Simple repetitive song often given in flight, and short, sharp call notes.

Family Bombycillidae (waxwings). Small, arboreal, soft-plumaged, crested, 15.5 to 19 cm (6 to 7.5 in.); distinguished by waxy red "droplets" at tips of secondary wing feathers of most individuals. Bill short, swollen, slightly hooked; nostrils almost concealed by feathers, rictal bristles absent. Long pointed wing with rudimentary 10th primary; square to slightly rounded, short tail. Velvety plumage blended with browns, grays, and yellow; tail tipped with yellow or red; sexes alike. Three species in temperate Northern Hemisphere, in evergreen or birch forests. Eat berries, fruit, buds, flowers, insects. Weak chattering song, soft lisping calls.

Family Ptilogonatidae (silky flycatchers). Small, arboreal, crested, 18 to 24.5 cm (7 to 9.5 in.). Bill short, broad, deeply cleft; nostrils exposed, bordered by membrane, rictal bristles present. Wings and legs rather short; 10th primary well-developed; tail long. Soft silky plumage of solid grays, black, brown, white spots in wings and tail, some with yellow markings. Four species, found from southwestern U.S. to Panama; in brush country, desert scrub, open forest. Rather shy, active birds that feed on berries and on insects caught in flight. Weak warbling song.

Family Dulidae (palm chat). Medium-sized about 19 cm (7.5 in.); deep laterally compressed bill, strongly curved culmen, circular nostrils wholly exposed. Wings rounded, 10th primary less than half as long as 9th; tail longish; legs and toes stout. Plumage stiff and harsh, olive brown above, yellowish-white streaked with brown below; sexes alike. The single species limited to cultivated areas and open woodlands of Hispaniola and Gonave Island, West Indies. Lives on fruits and flowers. Noisy, gregarious birds with harsh chattering notes.

Family Hypocoliidae (hypocolius). Medium-sized, 17.5 cm (7 in.); short, broad bill, moderately curved culmen, operculate nostrils; nasal bristles lacking and rictal bristles poorly developed. Tarsus rather short, heavily scutellated (scaled). Soft plumage blue-gray above, lighter below, with black facial, wing, and tail markings, white wing tips, slightly crested; sexes alike. The single species inhabits semi-arid scrublands, palm groves, gardens in Tigris-Euphrates valley of Iraq, wanders to Afghanistan, western India, Arabia, Red Sea coast. Fruit eater with low calls and weak song.

Family Artamidae (wood-swallows or swallow-shrikes). Chunky-bodied, medium-sized; 15 to 21 cm (6 to 8.5 in.); unique among oscines in having powder downs. Bill stout; broad at base, moderately long, decurved, pointed; legs short; feet strong. Wings long, pointed; tail short, nearly square. Plumage compact, soft, in plain solid browns, grays, or black above, usually lighter below; sexes similar. Ten species in open lands, forest clearings, Australia east to Fiji Islands, north to Philippines, Indochina, India. Live on insects caught in flight. Voice a harsh nasal twittering.

Family Vangidae (vanga shrikes). Small- to medium-sized, 12.5 to 32.5 (5 to 13 in.), 10-primaried, arboreal oscines of varied aspect; some resemble flycatchers, thrushes, nuthatches, or typical shrikes. Bill stout, heavy, hooked, notched; with enlarged ridge in one, thin, downcurved in another; legs and feet strong; wings fairly long, rounded; tail square or rounded, moderately long. Plumage rather soft, loose, typically black or blue above, white below, some with brown or gray patches; sexes alike or unalike. Thirteen species, restricted to forests, scrublands, mangrove swamps, of Madagascar. Eat insects, small reptiles, amphibians. Chattering calls, harsh notes, shrill whistles.

Family Laniidae (shrikes). Rather small- to medium-sized 15 to 36 cm (6 to 14 in.); 10-primaried, with proportionately large heads; stout, strong, sometimes toothed, sharply hooked bills; strong legs with sharp claws; tarsus scutellate anteriorly, lamellate (plated) laterally; wings medium; tail usually long, narrow. Plumage soft, black, gray, or brown above, usually paler below in true shrikes; yellows, reds, greens in some African bush shrikes; sexes alike or unalike. Essentially Old World, 64 species across temperate Eurasia, Africa, east to Philippines, and south to New Guinea and Timor; 2 species in North America to southern Mexico; usually solitary birds, in open forests, clearings, brushlands, cultivated areas. Food large insects, small vertebrates; some impale prey on thorns for storage and to tear apart. Voices varied, some with well-developed song, several species sing antiphonally.

Family Prionopidae (helmet shrikes). Medium-sized, 19 to 25.5 cm (7.5 to 10 in.); distinguished from preceding family by their fully scutellate tarsus and stiff forehead feathers projecting forward, almost concealing the nostrils; most with conspicuous wattles surrounding the eyes. Bill stout, hooked; legs short, strong; wings and tail medium to long. Plumage black above, white or buffy below in bold patterns; sexes similar. Nine species, in brushlands, scrub, and forests of Africa, south of the Sahara. Gregarious arboreal birds that hunt in small flocks for insects. Harsh chattering notes, nasal humming sounds, snap bills audibly.

Family Sturnidae (starlings, mynas, oxpeckers). Stocky, medium-sized, 16.5 to 42 cm (6.5 to 16.5 in.); very short but visible 10th primary. Bill pointed, straight, or slightly arched, swollen near the tip in oxpeckers (*Buphagus*); tongue flat, not tubular (as in some relatives). Legs and feet stout, strong, tarsus with unbroken plates behind. Typically dark coloured with metallic sheen, some brown or gray with white, yellow, or red markings, some with facial wattles, a few crested; sexes alike or unalike. About 107 living, 4 recently extinct species in all types of wooded and agricultural lands; temperate Eurasia through Africa, northern Australia, East Indies, east to Tuamotu Archipelago; a few introduced widely elsewhere (*e.g.*, North America). Almost omnivorous; wide range of vegetable and animal foods. Garrulous, varied notes, calls, whistles, some excellent mimics.

Family Meliphagidae (honeyeaters). Small- to medium-sized, 10 to 40 cm (4 to 16 in.); long, protractile, brush-tipped tongue curled at the sides to form a tube. Bill slender, pointed, downcurved, upper cutting edge serrated; nostrils unfeathered, with leathery operculum. Wings long, pointed, 10th primary about half the length of 9th; tail medium to long. Legs short to medium, tarsus scaled anteriorly. Drab browns, yellows, grays, or bold patterns of black, red, white; sexes alike or unalike. About 168 living, 4 recently extinct species, from Australia and New Zealand through the Papuan region, north to Marianas, east to Hawaii, 2 species in southern Africa. Habitat forests, brushlands, cultivated lands. Food chiefly nectar, insects, some fruit. Voices loud, varied; musical song in many.

Family Nectariniidae (sunbirds, spider hunters). Small, 9 to 22 cm (3.5 to 8.5 in.). Long, slender, downcurved, pointed bill, finely serrated near tip; tongue not brush-tipped but partly tubular, projectile, divided at tip; nostrils rounded, open, operculate; no rictal or nasal bristles. Wings rounded, short, 10th primary variable in length but always present; tail square,

medium to long, pointed, sometimes with elongated central feathers. Legs short, stout, tarsus with anterior transverse scales; hallux and claws short. Many sunbirds with vivid colours, metallic sheens, and bright patches in male; females much duller; other sunbirds and spider hunters dully coloured in both sexes. About 116 species, in forests, scrub country, mangroves; Asia and Africa south of Sahara, east to central China, Philippines, Solomons, south through Malaya to northern Australia. Active arboreal birds, sunbirds are Old World counterparts of New World hummingbirds but not as skilled fliers. Feed on nectar, insects, some small fruit. Weak call notes, faint songs.

Family Dicaeidae (flowerpeckers, pardalotes, diamond birds). Small chunky birds 7.5 to 18 cm (3 to 7 in.), with short necks, legs, tails. Bill usually short, stout, relatively straight; slender and curved in a few, edges of distal (outer) 3rd serrated; distal half of short tongue deeply cleft, the edges curled into 2 slender semitubular tips. Wings rather long, 10th primary usually vestigial, absent in a few. Bright-coloured males dark, glossy (rarely metallic) above, lighter below, often with red or yellow patches on crown, rump, breast; females usually much duller; some species dull-coloured in both sexes. About 58 species in forests, scrublands; India to southern China, Philippines, east through Solomons, and south to Australia. Active arboreal birds, feed around flowers on nectar, small insects, berries (particularly mistletoe). Sharp metallic twittering calls, a few have warbling song.

Family Zosteropidae (white-eyes). Relatively uniform group of little birds, 10 to 15 cm (4 to 6 in.) long, yellowish- or reddish-brown to olive green above, lighter below, typically with a conspicuous white eye-ring; sexes alike. Bill short, slender, pointed, slightly decurved; tongue brush-tipped; rictal and nasal bristles absent. Wings short, rounded, 10th primary usually lacking; tail medium, square; legs and claws short, feet strong. About 90 species in forests, brush country, mangroves; Africa south of Sahara, across southern Asia to Korea, Japan, south to Australia, New Zealand, east to Carolines, Samoa. Travel in restless flocks, feed on nectar, insects, fruit. Weak voices, pleasant twittering, warbling songs.

Family Cyclarhidae (pepper-shrikes). Rather heavily built, medium-sized, 14.5 to 17.5 cm (5.5 to 7 in.), with large heads; stout, strongly hooked, laterally compressed bills; nostrils rounded, exposed; rictal bristles present but weak. Wings short, rounded, 10th primary half length of 9th; tail medium; legs and feet strong. Loose-webbed plumage olive green above, yellow to buffy-white or gray below, rufous stripe over eye; sexes similar. One species southern Mexico to Guianas, Uruguay, northern Argentina; a second in Colombia and Ecuador; open forest, brushland. Weak fliers, travel singly or in pairs, moving deliberately through foliage. Eat large insects, fruit. Repetitive warbling song, harsh scolding calls.

Family Vireolaniidae (shrike-vireos). Small, poorly known; 14.5 to 18.5 cm (5.5 to 7 in.). Rather stout, hooked bill; nostrils oval, exposed; rictal bristles inconspicuous. Wings short, rounded, 10th primary reduced to half length of 9th; tail short to medium; legs short, stout. Plumage silky, loose-webbed, greenish above, yellowish to white below, head marked with yellow, gray, blue, black, or white; sexes similar. Three species; in rain-forest treetops from southern Mexico to central Brazil. Food fruits, insects. Loud repetitive whistled song.

Family Vireonidae (vireos, greenlets). Rather plainly coloured, small arboreal birds, 10 to 17.5 cm (4 to 7 in.), mostly brownish-gray to olive green above, yellow, grayish, or white below; plumage never streaked or spotted; some with light eye-rings, eye stripes, wing bars; sexes alike. Bill fairly heavy, slightly hooked and notched; nostrils ovate, operculate, partly exposed; rictal bristles inconspicuous; legs short, strong. Wings long, pointed to short, rounded, 10th primary very short or vestigial. About 37 species in all types of woodlands from central Canada to Uruguay and Argentina. Usually solitary inhabitants of forest edges; seek insects on leaves and branches, eat some berries, fruit. Compulsive singers; songs of repeated, often melodious, phrases; harsh scolding notes.

Family Drepanididae (Hawaiian honeycreepers). Small forest birds, 11.5 to 22 cm (4.5 to 8.5 in.), limited to the Hawaiian Islands; 14 species living, 10 of them rare and local, 8 species recently extinct. Bill extremely varied, short to long, thin to thick, straight to extremely downcurved, pointed to hooked, but never serrate or notched. Wings pointed, 10th primary vestigial or absent; tail medium, truncate or slightly forked; legs short to medium; feet strong. Plumage plain brown, olive green, yellow, red, gray, or black, never metallic or glossy; sexes unlike or similar with female smaller. Solitary or in small flocks; eat nectar, insects, seeds, fruit. Warbling songs, clear calls.

Family Parulidae (wood warblers, bananaquits). Dainty, small, 10 to 18.5 cm (4 to 7.5 in.); pointed wings of 9 primaries, medium 12-feathered tail. Bill usually slender, pointed, culmen slightly downcurved, flattened with pronounced rictal bristles in a few but never notched or hooked; tongue moderately slender, tip variably bifid (divided) or fimbriate (fringed). Legs medium, hindclaw never elongated. Colours varied but never metallic or glossy; grays and olive browns dominant ground colours, often brightly patterned with yellows, reds, blue, black, or white; sexes alike or unlike. About 119 species in forests and brushlands from tree line of North America to southern South America. Active, arboreal (a few terrestrial); feed mostly on insects, occasional fruits, berries, seeds, 1 group (bananaquits) sip nectar. Voices typically weak, thin, high-pitched; a few with loud and well-developed songs.

Family Zeledoniidae (wrenthrush). The single species is small, 11.5 (4.5 in.), with 9 primaries. Bill weak and flattened vertically; legs long, feet large, strong; wings and tail short, rounded. Plumage soft, olive brown above, slate gray below, crown black-bordered orange; sexes alike. Resident in humid cloud forests of Costa Rican, western Panamanian highlands. Shy, retiring birds with weak, fluttery flight, seldom fly more than 20 m (65 feet), creep and hop on forest floor. Food mainly insects. Voice thin, high-pitched, repetitive whistles.

Family Icteridae (orioles, troupials, oropendolas, caciques, blackbirds, grackles, cowbirds, meadowlarks). A heterogeneous group, medium to large, 16 to 54 cm (about 6 to 21 in.). Straight, pointed, unnotched, conical bill, with culmen extending more or less onto forehead; operculate nostrils, never concealed; rictal bristles lacking. Wings mostly long, pointed; tail short to rather long; legs and feet strong. Plumage often solid black with metallic gloss, or multicoloured in bold patterns of black and red, orange, yellow, brown; sexes usually differ, female smaller. About 88 species, throughout Western Hemisphere (except polar regions), in a wide variety of habitats. Many species gregarious; some colonial or semicolonial nesters. Eat a wide range of animal and vegetable foods. Usually loud-voiced, with harsh, whistling, or bubbling calls; well-developed song in some.

Family Tersinidae (swallow-tanager). Single tanager-like species, 15 cm (6 in.), differing from Thraupidae mainly in peculiar palate, expandable throat, and broad, flat, swallowlike bill; longer wings, shorter legs. Male shining turquoise with black throat, face, sides barred black, white below; female green with grayish-white barring. Lowland and mountain forests; Panama, Trinidad, south into Argentina. Somewhat gregarious; strong fliers; feed on insects, fruit. Voice monotonous chirps, almost no song.

Family Thraupidae (tanagers, euphonias, honeycreepers, and diglossas). Small to medium, compactly built, 9 primaries; 8 to 30 cm (about 3 to 12 in.), most under 20 cm. Bill with commissure (meeting of edges) not abruptly angled or bent at the base; mandibular tomium not distinctly angled (never toothed) near the base. Bill usually conical, short to medium, generally notched, toothed, or hooked at tip; rictal bristles present. Nectar-feeding species with specialized bill and brush-tipped tongue. Wings pointed, short to long; tail short to medium, truncate, emarginate, or rounded. Generally brightly coloured, in contrasting patterns of black, white, yellow, red, blue, green, brown; sexes alike or unalike. About 223 species in forests, brushlands; temperate North America to Brazil, Argentina. Arboreal birds; flight strong but not sustained in non-migratory species. Eat berries, fruit, insects, nectar. Voices varied, song well-developed in a few but usually short calls and warblings.

Family Catamblyrhynchidae (plush-capped finch). Finch-like, 9 primaries, 15 cm (6 in.), with stubby, conical, slightly hooked bill and a distinctive patch of short, stiff, velvety tipped, orange-yellow feathers on the forecrown. Wings short, rounded; tail medium; legs stout, feet strong. Male bluish-gray above, chestnut below, nape and sides of head black; female similar but duller. Mountain forests of Colombia, Venezuela south to Peru, Bolivia. A little-known species of uncertain relationships and systematic position.

Family Fringillidae (New World seedeaters). Small, 9 primaries; 10 to 27 cm (4 to 11 in.), with short, stout, conical, pointed bills. Commissure distinctly angled or deflexed at the base, mandibular tomium elevated, often toothed, gonys (midline ridge of lower bill) more than half length of upper bill; rictal bristles usually obvious. Wings short, rounded to long, pointed; tail short to long, tarsus relatively long. Colours and patterns varied; sexes alike or unlike. About 315 species almost cosmopolitan (absent from Oceania, East Indies, Madagascar, Antarctica; introduced in Australia, New Zealand) in all terrestrial habitats. Gregarious or solitary, terrestrial or arboreal birds; food mainly seeds, some other vegetable mat-

ter, insects. Call a simple chirp, often well-developed song; flight song rare. Family includes buntings, juncos, towhees, grosbeaks, cardinals, longspurs, Darwin's finches, chaffinch, yellow hammer, most sparrows.

Family Carduelidae (goldfinches, siskins, rosefinches, redpolls, crossbills, bullfinches, hawfinches, canaries). Small to medium-sized, 10 to 25 cm (4 to 10 in.), with short, stout to slender pointed bills (mandibles crossed in one genus); gonys less than half length of upper bill. Wings somewhat rounded, medium length, 10th primary present but less than half length of 9th; tarsus relatively short. Colours variable but browns, reds, and yellows predominate; some streaked or mottled, some plain patterns; sexes alike or unalike. About 112 species in woodlands, brushlands, worldwide except Pacific Islands; introduced in Australia. Gregarious birds with strong, undulating flight. Eat seeds, buds, berries, some insects. Highly developed songs, sing in flight.

Family Estrildidae (waxbills, some finches). Small, 7.5 to 15 cm (3 to 6 in.), bill stout, short, pointed; gonys less than half the length of upper bill. Wings rounded, short to pointed, medium length; 10th primary present but almost vestigial; tarsus relatively short. Great variety of bright and somber colours and patterns; sexes alike or unalike. About 107 species; forest edges, clearings, grasslands, and marshes of Africa, southern Asia, East Indies, Australia, some South Pacific islands. Highly gregarious, active, ground feeders. Food: seeds, some berries, insects. Song poorly developed, weak chirps, chatterings, buzzes.

Family Ploceidae (weaverfinches). Small, stoutly built seed-eaters, mostly 10 to 25 cm (4 to 10 in.), a few long-tailed species to 50 cm (males only). Short, stout, conical, pointed bills; gonys less than half length of upper bill; rictal bristles lacking; tarsus relatively short. Wings short to long, rounded to pointed, 10th primary present, much reduced; tail short to very long. Colours, patterns widely varied; sexes alike or unalike. About 156 species in all terrestrial habitats, mainly in Africa, also in Eurasia, Madagascar, Malaysia; a few (*e.g.*, the house sparrow, *Passer domesticus*) widely introduced elsewhere. Usually gregarious; food mainly seeds, some other vegetable matter, insects. Harsh, monotonous voices, song poorly developed.

Critical appraisal. The classification of the Passeriformes is currently in a state of flux. Of the several major works on the order since 1950, no two have agreed on the sequence of families, and each author has proposed innovations of his own. Ornithologists who are principally concerned with arranging museum displays, preparing national lists, or studying avian biology in the field may regard any change in classification as an unnecessary and unmitigated nuisance, and a few such workers have tried to prevent further change by having avian taxonomy formally "standardized." An attempt to do so at the XIVth International Ornithological Congress in 1966 failed in part because taxonomists opposed it vigorously and in part because its proponents could not agree on which of the several passerine sequences currently in use should be adopted.

Difficulties in preparing a lineal sequence

A main difficulty with any lineal sequence is the impossibility of making it reflect evolution's three-dimensional development, and it will be many years before enough evidence is in to satisfy all taxonomists. Most taxonomists place adjacent in the linear order those groups that they believe to be most closely related, but if three groups are believed to be equally interrelated, no linear sequence can express the relationship. The simplest solution for those who want a rigid, nonvarying sequence is to list all families alphabetically, and under each the genera and species alphabetically in turn. Even this system, which of course gives no indication of relationships, is not foolproof, for all scientific names are subject to change through the laws of priority as well as by inevitable changes in various taxa as more is learned about them. Hence it seems far better to retain the systems currently in use, leaving them fluid enough to be changed when changes are shown justifiable.

Since the late 19th century, when the many advances in taxonomic thought of the previous century began to crystallize and bear fruit, three main passerine sequences have dominated the world bird lists. The first, proposed originally by R.B. Sharpe of the British Museum in 1877, but based in part on the published and unpublished work of others, places the crows at the summit of avian evolution (hence at the end of a modern lineal sequence), ostensibly (although never so stated directly) on the basis of their alleged high intelligence. This sequence was adopted by the German ornithologist Ernst Hartert in his monumental *Die Vögel der paläarktischen Fauna*, published in 1903, and subsequently by most other European ornithologists.

The second sequence, placing the thrushes at the end, was in general usage, particularly in North America, until the late 1920s.

The third sequence, originally proposed in 1926 by two Americans, Alexander Wetmore and Waldron DeWitt Miller, but also based partly on the earlier work of others, places the crows, actually a rather generalized group, near the base of the oscine family tree, and in top place puts the so-called nine-primaried oscines, dominated by the seed-eating fringillids, because, as Wetmore explained in a 1960 paper (*A Classification for the Birds of the World*),

> this group is the modern expression of a main core or stem that through the earlier Tertiary periods has given rise to more specialized assemblages. . . . Further specialization is apparent in some parts of the existing fringilline assemblage that, if undisturbed, may lead to further differentiation.

This sequence immediately became standard for North American works and has remained so for these and certain international lists ever since, with only minor departures and rearrangements.

Within the suboscines, research is beginning to suggest that the three Old World families long included within the suborder Tyranni should be removed, leaving that suborder purely New World in distribution. It has been suggested that the asities (Philepittidae) may be more closely related to the broadbills, with which they share some features of the syrinx and sternum, than to the New World tyrannoid groups. The relationships of the asities, pittas, and New Zealand wrens will remain uncertain until further study can provide a better understanding of their biology and anatomy.

Nearly all recent classifications have grouped the eight to ten families of New World suboscines in a suborder, Tyranni, based on syringeal features studied by the 19th-century anatomist A.H. Garrod. An American ornithologist, P.L. Ames, following a re-examination of the syringeal anatomy, suggested in 1965 that one superfamily of the suborder, Furnarioidea (the antbird–ovenbird group), should not be united to the other, Tyrannoidea (tyrant flycatchers and allies), on the basis of the syrinx but should be elevated to subordinal rank, a view supported by features of the pterylosis and sternal and cranial osteology. This suggestion, made initially in 1907 by a British anatomist, W.P. Pycraft, but without supportive evidence, was taken up in a classification by R.W. Storer in 1971 (see the Table).

As more suboscines are studied in detail, some genera may be found to have been misclassified; particularly vulnerable are certain forms in the woodcreeper–ovenbird (Dendrocolaptidae–Furnariidae), antbird–tapaculo (Formicariidae–Rhinocryptidae), and cotinga–manakin–flycatcher (Cotingidae-Pipridae-Tyrannidae) complexes. A group of small South American ground birds known as gnateaters and antpipits was separated from the antbirds in 1882 as a family, Conopophagidae, and generally recognized for nearly 90 years. Careful study in 1968, however, revealed that the family was an artificial one; one genus (*Conopophaga*) has been returned to the Formicaridae, the other (*Corythopis*) placed in the Tyrannidae. Recent studies of egg white proteins by American biologist C.G. Sibley and colleagues suggest that the palm chat (*Dulus*) may be closer to the weaverfinches than to the waxwings; that the Cape sugarbird (*Promerops*) may be a starling rather than a honeyeater; and that the phainopepla (*Phainopepla*) may be most closely related to the solitaires (*Myadestes*), which are currently placed in the thrush family. A number of widely accepted oscine families invite similar examination. Only additional study will solve these problems and (hopefully) provide a better understanding of the position of aberrant genera, such as *Rupicola* and *Menura*.

Comparison of Systems of Passeriform Classification

Encyclopaedia article	Wetmore, 1960	Peters Checklist	Storer, 1971
Order Passeriformes	**Passeriformes**	**Passeriformes**	**Passeriformes**
Suborder Eurylaimi	*Eurylaimi*	*Eurylaimi*	*Eurylaimi*
Family Eurylaimidae (broadbills)	Eurylaimidae	Eurylaimidae	Eurylaimidae
Suborder Tyranni	*Tyranni*	*Tyranni*	*Furnarii*
Superfamily Furnarioidea	Furnarioidea	Furnarioidea	
Family Dendrocolaptidae (woodcreepers)	Dendrocolaptidae	Dendrocolaptidae	Dendrocolaptidae (incl. Furnariidae)
Furnariidae (ovenbirds)	Furnariidae	Furnariidae	
Formicariidae (antbirds) (incl. Conopophagidae, part)	Formicariidae Conopophagidae (antpipits)	Formicariidae Conopophagidae	Formicariidae (incl. Conopophagidae, part)
Rhinocryptidae (tapaculos)	Rhinocryptidae	Rhinocryptidae	Rhinocryptidae
Superfamily Tyrannoidea	Tyrannoidea	Tyrannoidea	*Suborder Tyranni*
Family Cotingidae (cotingas)	Cotingidae	Cotingidae	Cotingidae
Pipridae (manakins)	Pipridae	Pipridae	Pipridae
Tyrannidae (tyrant flycatchers)	Tyrannidae	Tyrannidae	Tyrannidae
Oxyruncidae (sharpbill)	Oxyruncidae	Oxyruncidae	Oxyruncidae
Phytotomidae (plantcutters)	Phytotomidae	Phytotomidae	Phytotomidae
Pittidae (pittas)	Pittidae	Pittidae	
Xenicidae (New Zealand wrens)	Acanthisittidae (=Xenicidae)	Xenicidae	*Suborder Menurae*
Philepittidae (asities)	Philepittidae	Philepittidae	Atrichornithidae Menuridae
Suborder Menurae	*Menurae*	*Menurae*	*Suborder?*
Family Menuridae (lyrebirds)	Menuridae	Menuridae	Xenicidae
Atrichornithidae (scrub-birds)	Atrichornithidae	Atrichornithidae	Pittidae Philepittidae
Suborder Passeres	*Passeres*	*Passeres*	*Passeres*
Family Alaudidae (larks)	Alaudidae	Alaudidae	Palaeospizidae
Palaeospizidae (fossil only)	Palaeospizidae		Alaudidae
Hirundinidae (swallows)	Hirundinidae	Hirundinidae	Hirundinidae
Dicruridae (drongos)	Dicruridae	Motacillidae	Campephagidae
Oriolidae (Old World orioles)	Oriolidae	Campephagidae	Pycnonotidae
Corvidae (crows and jays)	Corvidae	Pycnonotidae	Irenidae
Callaeidae (wattlebirds)	Cracticidae	Irenidae	Laniidae
Grallinidae (mudnest builders)	Grallinidae	Laniidae (incl. Prionopidae)	Vangidae
Cracticidae (bellmagpies)	Ptilonorhynchidae	Vangidae (incl. Hypositтidae)	Bombycillidae
Ptilonorhynchidae (bowerbirds)	Paradisaeidae	Bombycillidae (incl. Ptilogonatidae)	Dulidae
Paradisaeidae (birds of paradise)	Paridae	Dulidae	Motacillidae
Paridae (titmice)	Sittidae	Cinclidae	Cinclidae
Certhiidae (creepers)	Hypositтidae (coral-billed nuthatch)	Troglodytidae	Troglodytidae
Sittidae (nuthatches)	Certhiidae	Mimidae	Mimidae
Climacteridae (Australian treecreepers)	Paradoxornithidae (=Panuridae)	Prunellidae	Prunellidae
Panuridae (bearded tits, parrotbills)	Chamaeidae	Muscicapidae	Muscicapidae
Chamaeidae (wrentit)	Timaliidae	Subfamily Turdinae	Subfamilies not listed, but family constituted basically as in Peters Checklist arrangement.
Timaliidae (babblers)	Campephagidae	Orthonychinae (log runners)	
Campephagidae (cuckoo-shrikes)	Pycnonotidae	Timaliinae	
Pycnonotidae (bulbuls)	Palaeoscinidae	Panurinae	
Palaeoscinidae (fossil only)	Chloropseidae (leafbirds, ioras)	Picathartinae (rockfowl)	
Irenidae (leafbirds, ioras, fairy bluebirds)	Cinclidae	Polioptilinae	
Cinclidae (dippers)	Troglodytidae	Sylviinae	
Troglodytidae (wrens)	Mimidae	Malurinae	
Mimidae (mockingbirds and allies)	Turdidae	Muscicapinae	
Turdidae (thrushes)	Zeledoniidae	Platysteirinae (wattle-eyes)	
	Sylviidae	Monarchinae (monarch flycatchers)	
Sylviidae (Old World warblers, incl. Regulidae)	Regulidae (kinglets)	Pachycephalinae	
Polioptilidae (gnatcatchers)	Muscicapidae	Aegithalidae (long-tailed tits)	
Pachycephalidae (whistlers)	Prunellidae	Remizidae (penduline titmice)	Aegithalidae
Maluridae (wren-warblers)	Motacillidae	Paridae	Climacteridae
Muscicapidae (Old World flycatchers)	Bombycillidae	Sittidae	Rhabdornithidae
Prunellidae (accentors)	Ptilogonatidae	Certhiidae	Certhiidae
Motacillidae (wagtails)	Dulidae	Rhabdornithidae (Philippine creepers)	Sittidae
Bombycillidae (waxwings)	Artamidae	Climacteridae	Paridae
Ptilogonatidae (silky flycatchers)	Vangidae	Dicaeidae	Remizidae
Dulidae (palm chat)	Laniidae	Nectariniidae	Dicaeidae
Hypocoliidae (hypocolius)	Prionopidae	Zosteropidae	Nectariniidae
Artamidae (wood-swallows)	Cyclarhidae	Meliphagidae	Zosteropidae
Vangidae (vanga shrikes)	Vireolaniidae	Emberizidae	Meliphagidae
Laniidae (shrikes)	Callaeidae	Subfamily Emberizinae (buntings)	Oriolidae
Prionopidae (helmet shrikes)	Sturnidae	Catamblyrhynchinae	Dicruridae
Sturnidae (starlings)		Cardinalinae (cardinal-grosbeaks)	
	Meliphagidae	Thraupinae	Callaeidae
Meliphagidae (honeyeaters)	Nectariniidae	Tersininae	Grallinidae
Nectariniidae (sunbirds)	Dicaeidae	Parulidae	Artamidae
Dicaeidae (flowerpeckers)	Zosteropidae	Drepanididae	Cracticidae
Zosteropidae (white-eyes)	Vireonidae	Vireonidae	Ptilonorhynchidae
Cyclarhidae (pepper-shrikes)	Coerebidae (honeycreepers)	Icteridae	Paradisaeidae
Vireolaniidae (shrike-vireos)	Drepanididae	Fringillidae	Corvidae
Vireonidae (vireos)	Parulidae	Estrildidae	Sturnidae
Drepanididae (Hawaiian honeycreepers)	Ploceidae	Ploceidae	Ploceidae
Parulidae (wood warblers)	Icteridae	Sturnidae	Estrildidae
Zeledoniidae (wrenthrush)	Tersinidae	Oriolidae	Fringillidae
Icteridae (New World orioles and allies)	Thraupidae	Dicruridae	Vireonidae
Tersinidae (swallow-tanager)	Catamblyrhynchidae	Callaeidae	Drepanididae
Thraupidae (tanagers)	Fringillidae	Grallinidae	Parulidae
Catamblyrhynchidae (plush-capped finch)		Artamidae	Emberizidae
Fringillidae (New World seedeaters)		Cracticidae	Icteridae
Carduelidae (goldfinches and allies)		Ptilonorhynchidae	
Estrildidae (waxbills)		Paradisaeidae	
Ploceidae (weaverfinches)		Corvidae	

Within the oscines, one of the greatest problems is a satisfactory delineation and arrangement of the many superficially similar groups. Most taxonomists agree that the oscines contain three large groups (see the Table): (1) the crows, Old World orioles, birds of paradise, and allies; (2) the thrushes, babblers, Old World flycatchers, Old World warblers, kinglets, and allies; and (3) the finches, icterids, tanagers, and allies. Authorities disagree on how to subdivide these groups into "manageable" sections, precisely where to draw the lines between

some families, and even whether to assign them familial or lesser taxonomic rank. Other groups, currently assigned family rank, are still of moot affinities and defy all efforts to place them in a linear sequence completely acceptable to everyone; among these are the larks, swallows, titmice, wrens, dippers, mimid thrushes, honeyeaters, waxwings, and the palm chat.

As present-day students are delving more deeply into and reassessing the anatomical evidence on which most passerine families were erected a century ago, they are finding that much oscine classification was based originally on a few basic characters, weighted too strongly by the opinions of individual taxonomists. Bringing to bear such existing new lines of evidence as ethology, serology, parasitology, and protein analysis is helping to clear some of the long unchallenged beliefs that have clouded the main issues. Hopefully these developments will encourage taxonomists to abandon some of their tenacious opinions—*e.g.*, that the crows represent the apex of passerine evolution. They should then be able to draw familial lines at the point at which the evidence indicates they should be, regardless of the relative size of each group.

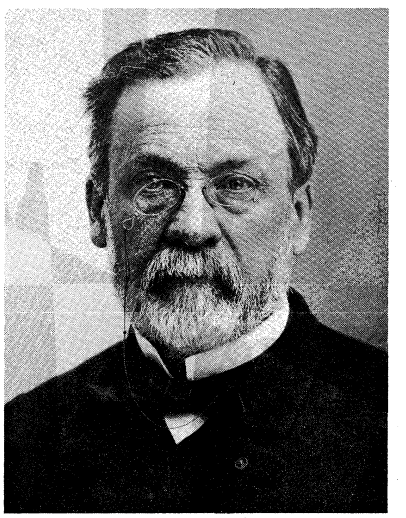
BIBLIOGRAPHY. General works, written in a popular style, containing a wealth of information on the biology of passerines, include O.L. AUSTIN, JR., *Birds of the World* (1961) and *Families of Birds* (1971); JAMES FISHER and R.T. PETERSON, *The World of Birds* (1964); and E.T. GILLIARD, *Living Birds of the World* (1958). Some regional faunal works contain much information on passerines. For the Eurasian region, see D.A. BANNERMAN, *The Birds of the British Isles*, vol. 6–12 (1957–63); CHARLES VAURIE, *The Birds of the Palearctic Fauna: A Systematic Reference, Order Passeriformes* (1959); H.F. WITHERBY *et al.*, *The Handbook of British Birds*, 5 vol. (1938–41). For tropical Asia, see E.C. STUART BAKER, *The Fauna of British India, Including Ceylon and Burma*, vol. 1–4 (1922–27); SALIM ALI and S.D. RIPLEY, *Handbook of the Birds of India and Pakistan, Together with Those of Nepal, Sikkim, Bhutan and Ceylon*, vol. 4 ff. (1970–); A.L. RAND and E.T. GILLIARD, *Handbook of New Guinea Birds* (1968); B.E. SMYTHIES, *The Birds of Burma* (1940) and *The Birds of Borneo* (1960). For Australasia, see G.M. MATHEWS, *The Birds of Australia*, 12 vol. (1910–26); D.L. SERVENTY and H.M. WHITTELL, *Birds of Western Australia*, 3rd ed. (1962); W.R.B. OLIVER, *New Zealand Birds*, 2nd ed. rev. (1955); ERNST MAYR, *Birds of the Southwest Pacific* (1945); J.E. DU PONT, *Philippine Birds* (1972). For Africa, see D.A. BANNERMAN, *The Birds of Tropical West Africa*, 8 vol. (1930–51) and *The Birds of West and Equatorial Africa*, 2 vol. (1953); J.P. CHAPIN, "The Birds of the Belgian Congo," *Bull. Am. Mus. Nat. Hist., 65, 75*, 4 pt. (1932–54); C.W. MACKWORTH-PRAED and C.H.B. GRANT, *Birds of Eastern and North-eastern Africa* (1955) and *Birds of the Southern Third of Africa* (1963). For North America, see A.C. BENT, "Life Histories of North American Birds . . . ," 8 *Bulletins* of the United States National Museum (1942–68); E.H. FORBUSH, *Birds of Massachusetts and Other New England States*, vol. 2–3 (1927–29); I.N. GABRIELSON and F.C. LINCOLN, *The Birds of Alaska* (1959); W.E. GODFREY, *The Birds of Canada* (1966); and volumes on birds of individual states. For the Caribbean and Central and South America, see ALEXANDER WETMORE and B.H. SWALES, "The Birds of Haiti and the Dominican Republic," *Bull. U.S. Natn. Mus. 155* (1931); JAMES BOND, *Birds of the West Indies*, 2nd ed. (1971); R. MEYER DE SCHAUENSEE, *The Birds of Colombia, and Adjacent Areas of South and Central America*, 2nd ed. (1971), *The Species of Birds of South America and Their Distribution* (1966), and *A Guide to the Birds of South America* (1970); ALEXANDER WETMORE, "Observations on the Birds of Argentina, Paraguay, Uruguay, and Chile," *Bull. U.S. Natn. Mus. 133* (1926); A.F. SKUTCH, *Life Histories of Central American Birds*, vol. 2–3 (Pacific Coast Avifauna, no. 34–35; 1960–69); *Life Histories of Central American Highland Birds* (1967); and *Studies of Tropical American Birds* (1972); General works, somewhat more technical in nature are ALFRED NEWTON, *A Dictionary of Birds* (1896); A.L. THOMSON (ed.), *A New Dictionary of Birds* (1964); JOSSELYN VAN TYNE and A.J. BERGER, *Fundamentals of Ornithology* (1959). Specialized works and technical review articles include P.L. AMES, "The Morphology of the Syrinx in Passerine Birds," *Bull. Peabody Mus. Nat. Hist. 37* (1971); C.G. SIBLEY, "A Comparative Study of the Egg-White Proteins of Passerine Birds," *ibid. 32* (1970); K.E.L. SIMMONS, "Anting and the Problem of Self-Stimulation," *J. Zool.*, 149: 145–162 (1966); E.F. POTTER, "Anting in Wild Birds. Its Frequency and Probable Purpose," *Auk*, 87:692–713 (1970); P. BRODKORB, "Origin and Evolution of Birds," in D.S. FARNER and J.R. KING (eds.), vol. 1, *Avian Biology* (1971). On passerine classification, see D. AMADON, "Remarks on the Classification of the Perching Birds," *Proc. Zool. Soc., Calcutta, Mookerjee Mem. Vol.*, pp. 259–268 (1957); J. DELACOUR and CHARLES VAURIE, "A Classification of the Oscines (Aves)," *Contr. Sci.*, no. 16, pp. 1–6 (1957); ERNST MAYR and D. AMADON, "A Classification of Recent Birds," *Am. Mus. Novit. 1496* (1951); ERNST MAYR and J.C. GREENWAY, JR., "Sequence of Passerine Families," *Breviora 58* (1956); J.L. PETERS, *Check-list of Birds of the World*, vol. 7–15, ed. by R.A. PAYNTER, JR., ERNST MAYR, and J.C. GREENWAY, JR. (1951–); R.W. STORER, "Classification of Birds," in D.S. FARNER and J.B. KING (eds.), vol. 1, *Avian Biology* (1971); ALEXANDER WETMORE, "A Classification for the Birds of the World," *Smithson. Misc. Collns.*, vol. 139, no. 11 (1960).

(M.H.C./O.L.A.)

Pasteur, Louis

The scientific contributions of Louis Pasteur, French chemist and microbiologist, were among the most varied and valuable in the history of science and industry. It was he who proved that micro-organisms cause fermentation and disease; he who originated and was the first to use vaccines for rabies, anthrax, and chicken cholera; he who saved the beer, wine, and silk industries of France and other countries; he who performed important pioneer work in stereochemistry; and he who originated the process known as pasteurization.

Archives Photographiques



Pasteur.

Born December 27, 1822, at Dole in eastern France, Pasteur was the descendant of generations of tanners. His great-grandfather had been an indentured labourer who had purchased his freedom. In his youth Pasteur showed little interest in anything but drawing and produced a number of pastels, portraits of his parents and friends. After attending primary and secondary schools in Arbois where his family had moved, and then in Besançon, Pasteur earned his *bachelier ès lettres* (bachelor of arts) in 1840 and *bachelier ès sciences* (bachelor of science) at the Royal College in Besançon in 1842. The following year he was admitted to the École Normale Supérieure, the famous teachers' college in Paris. He became *licencié ès sciences* (master of science) in 1845 and after acquiring an advanced degree in physical sciences, he won his *docteur ès sciences* (doctor of philosophy) in 1847. On May 22, 1848, at the age of 26, he presented before the Paris Academy of Sciences a paper reporting a remarkable discovery he had just made—that certain chemical compounds were capable of splitting into a "right" component and a "left" component, one component being the mirror image of the other. His discoveries arose out of a crystallographic investigation of tartaric acid, an acid formed in grape fermentation that is widely used commercially, and racemic acid—a new, hitherto unknown acid that had been discovered in certain industrial processes in the Alsace region. Both acids not only had identical

Stereochemical investigations

chemical compositions but also had the same structure; yet they showed marked differences in properties. The German chemist Eilhardt Mitscherlich (1794–1863) had shown that while ordinary commercial tartaric acid affects the rotation of plane polarized light, the unknown acid had no such effect. With the help of his own chemical methods Pasteur supplied the clue to this enigma by showing that the salts of the racemic acid consisted of two types of crystals that were mirror images of one another (like right- and left-hand gloves). When separated the two types of crystals rotated plane polarized light to the same degree but in opposite directions (one to the right, or clockwise, and the other to the left, or counterclockwise). One of the two crystal forms of racemic acid proved to be identical with the tartaric acid of the fermentation. As Pasteur showed further, one component of the racemic acid (that identical with the tartaric acid from fermentation) could be utilized for nutrition by micro-organisms, whereas the other, which is now termed its optical antipode, was not assimilable by living organisms. On the basis of these experiments, Pasteur elaborated his theory of molecular asymmetry, showing that the biological properties of chemical substances depend not only on the nature of the atoms constituting their molecules but also on the manner in which these atoms are arranged in space.

In 1848 Pasteur was appointed professor of physics at the Dijon Lycée (secondary school) but was shortly called to the University of Strasbourg as professor of chemistry. There, on May 29, 1849, he married the daughter of the rector of the university, Marie Laurent, by whom he was to have five children, only two of whom survived childhood.

In 1863 Pasteur became dean of the new science faculty at the University of Lille, where he initiated a highly modern educational concept: by instituting evening classes for the many young workmen of the industrial city, conducting his regular students around large factories in the area, and organizing supervised practical courses, he demonstrated the relationship that he believed should exist between theory and practice, between university and industry. At Lille, after receiving a query from an industrialist on the production of alcohol from grain and beet sugar, Pasteur began his studies on fermentation. In the course of his analysis he once again encountered—though in liquid form—new "right" and "left" compounds. From studying the fermentation of alcohol he went on to the problem of lactic fermentation, showing yeast to be an organism capable of reproducing itself, even in artificial media, without free oxygen—a concept that became known as the Pasteur effect.

Fermentation research

In 1857 he was named Director of Scientific Studies at the École Normale Supérieure. He continued his researches and announced that fermentation was the result of the activity of minute organisms and that when fermentation failed, either the necessary organism was absent or was unable to grow properly. Before this discovery, all explanations of fermentation had lacked experimental foundation; Pasteur showed that milk could be soured by injecting a number of organisms from buttermilk or beer but could be kept unchanged if such organisms were excluded.

He was elected to the Academy of Sciences in 1862, and the following year a chair at the École des Beaux-Arts was established for him for a new and original program of instruction in geology, physics, and chemistry applied to the fine arts.

As a scholar engaged in research, Pasteur eventually found his administrative duties as Director of Scientific Studies at the École Supérieure too irksome. He gave up the post in 1867 and, thanks to the support of Emperor Napoleon III, a laboratory of physiological chemistry was created for him at the same institution. As a logical sequel to his work on fermentation, he began research on spontaneous generation (the concept that bacterial life arose spontaneously), a question which at that time divided scientists into two opposing camps. Pasteur's recognition of the fact that both lactic and alcohol fermentations were hastened by exposure to air led him to wonder whether his invisible organisms were always present in the atmosphere or whether they were spontaneously generated. By means of simple and precise experiments, including the filtration of air and the exposure of unfermented liquids to the air of the high Alps, he proved that food decomposes when placed in contact with germs present in the air, which cause its putrefaction, and that it does not undergo transformation or putrefy in such a way as to spontaneously generate new organisms within itself.

After laying the theoretical groundwork, Pasteur proceeded to apply his findings to the study of vinegar and wine, two commodities of great importance in the economy of France; his pasteurization process, the destruction of harmful germs by heat, made it possible to produce, preserve, and transport these products without their undergoing deterioration.

Pasteurization

In 1865 he undertook a government mission to investigate the diseases of the silkworm, which were about to put an end to the production of silk at a time when it comprised a major section of France's economy. To carry out the investigation, he moved to the south of France, the centre of silkworm breeding. Three years later he announced that he had isolated the bacilli of two distinct diseases and had found methods of preventing contagion and of detecting diseased stock.

In 1870 Pasteur devoted himself to the problem of beer. Following an investigation conducted both in France and among the brewers in London, he devised, as he had done for vinegar and for wine, a procedure for manufacturing beer that would prevent its deterioration with time. British exporters, whose ships at the time had to sail entirely around the African continent, were thus able to send British beer as far as India without fear of its deteriorating.

Although Pasteur was partially paralyzed in 1868 and applied for retirement from the university, he continued his researches. In 1873 he was elected a member of the Academy of Medicine, and in 1874 the French Parliament provided him with an award that would ensure his material security while he pursued his work.

When, in 1881, he had perfected a technique for reducing the virulence of various disease-producing microorganisms, he succeeded in vaccinating a herd of sheep against the disease known as anthrax. Likewise, he was able to protect fowl from chicken cholera, for he had observed that once animals stricken with certain diseases had recovered, they were later immune to a fresh attack. Thus by isolating the germ of the disease and by cultivating an attenuated, or weakened, form of the germ and inoculating fowl with the culture, he could immunize the animals against the malady. In this he was following the example of the English physician Edward Jenner in his method for vaccinating animals against cowpox.

On April 27, 1882, Pasteur was elected a member of the Académie Française, at which point he undertook research that proved to be the most spectacular of all—the preventive treatment of rabies. After experimenting with inoculations of saliva from infected animals, he came to the conclusion that the virus was also present in the nerve centres, and he demonstrated that a portion of the medulla oblongata of a rabid dog when injected into the body of a healthy animal produced symptoms of rabies. By further work on the dried tissues of infected animals and the effect of time and temperature on these tissues, he was able to obtain a weakened form of the virus that could be used for inoculation. Having detected the rabies virus by its effects on the nervous system and attenuated its virulence, he applied his procedure to man; on July 6, 1885, he saved the life of a nine-year-old boy, Joseph Meister, who had been bitten by a rabid dog. The experiment was an outstanding success, opening the road to protection from a terrible disease. In 1888 the Pasteur Institute was inaugurated in Paris for the purpose of undertaking fundamental research, prevention, and treatment of rabies. Pasteur, although in failing health, headed the institute until his death on September 28, 1895.

Assessment

Louis Pasteur brought about a veritable revolution in the 19th-century scientific method. By abandoning his laboratory and by tackling the agent of the disease in its

natural environment, he was able through his investigations to supply the complete solution to a given question, not only identifying the agent responsible for a disease but also indicating the remedy.

A skillful experimenter, endowed with a great curiosity and a remarkable gift of observation, Pasteur devoted himself with immense enthusiasm to science and its applications to medicine, agriculture, and industry. He was prompt to defend his ideas with courage and often with considerable harshness—in writings as well as in speech—toward his opponents. It was chiefly in his work on spontaneous generation and on rabies that he encountered the strongest opposition to his ideas (which were, for the time, revolutionary) from medical circles and a section of the press. He was happy to accept the glory and honours that came his way, for he was well aware of his own value and of his scientific successes. A great friendship developed between Pasteur and the renowned British surgeon Sir Joseph Lister (1827–1912), who was quick to apply to his own discipline the discoveries of his eminent French colleague.

BIBLIOGRAPHY. E. DUCLAUX, *Pasteur: Histoire d'un esprit* (1896; Eng. trans., *Pasteur: The History of a Mind,* 1920), a scientific and philosophical work written by a collaborator of Pasteur; "Le laboratoire de Monsieur Pasteur," in *Centième Anniversaire* (1922); F. DAGOGNET, *Méthodes et doctrines dans l'oeuvre de Pasteur* (1967), primarily a detailed work of methodology; R.J. DUBOS, *Louis Pasteur: Freelance of Science* (1950), more philosophical than scientific; E. METCHNIKOFF, *Trois fondateurs de la médecine moderne: Pasteur, Lister, Koch* (1933; Eng. trans., 1939), an interesting work written by an important scholar who worked with Pasteur; JACQUES NICOLLE, *Louis Pasteur: A Master of Scientific Enquiry* and *Louis Pasteur: The Story of his Major Discoveries* (both 1961), two works giving a complete authoritative review of Pasteur's discoveries, and *Pasteur, sa vie, sa méthode, ses découvertes* (1969), the most complete work written to date about Pasteur's life and work; E. ROUX, "L'oeuvre médicale de Pasteur," in *Centième Anniversaire* (1922), a very interesting work written by one of Pasteur's collaborators; R. VALLERY-RADOT, *Histoire d'un savant par un ignorant* (1884), a very interesting work written by Pasteur's son-in-law, who was also his secretary, and *La vie de Pasteur* (1899; Eng. trans., 1960), a fundamental work on the life of Pasteur, but weak from the scientific point of view.

(J.M.R.N.)

Patagonian Desert

The Patagonian Desert covers the greater part of the region of Patagonia, that is to say, nearly all of the southern portion of mainland Argentina. With an area of about 260,000 square miles (673,000 square kilometres), it is the largest desert in the Americas, extending from latitudes 37° S to 51° S. It is bounded, approximately, by Andean Patagonia (the southern extension of the Andes mountains) to the west; by the Río Colorado to the north, except where the desert extends beyond the river into the borderlands of Mendoza and La Pampa provinces; by the Atlantic Ocean to the east; and by the Río Coig to the south. The desert thus covers most of the provinces of Neuquén, Río Negro, Chubut, and Santa Cruz; some geographers also include the area north of the island of Tierra del Fuego, which lies at the southern tip of South America and is divided between Argentina and Chile.

Origin of the name

The name Patagonia is derived from the Patagones, as the original inhabitants were called by Spanish explorers in the 16th century. It appears that Ferdinand Magellan, the Portuguese navigator and first European to reconnoitre the coast in that era, coined that name because the natives, with their thick furs, bushy hair, and painted faces, reminded him of Patagon, a dog-headed monster of the Spanish 16th-century romance *Amadís de Gaula.* The phrase *Patagonia, tierra maldita* ("Patagonia, land accursed") is proverbial. For an associated physical feature, see ANDES MOUNTAIN RANGES.

Landscape and environment. *Relief.* The desert covers the Patagonian tableland that extends from the Andes to the Atlantic Ocean. The general aspect of the tableland is one of vast steppelike (virtually treeless) plains, rising terrace fashion from high coastal cliffs to the foot of the Andes; but the true aspect of the plains is by no means as simple as such a general description would imply. Along the Río Negro, the land rises in a series of fairly level plains from about 300 feet (90 metres) at the coast to about 1,300 feet at the junction of the Limay and Neuquén rivers and 3,000 feet at the base of the Andes. The tableland region rises to an altitude of 5,000 feet. South of the Río Negro, the plains are much more irregular; volcanic eruptions have occurred in this area down to fairly recent times. Basaltic sheets, apparently only recently cooled, cover the tableland east of lakes Buenos Aires and Pueyrredón. On the Chico and Santa Cruz rivers, the plains have spread to within about 50 miles (80 kilometres) of the coast and reach almost to the coast south of the Coig and Gallegos rivers. In places, basaltic massifs (mountainous masses) are the salient features of the landscape.

The coast consists largely of high cliffs separated from the sea by a narrow coastal plain. Thus, the plateaus are formed of horizontal strata, some of sedimentary rocks, others of lava flows. Areas of hilly land, composed of resistant crystalline rocks, stand above the plateaus.

Rivers and lakes

Drainage. The deep, wide valleys bordered by high cliffs that cut the tablelands from west to east are all beds of former rivers that flowed from the Andes to the Atlantic, but only a few now carry permanent streams of Andean origin (the Colorado, Negro, Chubut, Senguerr, Chico, and Santa Cruz). The majority either have intermittent streams such as the Shehuen, Coig, and Gallegos, which have their sources east of the Andes, or, like the Deseado, are completely dry, except for salt ponds in the deeper depressions, and so altered by the combined effect of wind and sand as to afford little surface evidence of the rivers that once flowed in them. They serve an important purpose in the collection of the scanty surface water. Alluvial soils of considerable depth have been built up in them.

The line of contact between the Patagonian tableland and the folds of the Andes is marked by a chain of lakes, their upper ends deep in mountain canyons, their lower ends held in by glacial moraines.

From Lago Nahuel Huapí northward, the lakes—except for Lago Lácar—still drain to the Atlantic. South of Lago Nahuel Huapí, all of the lakes except Viedma and Argentino now drain to the Pacific through deep canyons cut across the Andean cordillera.

Climate. Patagonia is influenced by the South Pacific air current, which brings humid winds from the ocean to the continent. The winds, however, lose their humidity as they blow over the west coast of South America and over the Andes and are thus quite dry when they reach the Patagonian Desert.

The two climatic zones

Climatologically, the Patagonian Desert can be divided into two main zones, the northern and the southern, by a line drawn from the western half of Neuquén province to a point just south of the Península Valdés in northeastern Chubut province.

The northern zone, semi-arid, has annual mean temperatures between about 54° and 68° F (12° and 20° C); recorded maximum temperatures vary between about 106° and 113° F (41° and 45° C), and minimum temperatures vary between 12° and 23° F (−5° and −11° C). Sunshine, minimal along the coast, is most plentiful inland to the northwest. The rainfall varies from 3.5 to 17 inches a year. The prevailing winds, from the southwest, are dry, cold, and strong.

The climate of the southern zone of the Patagonian Desert is sharply distinct from the humid one of the Andean cordillera. In the north of this zone, the influence of the Atlantic is practically nonexistent—probably because of the heights of the coastal region (which reach 900 to 1,800 feet around the Golfo San Jorge)—although cold winds coming from the west and the cold Falkland Current both make themselves felt. In the southern part, which is practically peninsular, the Atlantic exerts some influence. The zone has a cold, dry climate, with temperatures that are higher along the coast than they are inland and with strong west winds. Mean annual temperatures range from 43° to 55° F (6° to 13° C), with maximum

temperatures at about 108° F (42° C), and minimum temperatures between 16° and −27° F (−9° and −33° C). There are heavy snows in winter and frosts throughout the year; and spring and autumn are so short that summer and winter are the only seasons worth counting. Average annual precipitation (rain and snow) ranges between about 4.5 and eight inches, though as much as 18.5 inches has been recorded. In the central areas of the desert there is less precipitation, and there is more sunshine than on the coast or in the Andean cordillera.

Vegetation and animal life. Whereas the long, narrow strip of Patagonia's western border has a vegetation like that of the adjacent cordillera, the arid region of the Patagonian Desert comprises two zones, each with its own characteristic vegetation.

The northern zone includes most of Neuquén province, most of Río Negro province, and the northeasternmost part of Chubut. About 84,000,000 acres (34,000,000 hectares) in area, it consists of open bushland, covered with widely spaced thickets between about three and seven feet high. Grasses flourish in the sandy areas, while salt-tolerant grasses and shrubs predominate in the salt flats. In irrigated areas profitable crops can be grown; these include peaches, plums, damsons, almonds, apples, pears, olives, grapes, hops, dates, vegetables, aromatic plants, and alfalfa. In the north, the zone merges into one of wooded steppe.

Crops of the irrigated areas

The second zone, covering the southwesternmost parts of Neuquén and Río Negro provinces, three-quarters of Chubut, and nearly all Santa Cruz province, has an area of 116,000,000 acres. The vegetation is low and very sparse and needs almost no water.

Among the birds of the desert are herons and other waders; predators such as the shielded eagle, the sparrow hawk, and the chimango (beetle-eater); and the almost extinct ñandú, or choique, a flightless bird similar to the African ostrich but not so tall.

The typical marsupial (animal having a pouch for carrying young) of the desert is the comadreja (a member of the weasel family). Species of bats include a long-eared variety. Armadillos, pichis (small armadillos), foxes, ferrets, skunks, mountain cats, and pumas are to be found, as also are maras, or Patagonian hares, and different kinds of burrowing rodents, such as the vizcacha and the tuco-tuco. Of the larger mammals, the most interesting is the guanaco, a kind of llama, hunted almost to extermination. There are a number of poisonous snakes as well as tortoises and lizards.

Vinchucas (winged bugs), bloodsucker insects (transmitters of American trypanosomiasis, or Chagas' disease), scorpions, and 14 kinds of spiders (including one kind called *Mecysmanchenius*, not found elsewhere) are also to be encountered.

The rivers and lakes are poor in fish, but sea fish and crustaceans and mollusks are plentiful off the coast.

The Patagones

Population. The original inhabitants of Patagonia seem to have been Indians from Tierra del Fuego. The most ancient artifacts, such as harpoons, found in the caves along the Strait of Magellan suggest that these people were moving up the mainland coast about 5,100 years ago. Robust and very tall, they constituted the two principal ethnic groups of Patagones—the Puelche-Guennakin in the north, and the Chonik or Tehuelche in the south. The surviving descendants of these Patagones, namely the Yámanes and the Alacalufes, are few in number; a kindred people, the Chono, died out in recent times. The Spanish explorers found the Patagones living as nomadic hunters of guanaco or of ñandú.

Toward the end of the 16th century the Spaniards attempted to colonize the Patagonian coastal region to clear it of English pirates; but a Jesuit settlement on the Golfo San Matías came to nothing. In 1778, however, the English tried to settle on the same bay, and the Spaniards reacted by founding Patagonia's first two towns, San José and Viedma (originally named Nuestra Señora del Carmen). A Spanish settlement at Puerto Deseado lasted from 1780 to 1807; but three years later this region was again devoid of European settlement.

After Argentina became independent, an attempt was made to populate Patagonia to make it part of the national state. Immigration, however, was not massive, though people came for various reasons: some to exploit the economic resources, others (for instance, the Welsh) to enjoy religious or political liberties. By the 1970s most of the immigrants were Chileans seeking temporary work rather than a fixed domicile. Such transitory immigrants form about 20 percent of the population. Apart from major concentrations at Comodoro Rivadavia and in the towns strung out along the upper valley of the Río Negro, Patagonia's sparse population is mostly rural.

Natural resources and their exploitation. Oilfields around Comodoro Rivadavia, Plaza Huincul, and Catriel are Patagonia's most conspicuous asset, apart from the great deposits of ore at Sierra Grande in Río Negro province, which supply all Argentina with iron. Río Negro province also has deposits of manganese, wolfram (or tungsten), fluorite (calcium fluoride, used as a flux in metallurgy), lead, and heavy spar (barite, the principal ore of barium); Neuquén province has deposits of copper and gold, vanadium (used to toughen steel), and zinc-lead; Chubut province has uranium and manganese in moderate quantities. There are also deposits of kaolin and gypsum.

Hydroelectric development

To exploit the hydroelectric potential of the Neuquén and Limay rivers, construction of a generating station to produce an estimated 1,650,000 kilowatts—the future centre of the El Chocón–Cerros Colorados complex—has been begun and is scheduled for completion by 1978. The project is also designed to irrigate the Comahue region, thus promoting farming and industry.

Transport. Argentina's National Route No. 3 runs for more than 1,860 miles from Buenos Aires and Bahía Blanca through the Patagonian coastal region southward to Comodoro Rivadavia. The roads of the desert proper, however, are few and of poor quality. Four railroads run east–west from the coastal region; two, which reach the foothills of the Andean cordillera, are connected with the Bahía Blanca–Buenos Aires line. Air services are based chiefly on the towns of the coastal region.

The chief ports are Rawson, Deseado, Santa Cruz, and Río Gallegos (south of the Río Coig); San Antonio Este, Puerto Madryn, Camarones, and San Julián, all on protected bays; and Comodoro Rivadavia, an outlet for petroleum products. San Antonio Este and Puerto Madryn are being developed for overseas traffic.

The economic potential

Prospects. The development of the Neuquén and Limay rivers has already been mentioned; the potential energy of other rivers could also be exploited. Better roads and a co-ordinated policy for the ports are also desirable. Fisheries could be organized on an economic basis. The wool trade, however, requires protection against unfavourable trends in the international market. The clay deposits of Neuquén province, on the Meseta de la Barda Negra, around Picún Leufú, and around Cerro Bandera, could provide the basis for a ceramics industry; an aluminum plant at Puerto Madryn could process imported ores and also Patagonia's own aluminum minerals. Patagonia's oilfields are capable of greater production.

It is also to be noted that the Patagonian Desert region, in general, and its arid subregion, in particular, are in the process of expanding. This is due to the destruction of the vegetational cover by animals pastured on the land, which in turn leads to erosion (gullying). Because of the lack of control over the land being used for pasture, the advance of this destruction is rapid.

BIBLIOGRAPHY. ANSELM WINDHAUSEN, "The Birth of Patagonia," *Rev. Soc. Arg. Geol.*, vol. 2 (1947), is a relatively nontechnical article on the geological history of Patagonia, with emphasis on the more prominent geological stratifications and tectonic events that led to the present configuration of the region. On the general structure, see A.D. YGOBONE, *Planificación general de la Patagonia* (1947).

(E.F.G.D.)

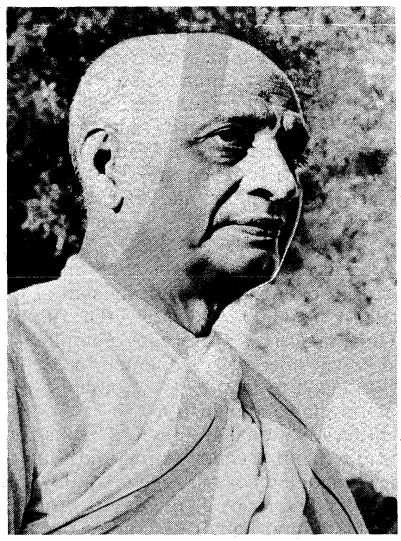
Patel, Vallabhbhai Jhaverbhai

Vallabhbhai Jhaverbhai Patel, barrister and statesman, stood in the forefront leadership of the Indian National Congress in the struggle for Indian independence; as

home minister and deputy prime minister of independent India, he smoothed the integration of the Indian states and helped mold the Indian Union.

Patel was born in Nadiād, near Surat, in Gujarāt, India, on October 31, 1875, into a self-sufficient landowning family of the Leva Patidar caste, renowned for their assertiveness and resourcefulness. Reared in an atmosphere of traditional Hinduism, he attended primary school at Karamasad and high school at Petlād, but was mainly self-taught. In contrast to many of his fellow students, Patel was of strong constitution; he enjoyed athletics and expressed a self-confident and stubborn leadership.

By courtesy of the Information Service of India, London



Patel.

Legal career

He married at the age of 16, matriculated at 22, and passed the district pleader's examination, which enabled him to practice law. In 1900 he set up an independent office of district pleader in Godhra, and two years later he moved to Borsad, in Kheda District.

As a lawyer, Patel distinguished himself in presenting an unassailable case in a precise manner, and in challenging police witnesses and British judges. He was immune to flattery and aloof from servile Indian lawyers. In 1908 Patel lost his wife, who had borne him a son and daughter, and thereafter remained a widower.

Determined to enhance his career in the legal profession, Patel travelled to London in August 1910 to study at the Middle Temple. There he studied diligently and passed the final examinations with high honours.

Returning to India in February 1913, he settled in Ahmadābād, rising rapidly to become the leading barrister in criminal law at the Ahmadābād bar. Reserved and courteous, he was noted for his superior mannerisms, his smart, English-style clothes, and his championship in bridge at Ahmadābād's fashionable Gujarāt Club. He was, until 1917, indifferent to Indian political activities.

In 1917 Patel, like many other Indians at the time, found the course of his life changed after having been influenced by Mahatma Gandhi. Patel adhered to Gandhi's *satyāgraha* (policy of nonviolence) insofar as it furthered the Indian struggle against the British. But he did not identify himself with Gandhi's moral convictions and ideals, and he regarded Gandhi's emphasis on their universal application as irrelevant to India's immediate political, economic, and social problems. Nevertheless, having resolved to follow and support Gandhi, Patel changed his style and appearance. He quit the Gujarāt Club, dressed in the white cloth of the Indian peasant, and ate in the Indian manner.

From 1917 to 1924 Patel served as the first Indian municipal commissioner of Ahmadābād and was its elected municipal president from 1924 to 1928. Patel first made his mark in 1918, when he planned mass campaigns of peasants, farmers, and landowners of Kaira District, Gujarāt, against the decision of the Bombay government to collect the full annual revenue taxes despite crop failures caused by heavy rains.

In 1928 Patel successfully led the landowners of Bārdoli in their resistance against increased taxes. His efficient leadership of the Bārdoli campaign earned him the title *sardār* (leader), and henceforth he was acknowledged as a nationalist leader throughout India. He was considered practical, decisive, and even ruthless; and the British recognized him as a dangerous enemy.

Political philosophy

Patel, however, was no revolutionary. In the crucial debate over the objectives of the Indian National Congress during the years 1928 to 1931, Patel believed (like Gandhi and Motilal Nehru, but unlike Jawaharlal Nehru and Subhas Chandra Bose) that the goal of the Indian National Congress should be dominion status within the British Commonwealth—not independence. In contrast to Jawaharlal Nehru, who condoned violence in the struggle for independence, Patel ruled out armed revolution, not on moral but on practical grounds. Patel held that it would be abortive and would entail severe repression. Patel, like Gandhi, saw advantages in the future participation of a free India in a British Commonwealth, provided that India was admitted as an equal member. He emphasized the need to foster Indian self-reliance and self-confidence, but, unlike Gandhi, he did not regard Hindu–Muslim unity as a prerequisite for independence.

Patel disagreed with Jawaharlal Nehru on the need to bring about economic and social changes by coercion. A conservative rooted in traditional Hindu values, Patel belittled the usefulness of adapting Socialist ideas to the Indian social and economic structure. He believed in free enterprise, thus gaining the trust of conservative elements and thereby collected the funds that sustained the activities of the Indian National Congress.

Patel was the second candidate after Gandhi to the presidency of the 1929 Lahore session of the Indian National Congress. Gandhi shunned the presidency in an attempt to prevent the adoption of the resolution of independence and exerted pressure on Patel to withdraw, mainly due to Patel's uncompromising attitude towards the Muslims; Jawaharlal Nehru was elected. During the 1930 Salt Satyāgraha, Patel served three months' imprisonment. In March 1931 Patel presided over the Karāchi session of the Indian National Congress. He was imprisoned in January 1932. Released in July 1934, he marshalled the organization of the Congress Party in the 1937 elections and was the main contender for the 1937–38 Congress presidency. Again, because of Gandhi's pressure, Patel withdrew and Jawaharlal Nehru was elected. Along with other Congress leaders, Patel was imprisoned in October 1940, released in August 1941, and imprisoned once more from August 1942 until June 1945.

During the war Patel rejected as impractical Gandhi's nonviolence in the face of the then expected Japanese invasion of India. On the transfer of power, Patel differed with Gandhi in realizing that the partition of the subcontinent into Hindu India and Muslim Pakistan was inevitable, and he asserted that it was in India's interests to part with Pakistan.

Patel was the leading candidate for the 1945–46 presidency of the Indian National Congress, but Gandhi once again intervened for the election of Nehru. Nehru, as president of the Congress, was invited by the British viceroy to form an interim government. Thus, in the normal course of events, Patel would have been the first prime minister of India. During the first three years of independence, Patel was deputy prime minister, minister of home affairs, minister of information, and minister of states; above all his enduring fame rests on his achievement of the peaceful integration of the princely Indian states into the Indian Union and the political unification of India. He died in Bombay on December 15, 1950.

BIBLIOGRAPHY. D.V. TAHMANKAR, *Sardar Patel* (1970), provides the most comprehensive biography of Patel to date. MICHAEL BRECHER, *Nehru: A Political Biography* (1959), includes the best analysis of the confrontation between Nehru and Patel. V.P. MENON, *The Story of the Integration of the Indian States* (1956), highlights the role of Patel.

(D.Ar.)

Patent Law

The word patent is derived from "letters patent," which earlier designated the document by which a sovereign conferred a privilege or right on someone. The name was a reference to the fact that the document, addressed to the public at large, was sealed with the great seal in such a way that the document could be unfolded and read without breaking the seal ("patent," from the Latin meaning "opened" or "exposed"). Modern usage of the term is confined primarily to grants of certain rights in inventions, and the solemn formalities of a royal grant generally no longer exist.

The patent of invention is a grant of specified rights by the government of a particular country. These rights generally consist in the exclusive right to make (that is, manufacture) and to deal with the subject matter of the patent for a limited period of time. A distinction must be drawn between the grant of the positive right to make and the grant of the right to exclude others from making. In older times probably both were usually intended, since restrictions existed on the freedom of anyone to engage in particular occupations. Modern development, however, has been toward the concept that only the right to exclude others is granted by the patent, since the patentee's own right to make and deal with the subject matter exists without any grant, subject to any superior rights of others or any applicable statutes.

The "exclusive" rights of patents

The right to exclude, conferred by the patent, is expressed in various ways and may differ somewhat in scope in the statutes of different countries. In the United States, for example, it is specifically expressed as "the right to exclude others from making, using or selling the invention." In the United Kingdom the granting document gives the patentee the sole right to "make, use, exercise and vend" the invention. Some statutes—*e.g.*, that of France in 1968—list specific acts, such as manufacturing, making use of, introducing into the country, selling, offering for sale, putting on the market, etc. Hence, the person granted the patent has a monopoly over the specific subject matter of the grant for a period of time.

The system of granting patents for inventions has developed with a complex of objectives including rewarding the inventor and thereby stimulating inventive activity; encouraging the disclosure of the subject matter so that the public will be in possession of the knowledge; and encouragement to produce the item. By having the exclusive right, the patentee is encouraged to manufacture and put the invention into use since he is free of competition for a time; and investment in such enterprises is also stimulated. And, incidentally, some competition in making inventions may be introduced.

HISTORY OF PATENTS

Italian origin of patents

Patents for inventions were first introduced in the 15th century in certain Italian states—the first known grant by a state to an inventor having occurred in the Republic of Florence in 1421 and an ordinance relating to patents having been enacted in Venice in 1474. The concept spread to other European states, and during the following two centuries the grant of invention patents became more and more common and regular. At first these grants were ordinarily made without any specific statutes on the subject but by virtue of the general authority and power of the ruler. In England, for example, Queen Elizabeth I issued grants relating to inventions, grants of exclusive rights and privileges regarding the importation and establishment of industries new to the realm, and monopoly grants relating to known commodities. Dissatisfactions over the monopoly grants led to the Statute of Monopolies of 1623, an act of Parliament that declared such grants to be unlawful but, as an exception, confirmed the authority to grant exclusive rights for new inventions for a term of 14 years. Thereafter, invention patents were granted fairly continuously in England (more so than in other European countries) under a general system developed by custom, rather than by statute.

First comprehensive patent statutes

Formal, comprehensive patent statutes did not appear until near the close of the 18th century—in the United States in 1790 and in Revolutionary France in 1791. The former was an expression of democratic action by the new republic and the latter a change from what had been a royal prerogative to one of the rights of man. The French statute of 1791 was explicit in declaring the natural right of an inventor to the exclusive right to his invention. In the United States it was necessary, when the individual states formed the union, to confer the authority to grant patents specifically upon the central government. The Constitution of 1787 gave to the Congress the power "to promote the progress of . . . the useful arts, by securing for limited times to . . . inventors the exclusive right to their . . . discoveries." In accordance with this provision, the first federal patent law was enacted. In the 19th century the enactment of specific statutes continued elsewhere. Early ones include those in the Netherlands (1809), Austria (1810), Russia (1812), Bavaria (1812), Prussia (1815), Sweden (1826), Spain (1826), two Canadian provinces (1823, 1826), Mexico (1832), Brazil (1840), Chile (1840), and even the Republic of Texas (1839) before it joined the United States. There was no general statute in Great Britain until 1852. Eventually, most of the individual German states enacted provisions for the granting of patents, and, after unification, the new Germany passed a general patent law (effective in 1877). The Sardinian patent law of 1859 was extended to the rest of Italy in stages as the country became unified from 1860 to 1870.

By the end of the 19th century a large number of countries had enacted patent laws and some had even revised earlier laws to improve them. At present there are about 100 independent jurisdictions having separate patent statutes. In addition, about 20 countries rely solely on the registration therein of a patent obtained in a specified other country; for most, this is the United Kingdom, of which they were formerly colonies, and about two dozen British colonies have patent ordinances that provide for registering British patents therein and, in some instances, for independent patents also. Only about a half-dozen countries have no patent law.

THE GRANTING OF PATENTS

Types of patents

The most common type of patent is one granted to an inventor for an invention or one granted to another person (including a "legal person" or corporation) who derives rights from the inventor. A "patent of addition" refers to an improvement on a patented invention; it is issued to the same patentee and expires with the main patent. So-called secret patents may be issued for inventions that are to be kept secret for purposes of national security, but usually in such cases the patent is withheld; the invention is simply kept secret. A small number of countries, including Germany, Italy, Japan, Spain, Poland, and the Philippines, provide for a kind of second-class patent for inventions of useful articles, referred to as a *Gebrauchsmuster* in Germany or, generally, as a utility-model patent. Its chief distinction is that its term is usually much shorter and processes and compositions may be excluded. Some countries provide for what are called confirmation, revalidation, or importation patents, whereby a person who has obtained a patent in a foreign country may more easily obtain a local patent. The chief advantage is that all the procedural efforts toward issuing the patent having already occurred, the second country can merely enter a record of the first patent. A number of former British colonies provide only for the recording of a British patent. It is possible in a few countries for one who is not the inventor or who did not derive the rights from the inventor to obtain a patent by importing the invention into the country; this is the case in the United Kingdom.

Inventors' certificates in Socialist countries

A different system exists in the U.S.S.R. and in several other Socialist countries (Albania, Bulgaria, China, Romania), which provides not only patents of the usual kind but also what are referred to as authors' or inventors' certificates. The author's certificate does not confer any exclusive rights to the invention but records the inventor's contribution, recognizes his authorship, and entitles him to receive some compensation for the use of the invention (if it is used) and also to receive some per-

quisites. Nationals of these countries obtain very few or no patents; and the right conferred by a patent, though it may be expressed in terms of exclusive rights, does not really have the same effect as in other countries because of the absence of private enterprise.

Subject matter for patents. The subject matter for which patents can be obtained, as has been stated, relates to inventions. In general, this includes practically every type of invention of an industrial character. Different statutes express the subject matter in various ways—in the United States as "any new and useful process, machine, manufacture, composition of matter or improvement thereof," and in the United Kingdom as "any manner of new manufacture" (with certain elaborations).

In a few countries, such as the United States, patents may also be granted for new and distinct varieties of plants, which the inventor, or discoverer, himself has asexually reproduced. Various items, such as scandalous or immoral matter, matter contrary to law, scientific principles, and so on, commonly cannot be patented; indeed, processes involving only mental steps are not ordinarily considered proper subject matter for patent at all. Some countries exclude the patenting of substances produced by chemical action and of medicines, although the process for making such substances can be patented and the patent would cover the product when made by the process. This type of exclusion was removed recently in Germany and in the Scandinavian countries, and its removal is finding greater support elsewhere. It does not appear in the laws of France, the United States, the United Kingdom, and many other countries.

Principle of novelty

To qualify for a patent, an invention must be new, and novelty is ordinarily defined by the statute. In most countries, if the invention has become known to the public (*e.g.*, by having been described in a printed publication, by having been put into public use or otherwise having become known to the public) before the application for patent is made, it is no longer considered novel and a valid patent cannot be obtained. In a few countries the inventor's own publication of his invention for a limited period of six months or a year is not considered derogatory to his rights. In only a few countries, too, is there a period within which any publicity relating to the invention will not defeat the right of an inventor to a patent. Usually, any prior use outside the country concerned, without printed publicity, is not considered as defeating novelty. In the United Kingdom all relevant events occurring outside the kingdom are disregarded.

Principle of inventive merit

In addition to the requirement for novelty, there has developed, primarily in the more industrialized countries, a further requirement of a somewhat subtle character concerning the "inventiveness," "unobviousness," or, in Germany, "inventive level" or "inventive height" in comparison with the "state of the art" or "prior art." In other words, the new subject matter must sufficiently advance the state of the art for a patent to be warranted.

The form of language proposed by the Council of Europe and coming into use in European countries is that an invention, to be patentable, must be new and "involve an inventive step," with the further definition that an inventive step is involved if the subject matter is "not obvious having regard to the state of the art." The intention of these provisions, of course, is to exclude slight or trivial changes and obvious changes from the protection of patents; however, the determination of adequate inventiveness has been one of the most difficult aspects of patent law and has involved the greatest number of conflicts of opinion.

Application for patents. In most countries the inventor or person (including companies) deriving the right from the inventor may apply for a patent; in the United States usually only the inventor may apply, although the patent may be granted to the assignee. The application for a patent must comply with the formal requirements of the country, including specified forms, documents, procedures, and fees. The application must contain a description of the invention so presented that it can be accomplished through its directions; and it must conclude with one or more statements referred to as "claims," which define the invention succinctly and serve to delimit the scope of the rights requested.

Formal requirements and filing dates

An application for patent must be directed to one invention, but several related inventions, linked by a single inventive concept or otherwise closely related and interdependent, may be permitted in the same application. Practices vary on this point. If the application violates the rules or practices relating to unity of invention, the applicant will be required to "divide" or "restrict" the application to but one invention.

The filing date of the application is important since, in general, it marks the point in time at which novelty of the invention must exist and also determines who obtains the valid patent in cases in which two or more different persons make application for the same invention. Although the filing date is controlling in most countries, a few countries, such as the United States, Canada, and the Philippines, permit recourse to the date of invention.

In general, applications for patent are maintained in confidence by the patent office until after the patent is granted, or at a certain stage in the proceedings, or after a certain time interval from the filing date of the application, as may be specified in the statute.

The practice of opening applications to public inspection at a certain time after the filing date, 18 months, was first followed in Australia and has spread to such countries as The Netherlands (1964), Denmark, Finland, Norway, and Sweden, Germany (1968), France (1969), and Japan (1970). In The Netherlands, Germany, and Japan, this practice was adopted as part of the deferred examination system referred to below. In these countries, as well as in France, the specification is issued in printed form at or about the same time; in the other countries, lists may be published, with the public then being able to inspect or obtain copies. The practice was adopted in the Scandinavian countries in view of the long period of pendency of applications before the patent is issued in order that domestic industry would not be taken by surprise. The applicant has provisional patent rights on the publication of his application, which become ineffective if a patent is not ultimately obtained.

Administrative procedure. Broadly, two main types of administrative procedures govern the granting of patents. Under the first system, called the registration system, if the application papers are deemed to be formally in order, the patent is granted in due course. Whether the patent complies with substantive provisions of the law is determined later by a court if the question arises, as in a suit on the patent. This system is followed in Belgium, Italy, Spain, Switzerland (except for certain classes of inventions), and a large number of other countries. In some countries there may be a refusal of a patent for reasons apparent on the face of the papers (that the invention is scandalous, for example) or an objection based on a charge of lack of unity. The French law of 1844 was the prototype of this system.

The registration and the examination systems

The United States adopted the French system in 1793, but a new patent law in 1836 introduced what is usually referred to as the examination, or pre-examination, system. This was adopted by Germany in 1877 and since then by many other countries. Under this system, an attempt is made in advance to determine whether or not the application complies with the substantive requirements for patents. An examiner searches through relevant prior patents, possibly including those of other countries, and publications available to him to determine whether or not the invention being claimed is, in fact, new and also whether or not the degree of novelty is sufficient to support a patent. The U.S., Germany, The Netherlands, Sweden, Austria, Japan, the U.S.S.R., and a few other countries make the fullest pre-examination in this respect. But even in these countries the examination cannot be exhaustive, and matters such as the existence of prior public use are not investigated before grant.

Typically, under the examination system, an examiner, after making a search of the prior art, writes to the applicant communicating the results and indicating whatever action is considered appropriate at that time. All or some claims may be indicated as not patentable (the precise

terminology and procedure differ according to different national laws), which means that a patent cannot be granted on the stated claims; and specific formal objections can be raised. If the decision is adverse in any respect, the applicant may present arguments and amendments (though in general new matter cannot be introduced) seeking to overcome the examiner's position. The application is then re-examined. There may be several of these exchanges of actions and responses. If the examiner concludes in the re-examination that all or some of the claims cannot be allowed, the applicant can appeal to a higher body—such as the patent office's board of appeals and then the courts (as in the United States) or to a federal patent court (as in West Germany).

The French system

Under a new French law that came into effect in 1969, a search of the prior art is made for the French office by the International Patent Institute (referred to below) and a report on the results made. The applicant may amend and comment on the report, and members of the public can also file statements, after which the patent office draws up a final report and issues the patent. There is no refusal of a patent even if the report is adverse, in strict conformity with traditional French principles, but the patent with its record is left to be adjudicated by the courts. The 1969 law also introduced a second type of patent, called a "utility certificate," which an applicant obtains at his own option; this patent does not involve a search and report, and its term is six years instead of 20.

Most countries employing the examination system carry out merely a limited examination, searching primarily only through their own prior patents; few countries have arrangements for obtaining the results of searches made in other countries. The chief reason for this is the formidable expenditure of time and effort involved in such full searches. A complex of documentation must be built up and maintained, and a staff of persons skilled in science and technology is needed to conduct the examinations. In the United Kingdom only a limited search and a limited examination are made through initial reluctance to give public officials too much authority.

In most examining countries (the United States and Canada excluded), when the examiner concludes that a patent may be granted, the application is published; and the public, during a period of generally three or four months, is given an opportunity to oppose the granting of the patent. In West Germany and some other countries, any person, regardless of his interest, may oppose; whereas in the United Kingdom the opposer must have an interest for so doing. But, in any event, the opposer can object on virtually any grounds and produce evidence additional to what the examiner had considered or could consider.

Systems of deferred examination

Because of the heavy workload and increasing backlog of applications, various schemes have been developed to alleviate or expedite the process of examination. In The Netherlands, a procedure introduced in 1964 represents an attempt to reserve examinations only for applications that prove to have enduring importance or interest. Under this system, an application is published 18 months after the filing date, and the applicant receives certain provisional rights. To maintain these rights he must pay an annual fee. At any time during a period of seven years after the filing date, the applicant may request initiation of the examination procedure, paying a special examination fee. If the seven-year period elapses, however, without any such request, the application is considered no longer pending, and the provisional rights disappear retroactively. In this manner, only those applications in which the applicant demonstrates a persistent interest are examined, and considerable work on the part of the patent office is saved. The deferred examination system was adopted in Germany in 1968 and in Japan in 1970. Other efforts to relieve procedural pressures, particularly with respect to the time required for searching, continue; the possibility of mechanical or computer searching, under study for some time, has as yet yielded inadequate results.

Publication of specifications. About two dozen countries that grant patents publish the specification (and drawings) in printed form—either some time during or before examination or at the time the patent is granted or shortly thereafter. These publications constitute a large body of published technology. Patent offices in some countries have put considerable effort into developing and maintaining classifications of this material so that it may be consulted not only by their official examiners but also by the public generally. Many countries exchange copies of their printed specifications, which enables them conveniently to build up their files. Most countries publish an official journal in which notices and other information relating to patents and allied subjects appear.

Term and maintenance fees. A patent is granted for a limited period of time that varies among different countries. The most common practice is for the patent to expire a certain number of years after the filing date—16 years in the United Kingdom, Australia, Ireland, New Zealand, and South Africa; 17 years in Denmark, Finland, Norway, and Sweden; 18 years in Germany and Switzerland; and 20 years in Belgium, France, Hungary, Israel, and a number of other countries. In some countries, however, the term is a specified number of years from the granting date (in the United States and Canada, 17 years; in Spain, 20 years; and in some other countries, 14 or 15 years) or from the date of publication of the application (in Austria, 18 years).

Renewal fees

A fee is required when an application is filed and, in general, at some later stage also. The large majority of countries (the United States and Canada excepted) require the payment of periodical fees to maintain a patent in force. These fees may become due beginning one, two, three, or four years from the filing date and annually thereafter. One result of the renewal-fee system is that patents in which the owner is no longer interested drop out of consideration as subsisting monopolies. If a renewal fee is not paid, the patent lapses, but usually a period of grace for late payment of a fee is provided; and in some countries (the United Kingdom for one) restoration of a lapsed patent is possible.

PATENTS AS PROPERTY

Patents constitute objects of intangible personal property, and this is expressly stated in some statutes. A patent can be transferred to another owner by assignment or passed to another by inheritance or bequest.

Licensing

The patentee may grant licenses to others to use his invention or discovery; this license, being a contract, is governed by the laws relating to contracts (see CONTRACTS, LAW OF). Customarily, the license is based upon royalty payments, though there may be some other consideration. Terms and conditions of various kinds can be included, but here violation of other laws, such as antitrust laws, must be avoided. A "tying clause," whereby the licensee is required to purchase materials and supplies from the licensor (patentee), may be illegal.

Patent infringement. Infringement of a patent consists in the unauthorized performance of any of the acts reserved for the patentee—such as manufacturing, using, and selling and, in some countries, other activities such as importing or stocking for trade. The laws of some countries—Germany, for example—specifically state that the activities must be for industrial or commercial purposes and that personal use of a patented invention for noncommercial purposes would not constitute infringement. When his exclusive rights have been infringed, the owner of the patent may file suit in the appropriate court and recover damages as well as obtain an injunction prohibiting future infringements. In some countries a criminal action may also be taken against the infringer, although this is not often done. The government may itself exercise the right to use the patented invention but with just compensation to the patentee; the latter may usually file suit in an attempt to adjust a compensation that he considers unsatisfactory.

The court, in a suit for infringement, may not only adjudicate the infringement question but also consider the validity of the patent itself. A few countries (West Germany, for example) separate the two questions: one court handles the infringement suit; the federal patent court considers the question of validity—in a separate suit to

revoke or nullify the patent. Specific actions in court to revoke a patent are possible in many countries. In the United Kingdom an interested party can seek revocation in the patent office itself during the first year after the patent is granted.

Working and compulsory licenses. In the 19th century, in most countries, the patentee was obliged to "work" his invention fairly continuously within a specified period, such as two or three years; if he did not, the patent became void or was revoked (provided the patentee offered no adequate excuse). "Working" meant either some measure of actual manufacture within the country or some public offer to grant licenses. Later, however, there developed a compromise procedure: the patent is not revoked for nonworking; it simply, after a specified period, becomes subject to the issuing of what are called *compulsory* licenses. If the invention has not been worked, then a qualified applicant may seek the award of a license from the patent office or a court. He must pay the patentee whatever royalties may be determined, and he cannot transfer his license to another. After two years, if the compulsory licenses themselves do not prove adequate to encourage production, the patent can finally be revoked for nonworking.

Compulsory licenses

Compulsory licenses serve a number of purposes. They may be intended to advance the public good, by making certain that an invention that is deemed useful is indeed used. Several countries—France, the United Kingdom, Canada, and the Philippines, for instance—specifically provide that medicines or inventions useful in preparing medicines be subject to compulsory licenses. The United States has such a provision for patents relating to atomic energy (though atomic weapons are not patentable at all). They may be provided for in cases of abuse of patent rights. Although the United States has no general law relating to working or compulsory licenses, the courts have used the device of compulsory licenses in deciding antitrust suits; a corporation judged monopolistic can be forced, as part of the remedy, to grant licenses to others. Finally, compulsory licenses are often issued to those who have dependent patents—patents, that is, that cannot be worked without using someone else's patent. A compulsory license is granted for the main patent so that the dependent one can be properly worked. Although it can be said that the frequency of recourse to the provisions of compulsory license laws is not very great, such laws may serve as safeguards.

In the United Kingdom, West Germany, and a few other countries, the patentee can file a declaration stating that he will grant licenses to anyone, whereupon his annual fees are reduced by half. Some countries contain a provision for the expropriation of a patent when required in the public interest, with the payment of compensation to the patentee. Such provisions are seldom, if ever, invoked, however.

INTERNATIONAL DEVELOPMENTS IN PATENT LAW

International Convention for the Protection of Industrial Property. Although there are a number of bilateral treaties on patent matters and a few treaties involving several countries, the most important general patent treaty is the International Convention for the Protection of Industrial Property, first signed in Paris in 1883 and coming into effect on July 7, 1884, between 11 countries. This treaty was revised at Brussels in 1900, at Washington in 1911, at The Hague in 1925, at London in 1934, and at Lisbon in 1958, each revision superseding the former version. A further revision at Stockholm in 1967 related mainly to organizational matters. By 1970 the number of participating countries had reached 78. The large increase in recent decades was due mainly to the adherence of newly established countries, but the normal increase continued as well, the U.S.S.R. adhering in 1965 and Argentina in 1967. Under the first article of the convention the countries constituted themselves a Union for the Protection of Industrial Property.

The convention establishes rights in industrial property —namely patents, designs, trademarks, unfair competition, and related matters—of the nationals (and residents) of the participating countries and in some respects has the aspect of international legislation on these subjects. One of the basic articles of the convention provides that the nationals (and residents) of each country enjoy in the other countries the same advantages that the other countries grant to their own nationals, without prejudice to any rights specially provided for by the convention.

Major provisions of the International Convention

One of the most important articles of the convention deals with what is called the "right of priority." By the operation of this provision, a person who has duly applied for a patent in one of the countries of the union and then, later, within 12 months, applies for a patent for the same invention in another of the countries can be considered to have applied in the second country on the earlier date. This article is important to inventors in view of the fact that in most countries publication of the invention before applying for a patent will defeat the right to a patent.

Other articles of the convention provide that patents obtained in the different countries of the union shall be independent of one another; that the inventor shall have the right to be named in the patent (this being important in those countries in which the owner, as distinguished from the inventor, may apply for the patent); that a period of grace shall be provided for the payment of the fees for maintaining a patent in force; that the use of an invention on board ships or on land vehicles or aircraft temporarily or accidentally in the territory of a country shall not be considered as infringing the rights of a patentee in that country, and so on. One article prescribes certain criteria that must be observed if a patent is to be revoked or if compulsory licenses are to be granted. The convention also includes provisions dealing with trademarks, trade names, and unfair competition.

Under the convention, a central office known as the International Bureau for the Protection of Industrial Property, located in Geneva and under the general supervision of the Swiss government, serves as a clearinghouse of information on patent and trademark laws and issues a monthly periodical in French and English, *La Propriété industrielle*, or *Industrial Property*, dealing with patents, trademarks, and related matters. It issues other publications, arranges and serves as the secretariat for international conferences and meetings, and engages in other activities relating to patents and trademarks. A similar bureau was established under the Berne Copyright Convention of 1886, which is administered jointly with the patent bureau—the combined administration being called the United International Bureaux for the Protection of Intellectual Property (BIRPI). The Stockholm Conference of 1967 also produced a new treaty establishing a World Intellectual Property Organization (WIPO), which entered into force in 1970. This new organization is to replace the earlier administrative framework and is somewhat wider in scope.

International Patent Institute

Another intergovernmental organization is the International Patent Institute (Institut International des Brevets) at The Hague, established in 1949 by agreement between Belgium, France, Luxembourg, and The Netherlands, with the object of maintaining a central place that would preserve documents and perform the function of making a search of the prior art for member countries. Since 1949, Monaco, Switzerland, Turkey, and the United Kingdom have also joined. At present the institute performs searching work for the Swiss, French, and Dutch patent offices. It also makes searches for private parties, for a fee.

Numbers of patents. The ten countries that in recent years have issued the greatest number of patents are, in alphabetical order, Belgium, Canada, France, Federal Republic of Germany, Italy, Japan, Switzerland, the Soviet Union (inventor's certificates), the United Kingdom, and the United States. The numbers for these countries range from 65,000 to 15,000 per year, though the numbers are not strictly comparable owing to the different systems. Countries issuing from 5,000 to 15,000 patents per year in recent years include Argentina, Australia, Austria, India, Mexico, South Africa, Spain, and

Sweden. Countries issuing from 1,000 to 5,000 include Brazil, Chile, Denmark, Finland, Greece, Hungary, Israel, Luxembourg, The Netherlands, New Zealand, Norway, Pakistan, Poland, Portugal, and Romania.

During the decade before World War II, the total number of patents granted by all countries averaged approximately 180,000 a year. This number decreased considerably during the war and the years immediately following, sinking to less than half; but by 1950 the number had risen again to above the prewar average. The increase continued sharply until by 1970 over 400,000 patents were being issued annually. The increase is attributable partly to a gradual increase in the total number of inventions being patented; but more largely it is due to the expansion in international patenting. Of the more than 400,000 patents issued in 1969, 40 percent were granted to residents of the countries issuing them and 60 percent to residents of other countries. Thus a large proportion of the 400,000 patents were simply duplications of the same essential invention patented in different countries. Whereas on the national level the increase in patent applications has led to attempts to ease the increasing burdens by changes in systems, on the international level there has been an attempt among countries to increase cooperation and develop various plans for curtailing the considerable duplication of effort.

International duplication of patents

International patents. The rights granted by a patent extend only within the territorial jurisdiction of the state granting the patent, and thus one who desires patent rights in a number of different countries must obtain a separate patent in each of them. The burden of obtaining international protection can become quite heavy.

Although the laws of some countries provide that under certain conditions a patent granted in another country may be registered and become effective without the usual patenting procedures, there is no international patent law—with one exception. In 1964, 13 French African countries concluded a treaty adopting a common patent law, with a single patent office issuing patents, each patent effective in all countries of the group. The countries—Cameroon, Chad, Central African Republic, Congo (Brazzaville), Dahomey, Gabon, Ivory Coast, Malagasy Republic, Mauritania, Niger, Senegal, Togo, and Upper Volta—maintain their common patent office at Yaoundé, Cameroon, and have a patent law that is nearly the same as the French law before the recent changes. The countries also provide for a common trademark registration.

The Council of Europe early considered the question of a common patent but decided first to take up problems involving the lack of uniformity in national laws. The first result, in 1955, was a treaty governing the formalities required in filing applications; it has since been adhered to by 18 countries, two outside Europe. Next, again in 1955, was a treaty providing for an international classification of patents (now adhered to by 15 countries and also adopted by others). Third was a Convention on the Unification of Certain Points on Substantive Law on Patents for Inventions, signed in 1963 by 11 countries, which envisioned uniform provisions relating to the subject matter for patents and to such substantive conditions of patentability as novelty and inventiveness. Although this treaty never went into effect, it influenced several countries to enact laws following its provisions.

After World War II, the Scandinavian countries, Denmark, Finland, Norway, and Sweden, commenced efforts toward greater cooperation in patent matters, as in other affairs. The first part of the plan involved enactment of uniform patent laws, accomplished finally in 1967; these are still separate national laws, however. Under the second part of the plan, not yet achieved, there will be an international law and treaty, and it will be possible to obtain a single patent valid in all of the member countries.

Although the treaty establishing the Common Market, or European Economic Community, recognized the existence of separate national patent laws, there have also been efforts to internationalize patent rights in these countries. Early efforts failed by reason of various disagreements, but negotiations begun in 1968 among a larger number of European countries led to the publication in 1971 of the second preliminary draft of a treaty establishing a European patent. The general plan is for the creation of a European patent office in which applications for patents, filed by anyone, can be examined in accordance with the patent law and procedure embodied in the treaty. The single patent, when issued, will operate as a national patent in each of the participating countries, subject to respective national laws but bound by the law of the treaty with regard to certain basic requirements. The single patent will represent a bundle of separate patents, issued, however, as a single act by one authority and bound by common conditions for patentability. Each country may still issue national patents to persons who apply directly in that country. A second treaty between the Common Market countries alone is part of the plan; as to these countries the European patent will operate as a single patent for a combined territory.

Patent Cooperation Treaty of 1970

A new treaty called the Patent Cooperation Treaty was drawn up and signed in Washington in June 1970 by 20 adherents of the International Convention. Its main objective is to facilitate the filing of patent applications in a number of countries. A single application can be filed in a single office (branch offices existing in each nation), and the application can be effective in any of the participating countries that the applicant designates. The application will be subject to an initial examination as to form and will then receive a search of the prior art, with a report on the results. After this stage, the individual countries take over, having been supplied official copies; and they proceed with their own further steps toward granting the patent. A separate chapter of the treaty provides for examination by a single examining office, but this provision is optional and pertains only to those countries that may elect to have such a service. One chapter relates to supplying information services and technical services in patent matters. The treaty will come into force when ratified by eight countries, at least four of which meet certain conditions as to the volume of patent activity.

BIBLIOGRAPHY. There is no recent general legal treatise on patent law on an international and comparative level. The *Manual for the Handling of Applications for Patents, Designs, and Trademarks Throughout the World*, 2 vol. (looseleaf with annual replacement pages); WILLIAM WALLACE WHITE and BYFLEET G. RAVENSCROFT, *Patents Throughout the World* (looseleaf); and KONSTANTIN KATSAROV, *Manual and Directory on Industrial Property All Over the World*, 7th ed. (1970), are established works presenting summaries of the patent laws of every country, concentrating on details and procedures for obtaining and maintaining patents; the first one mentioned is probably the fullest. J.W. BAXTER, *World Patent Law and Practice* (1971, looseleaf), is a topical compilation. HOBART NOBLE DURHAM, *World Patent Litigation* (1967, looseleaf), separately treats patent suits in 18 countries.

Articles of international scope are published in *International Review of Industrial Property and Copyright Law* (quarterly since 1970), which also includes a section of abstracts of decisions and articles. *Industrial Property*, official publication of the United International Bureau for the Protection of Intellectual Property (English-language edition monthly since 1962), includes articles relating to individual countries and general studies from time to time, in addition to official notices and texts, and annual statistical reports.

The UNITED NATIONS publication, *The Role of Patents in the Transfer of Technology to Developing Countries: Report* (1964), is an economic study focussed on the problems of developing and underdeveloped countries. The most comprehensive and scholarly work on the Paris Convention is STEPHAN P. LADAS, *The International Protection of Industrial Property* (1930), but it is quite out of date. G.H. BODENHOUSEN, *Guide to the Application of the Paris Convention for the Protection of Industrial Property* (1968), is a short explanatory treatment.

There is no recent one-volume legal treatise on patent law in the United States; the section on "Patent Law" in *Corpus Juris Secundum*, vol. 69, pp. 49–59, 160–1126, 1134–1211, with cumulative annual pocket parts, can serve this purpose for want of one. Legal works on British patent law are *Terrell on the Law of Patents*, 11th ed. (1965); and THOMAS ANTHONY BLANCO WHITE, *Patents for Inventions and the Registration of Industrial Designs*, 3rd ed. (1962).

The most recent Canadian text is HAROLD D. FOX, *The Canadi-*

an Law and Practice Relating to Letters Patent for Inventions, 4th ed. (1969). Textbooks in English, other than compendia of the type first mentioned, on the laws of non-English-language countries are scarce. For Japan there is ROBERT WHEELER RUSSELL, *Patent and Trademarks in Japan*, 2nd ed. (1966); and KUKIMOTO, *Summary of the Japanese Patent Law* (1971). There is a good deal of patent law writing in French and considerable in German. The Max-Planck Institute for Foreign and International Patent, Copyright, and Competition Law, in Munich, is responsible for the publication of many studies.

(P.J.F.)

Patna and Pāṭaliputra

Patna is the capital of the state of Bihār in northern India; it is located in the district of the same name. The ancient city of Pāṭaliputra, capital of the Maurya and Gupta empires, occupied approximately the same site between the 6th century BC and the 6th century AD.

History. Pāṭaliputra was founded by Ajātaśatru, king of Magadha (South Bihār) in the 5th century BC. He built a fort there named Pāṭaligrāma to repel the possible attack of the Vṛjjis from the other side of the river Ganges. In the time of his son Udāya (Udāyin), the capital of Magadha was moved there from Rājagṛha (Rājgīr in Bihār subdivision). Between the mid-4th and the late 1st century BC, Pāṭaliputra was the metropolis of three successive Magadha dynasties: the Nanda, Maurya, and Śuṅga.

The Mauryan period

No descriptions of the city survive from the Nanda Period. The Greek ambassador Megasthenes, who stayed there for a number of years in the time of Candragupta Maurya (*c.* 321–*c.* 297 BC), describes the location of Palibothra (Pāṭaliputra) at the confluence of the river Ganges with Erannaboas. According to him, its shape was that of a parallelogram, and it was surrounded by a wooden wall, pierced with loopholes for the discharge of arrows. In his words, "even Susa and Ecbatana could not rival the beauty and grandeur of Pāṭaliputra." The city was the scene of the third Buddhist Council during the reign of the Maurya emperor Aśoka (*c.* 265–238 BC). Pāṭaliputra was sacked by Indo-Greek forces under Demetrius about 185 BC. Their retreat, caused by trouble at home, was immediately followed by a coup organized by the Brahmin commander in chief Puṣyamitra, who killed his master Bṛhadratha (the last Maurya ruler) and ascended the throne as founder of the Śuṅga dynasty. The city lost its political importance after the fall of the Śuṅga (*c.* 73 BC) but continued to be a centre of learning.

Early in the 4th century AD, Pāṭaliputra became the capital of the Gupta state. According to the Chinese traveller Fa-hsien, who visited the city during the reign of Candra Gupta II (reigned *c.* AD 380–*c.* 415), "the royal palace and the halls in the midst of the city, the walls and the gates with the inlaid sculpture work seemed to be the work of superhuman spirits." Shortly thereafter the capital was shifted westward. Pāṭaliputra once again declined and was finally deserted. At the time of the Chinese writer Hsüan-tsang's visit in AD 637, the city lay in ruins. Ashy patches noticed in excavated sites suggest that it may have been destroyed by fire.

Nothing is known of its history from the 7th century until 1541, when it was refounded as Patna by the Āfghan ruler Sher Shāh. Under the Mughals, it regained its old status as the metropolis of Bihār. The Englishman Ralph Fitch, who visited Patna in 1586, described it as "a very long and great town." The emperor Aurangzeb (1659–1707) renamed it ʿAzīmābād for his grandson, ʿAzīm. The city passed into the hands of the British East India Company in 1765 and became the seat of the provincial council for the collection of the company's revenue. In 1865 Patna and Gayā formed the two districts of Bihār. When the new province of Bihār and Orissa was carved out of the old presidency of Bengal in 1912, Patna was selected as its capital. Orissa, however, became a separate province in 1936, and since then Patna has been associated with Bihār alone. The greater part of the city proper, stretching along the river bank for nine miles, has considerably increased in population and expanded in size within the last 50 years. A high Court was set up here for the new province in 1916, followed by a university in 1917.

The modern city

Prominent among its modern structures are the Government House, the Assembly Chambers, the Oriental Library, a medical college, and an engineering college. Its historic monuments include the mosque of Ḥusayn Shāh of Bengal (1499); the Sikh Temple associated with the tenth Gurū, Govinda Singh; and the granary at Bankipore (now Bānkipur; 1786), popularly called the Golghar.

Archaeology. Excavations at Bulandibagh and Kumrahar near Patna have partly exposed the ancient sites of Pāṭaliputra and the Mauryan Palace. L.A. Waddell in 1893 found fragments of a polished column, carved stones, and certain other objects on the basis of which he presumed that Kumrahar represented the site of the palace. His work was continued by D.B. Spooner, who carried out excavations at both sites between 1912 and 1915, locating remains dating from four ancient periods between *c.* 600 BC and AD 600. The earliest period included the characteristic Northern Black Polished Pottery (NBP), with red and grey wares, iron implements, and terra-cotta figurines. Fragmentary architectural pieces bearing the high Mauryan polish were also found in this period. In the second period the same ceramic industries continued, with some deterioration in workmanship. In the third period (AD 100–300), NBP pottery disappears altogether, red and grey ware predominating. A gold amulet modelled on the coins of the Kuṣāṇa king Huviṣka with the legend on one side and Ardochṣo, goddess of abundance, on the other, was also found at this level. The last period (*c.* AD 300–600) yielded plain red ware, terra-cotta figurines, and certain objects of Gupta workmanship. This was followed by a long break until 1600, when the site again showed habitation with finds of glazed pottery, glass beads, ivory dice, and a coin of Shāh ʿĀlam.

Spooner's excavations at Kumrahar led to the discovery of the Mauryan Pillared Hall. Here eight heaps of polished stones in eight rows of ten each were discovered amidst a deposit of charcoal and ash, suggesting that the wooden structure of the hall had been destroyed by fire. Spooner also exposed structures from the post-Mauryan and Gupta periods. Some brick construction of the latter period also was uncovered at Bulandibagh in 1926–27, with a unique wooden construction of long planks at the bottom. A similar one, without the wood planks, was traced at Goshain-Khan, half a mile east of Bulandibagh.

Further excavations were carried out in 1951–55 at Kumrahar in an effort to trace the extension of the Pillared Hall, but all that was found was the remains of several monasteries. The hall was probably burned early in the Śuṅga period, as the structural remains of this period were traced over the ashy layer. The entire area was in continuous occupation from the Mauryan period until *c.* AD 600.

The surrounding district of Patna also has many ancient sites. These include Rājgīr, associated with the life of Buddha; Pāvāpurī, where Mahāvīra, the founder of Jainism, died; and Nālandā, the great intellectual centre of the early medieval period in India. Excavations here revealed seven levels of occupation and buildings of nine different periods from the 5th to 12th centuries AD. The division of Patna includes the area to the south of the Ganges covering 11,336 square miles comprising the districts of Patna, Gayā, and Shāhābād.

BIBLIOGRAPHY. L.A. WADDELL, *Pāṭaliputra* (1892), and his *Report on the Excavations at Pāṭaliputra* (1903), are the two pioneer works. D.B. SPOONER'S fairly exhaustive report on his excavations is published in the *A. Rep. Archaeol. Surv., India* pp. 55–61 (1912–13). The results of the excavations at Kumrahar (1951–55) are published by A.S. ALTEKAR and V. MISHRA in their *Report on Kumrahar* (1959). The history of the Patna district and city is given in the *Imperial Gazetteer of India*, vol. 20 (1908), and also in the Bihār Census Report (Patna District) of 1961. The history of the city and excavations at Pāṭaliputra are also described in B.N. PURI, *Cities of Ancient India* (1966).

(B.N.P.)

Patrick, Saint

The last missionary enterprise of the British church before it sank under the impact of the Anglo-Saxon inva-

sion was the firm establishment of the Christian faith in Ireland. The leader and moving spirit of this undertaking was the bishop Patricius, or Patrick, who thus may also be credited with having made possible the Christianizing by the Irish of their Pictish and Anglo-Saxon neighbours. Accordingly, St. Patrick is venerated in ecclesiastical and popular tradition as the national apostle of Ireland.

Confessio and *Epistola*

Patrick is known only from his own two short works, the *Confessio* ("Confession," or spiritual autobiography), set down in his old age, and his *Epistola* ("Letter"), written in denunciation of the ill-treatment inflicted on some Irish Christian captives by the soldiers of a British chieftain, Coroticus. The authenticity of these sources is unquestioned.

Born in Britain of a Romanized family, at the age of 16 he was torn by Irish raiders from the villa of his father, Calpurnius, a deacon and minor local official, and carried into slavery in Ireland, where, during six bleak years spent as a herdsman, he turned with fervour to the Lord. Hearing at last in a dream that the ship in which he was to escape was ready, he fled his master and found passage to Britain. There he came near to starvation and suffered a second brief captivity before he was finally reunited with his family.

The best known passage in the *Confessio* tells of a dream, after his return to Britain, in which one Victoricus delivered him a letter headed "The Voice of the Irish." As he read it he seemed to hear a certain company of Irish beseeching him to walk once more among them. "Deeply moved," he says, "I could read no more." Nevertheless, because of the shortcomings of his education he was reluctant for a long time to respond to the call. Even on the eve of re-embarkation for Ireland he was beset by doubts of his fitness for the task. Once in the field, however, his hesitations vanished. Utterly confident in the Lord, he journeyed far and wide, baptizing and confirming with untiring zeal. In diplomatic fashion he brought gifts to a kinglet here and a lawgiver there but accepted none from any. On at least one occasion he was cast into chains. On another, he addressed with lyrical pathos a last farewell to his converts who had been slain or kidnapped by the soldiers of Coroticus. Careful to deal fairly with the heathen, he nevertheless lived in constant danger of martyrdom. The evocation of such incidents of what he called his "laborious episcopate" was his reply to a charge, to his great grief endorsed by his ecclesiastical superiors in Britain, that he had originally sought office for the sake of office. In point of fact, he was a most humble-minded man, pouring forth a continuous paean of thanks to his Maker for having chosen him as the instrument whereby multitudes who had worshipped "idols and unclean things" had become "the people of God."

The phenomenal success of Patrick's mission is not, however, the full measure of his personality. Since his writings have come to be better understood, it is increasingly recognized that, despite their occasional incoherence, they mirror a truth and a simplicity of the rarest quality. No diarist has ever bared his inmost soul to the same degree as did the patron saint of Ireland. As D.A. Binchy, the most austerely critical of Patrician scholars, has put it, "The moral and spiritual greatness of the man shines through every stumbling sentence of his 'rustic' Latin."

Uncertainty about the exact dates of his birth and death

It is not possible to say with any assurance when Patrick was born. There are, however, a number of pointers to his missionary career having lain within the second half of the 5th century. In the Coroticus letter, his mention of the Franks as still heathen indicates that the letter must have been written between 451, the date generally accepted as that of the Franks' irruption into Gaul as far as the Somme River, and 496, when they were baptized en masse.

Patrick, who speaks of himself as having evangelized heathen Ireland, is not to be confused with Palladius, sent by Pope Celestine in 431 as "first bishop to the Irish believers in Christ."

Before the end of the 7th century Patrick had become a legendary figure, and the legends have continued to grow. One of these would have it that he drove the snakes of Ireland into the sea to their destruction. Another, probably the most popular, is that of the shamrock, which has him explain the concept of the Holy Trinity, three Persons in one God, to an unbeliever by showing him the three-leaved plant with one stalk. Today Irishmen wear shamrocks, the national flower of Ireland, in their lapels on St. Patrick's Day, March 17.

BIBLIOGRAPHY. For particulars as to the extant manuscript sources, see the definitive edition by L. BIELER of the *Confessio* and *Epistola* in *Libri Epistolarum Sancti Patricii Episcopi*, 2 vol. (1952), and for an English translation of the writings, Bieler's *Works of St. Patrick* (1953). Bibliographical references are covered exhaustively up to 1929 in J.K. KENNEY (ed.), *Sources for the Early History of Ireland*, vol. 1 (1929); studies published from then till 1962 are amply treated in D.A. BINCHY, "Saint Patrick and His Biographers, Ancient and Modern," *Studia Hibernica*, 2:7–173 (1962), an incomparable critical analysis of the Patrician problem. Useful works appearing since 1962 include: L. BIELER, *St. Patrick and the Coming of Christianity* (1967); R.P.C. HANSON, *Saint Patrick: His Origins and Career* (1968); and T. O'RAIFEARTAIGH, "Saint Patrick's Twenty-Eight Days Journey," *Irish Historical Studies*, vol. 16, no. 64 (1969).

(T.O'R.)

Patristic Literature

Patristic literature includes the writings, both orthodox and heretical, of the Fathers of the early Christian Church, from the late 1st century AD to the early 8th century.

This article is divided into the following sections:

NATURE AND SIGNIFICANCE

The concept and periodization of patristic literature. Patristic literature is generally identified today with the entire Christian literature of the early Christian centuries, irrespective of its orthodoxy or the reverse. Taken literally, however, patristic literature should denote the literature emanating from the Fathers of the Christian Church, the Fathers being those respected bishops and other teachers of exemplary life who witnessed to and expounded the orthodox faith in the early centuries. This would be in line with the ancient practice of designating as "the Fathers" prominent church teachers of past generations who had taken part in ecumenical councils or whose writings were appealed to as authoritative. Almost everywhere, however, this restrictive definition has been abandoned. There are several reasons why a more elastic usage is to be welcomed. One is that some of the most exciting Christian authors, such as Origen, were of questionable orthodoxy, and others—Tertullian, for example—deliberately left the church. Another is that the undoubtedly orthodox Fathers themselves cannot be properly understood in isolation from their doctrinally unorthodox contemporaries. Most decisive is the consideration that early Christian literature exists, and deserves to be studied, as a whole, and that much will be lost if any sector is neglected because of supposed doctrinal shortcomings.

Reasons for a broad concept of patristic literature

There has been much difference of opinion about the period that the patristic literature, understood in this more flexible sense, should be held to cover. The editor of

the largest and best known collection of Greek and Latin patristic texts took an exceptionally inclusive line, carrying his Latin series down to the time of Pope Innocent III (d. 1216), and his Greek series down to approximately the fall of Constantinople (1453). It seems more realistic, however, to make the patristic epoch coterminous with the life of the ancient, or premedieval, Catholic Church. Leaving out the New Testament itself, this scheme takes Clement of Rome (bishop, *c.* 95) as a starting point and concludes in the West with Pope Gregory the Great (d. 604) and in the East with John of Damascus (d. *c.* 749). For convenience it is useful to treat the ecumenical Council of Nicaea (325) as dividing this vast stretch of centuries into two manageable sections. Such a cleavage is by no means arbitrary, for the council was much more than a watershed in the evolution of dogmatic theology. In the ante-Nicene era (before 325) the church was relatively unimportant in the Roman empire and perforce on the defensive, but the emperor Constantine's convening of the council was itself a token of the triumph of Christianity and foreshadowed the increasingly influential role it was to play in the post-Nicene era (after 325).

Languages and types of patristic literature. *Languages.* The language of the earliest patristic writings, as of the primitive liturgy, was Greek in both West and East. This is explained partly by the wide diffusion of Greek throughout the whole Mediterranean basin and the prestige it enjoyed in the larger western towns, partly by the fact that Greek-speaking missionaries had first carried the new religion westward. Only toward the end of the 2nd century did Latin make headway, first in North Africa and then in Rome; but after a hesitant start it completely ousted Greek in the West. In the East, Greek was liable to be displaced by Syriac, Coptic, and Armenian, where these were the languages spoken by the educated. The Greek of the Fathers, though incorporating biblical expressions and impressing fresh meanings on certain terms, was substantially identical with contemporary secular Greek; but Christian Latin differed rather more, in vocabulary if not in syntax, from the everyday speech. This was because, being less rich and flexible than Greek, Latin needed a great importation of new words and adaptation of old ones to become an adequate vehicle for Christians.

Almost all of the writings classified as patristic either are theological in content or have a theological colouring. The modern reader who feels put off by this should reflect that the ante-Nicene authors were in the main striving to defend, understand, and explain their faith, while in the post-Nicene period the interest of people of all sorts and conditions in theology and theological debate reached a pitch unparalleled except, perhaps, at the Reformation. Though this theological preoccupation is everywhere evident, the patristic writings exhibit an extraordinary variety of literary types.

Apologetic and polemical literature

Types. From the pre-Nicene centuries there have survived specimens of the early apologetical and polemical literature of the church, the first tentative essays in constructive theology and Christian philosophizing, official and personal letters, and essays in biblical interpretation; there are also fragments of liturgy and hymnody, accounts of martyrdoms, and blueprints for the organization of church life. The post-Nicene literature, considered as literature, reaches a much higher level; this was to be expected, for Christianity by this time was attracting the highly educated and sophisticated in increasing numbers. In both East and West a literary golden age dawned, and even though commonplace works of ecclesiastical controversy and biblical exegesis (critical interpretation) abound, these later centuries can boast magnificent examples of dogmatic expositions, set-piece sermons, historiography, exchanges of correspondence, mystical treatises of elaborate artfulness, and religious poetry.

In the past, classical students were sometimes inclined to dismiss the patristic literature as secondary, the product of a decadent age, but the shortsightedness of this judgment is now generally realized. The writings of the Fathers make a special appeal to Christians because they so graphically recall the development of ecclesiastical institutions and doctrines, indeed the whole life of the adolescent church. But they are equally interesting both to those who explore the ancient Greco-Roman world for its own sake and to those who are attracted to it as the womb out of which Western civilization was to emerge. The interplay of pagan and Christian cultures in these fateful centuries makes a fascinating study; if Christian thinkers were sometimes more powerfully affected by the secular intellectual environment than they cared to acknowledge, those who upheld the views of this latter-day paganism, with its insecurity and yearning for salvation, had more in common with them than appeared on the surface. But the slow, all-embracing transformation of society which, through the medieval milieu and the Renaissance, was to issue in the modern world was already underway. The writings of the Fathers uniquely illuminate both the germinative ideas at work in the process and the new structures that were taking shape.

THE ANTE-NICENE PERIOD

During the first three centuries of its existence, the Christian Church had first to emerge from the Jewish environment that had cradled it and then come to terms with the predominantly Hellenistic (Greek) culture surrounding it. Its legal position at best precarious, it was exposed to outbursts of persecution at the very time when it was working out its distinctive system of beliefs, defining its position vis-à-vis Judaism on the one hand and Gnosticism (a heretical movement that upheld the dualistic view that matter is evil and the spirit good) on the other, and constructing its characteristic organization and ethic. It was a period of flux and experiment, but also one of consolidation and growing self-confidence, and these are all mirrored in its literature.

Earliest patristic writings

The Apostolic Fathers. According to conventional reckoning, the earliest examples of patristic literature are the writings of the so-called Apostolic Fathers; the name derives from their supposed contacts with the Apostles (intimate disciples of Jesus) or the apostolic community. These writings include the church order called the *Didachē*, or *Teaching of the Twelve Apostles* (dealing with church practices and morals), the *Letter of Barnabas*, and the *Shepherd of Hermas*, all of which hovered at times on the fringe of the New Testament canon in that they were used as sacred scripture by some local churches; the *First Letter of Clement*, the seven letters that Ignatius of Antioch (d. *c.* 110) wrote when being escorted to Rome for his martyrdom, the related *Letter to the Philippians* by Polycarp of Smyrna (d. *c.* 156 or 168), and the narrative report of Polycarp's martyrdom; some fragmentary accounts of the origins of the Gospels by Papias (fl. late 1st or early 2nd century AD), bishop of Hierapolis in Phrygia, Asia Minor; and an ancient homily (sermon) known as the *Second Letter of Clement*. They all belong to the late 1st or early 2nd century and are all to a greater or lesser extent influenced (sometimes by way of reaction) by the profoundly Jewish atmosphere that pervaded Christian thinking and practice at this primitive stage. For this reason alone, modern scholars tend to regard them as a somewhat arbitrarily selected group. A more scientific assessment would place them in the context of a much wider contemporary Jewish-Christian literature that has largely disappeared but whose character can be judged from apocryphal (or noncanonical) works such as the *Ascension of Isaiah*, the *Odes of Solomon*, and certain extracanonical texts modelled on the New Testament.

Even with this qualification, the Apostolic Fathers, with their rich variety of provenance and genre (types), illustrate the difficult doctrinal and organizational problems with which the church was grappling in those transitional generations. Important among these problems were the creation of a ministerial hierarchy and of an accepted structure of ecclesiastical authority. The *Didachē*, which is Syrian in background and possibly the oldest of these documents, suggests a phase when Apostles and prophets were still active but when the routine ministry of bishops and deacons was already winning recognition. The *First*

Letter of Clement, an official letter from the Roman to the Corinthian Church, reflects the more advanced state of a collegiate episcopate, with its shared authority among an assembly of bishops. This view of authority was supported by an emergent theory of apostolic succession in which bishops were regarded as jurisdictional heirs of the early Apostles. The *First Letter of Clement* is also instructive in showing that the Roman Church, even in the late 1st century, was asserting its right to intervene in the affairs of other churches. The letters of Ignatius, bishop of Antioch at the beginning of the 2nd century, depict the position of the monarchical bishop, flanked by subordinate presbyters (priests) and deacons (those serving priests in the liturgy), which had been securely established in Asia Minor.

Relationship of Judaism and Christianity

Almost more urgent was the question of the relation of Christianity to Judaism, and in particular of the Christian attitude toward the Old Testament. In the *Didachē* there is little sign of embarrassment; Jewish ethical material is taken over with suitable adaptations, and the Jewish basis of the liturgical elements is palpable. But with *Barnabas* the tension becomes acute; violently anti-Jewish, the Alexandrian author substitutes allegorism (use of symbolism) for Jewish literalism, and thus enables himself to wrest a Christian meaning from the Old Testament. The same tension is underlined by Ignatius' polemic against Judaizing tendencies in the church. At the same time all these writings—especially those of Ignatius, Polycarp, and Papias—testify to the growing awareness of a specifically Christian tradition embodied in the teaching transmitted from the Apostles.

Almost all the Apostolic Fathers throw light on primitive doctrine and practice. The *Didachē*, for example, presents the Eucharist as a sacrifice, and *I Clement* incorporates archaic prayers. *II Clement* invites its readers to think of Christ as of God, and of the church as a preexistent reality. Hermas seeks to modify the rigorist view that sin committed after Baptism cannot be forgiven. But the real key to the theology of the Apostolic Fathers, which also explains its often curious imagery, is that it is Jewish-Christian through and through, expressing itself in categories derived from latter-day Judaism and apocalyptic literature (that depicting the intervention of God in history in the last times), which were soon to become unfashionable and be discarded.

The Gnostic writers. Hardly had the church thrown off its early Jewish-Christian idiosyncrasies when it found itself confronted by the amorphous but pervasive philosophical-religious movement known as Gnosticism. This movement made a strong bid to absorb Christianity in the 2nd century, and a number of Christian Gnostic sects flourished and contributed richly to Christian literature. Although the church eventually maintained its identity intact, the confrontation forced it to clarify its ideas on vital issues on which it differed sharply from the Gnostics. Chief among these were the Gnostics' distinction between the unknown supreme God and the Demiurge (identified with the God of the Old Testament) who created this world; their dualist disparagement of the material order and insistence that the Redeemer became incarnate in appearance only; their belief in salvation by esoteric knowledge; and their division of mankind into a spiritual elite capable of achieving salvation and, on a lower plane, inferior grades of "psychic" and "material" men.

Gnostic emphasis

Among the leading 2nd-century Christian Gnostics were Saturninus and Basilides, reputedly pupils of Menander, a disciple of Simon Magus (late 1st century), the alleged founder of the movement; they worked at both Antioch and Alexandria. But most famous and influential was the Egyptian Valentinus, who acquired a great reputation at Rome (*c.* 150) and founded an influential school of thought. Basilides and Valentinus are reported to have written extensively, and their systems can be reconstructed from hostile accounts by Irenaeus (*c.* 120/140–*c.* 200/203), Clement of Alexandria (*c.* 150–*c.* 215), and other orthodox critics. But the Gnostics generally seem to have been prolific writers, and as they needed their own distinctive scriptures they soon created a body of apocryphal books patterned on the New Testament. It was a Syrian Gnostic convert, Tatian, who compiled (late 2nd century) the first harmony of the Four Gospels (the *Diatessaron*)—a single gospel using the material from the Gospels; and an Italian Gnostic, Heracleon (2nd century), who prepared the earliest commentary on St. John's Gospel (extracts from it have been preserved by Origen). Epiphanius (*c.* 315–403) preserved a *Letter to Flora*, by the Valentinian Gnostic Ptolemaeus (late 2nd century), supplying rules for interpreting the Mosaic Law (the Torah) in a Christian sense; and another disciple of Valentinus, Theodotus (2nd century), published an account of his master's system that was excerpted by Clement of Alexandria.

Anti-Gnostic sources

Almost the entire vast literature of Gnosticism has perished, and until recently the only original documents available to scholars (apart from extracts like those already mentioned which were preserved by orthodox critics) were a handful of treatises in Coptic contained in three codices (manuscript books) that were discovered in the 18th and late 19th centuries. The most interesting of these are *Pistis Sophia* and the *Apocryphon of John*, the former consisting of conversations of the risen Jesus with his disciples about the fall and redemption of the aeon (emanation from the Godhead) called Pistis Sophia, the latter of revelations made by Jesus to St. John explaining the presence of evil in the cosmos and showing how man can be rescued from it.

The Naj' Ḥammādī codices

Since 1946, however, this meagre store has been richly supplemented by the discovery near Naj' Ḥammādī, in Egypt on the Nile about 78 miles northwest of Luxor, of 13 codices containing more than 100 Christian Gnostic treatises in Coptic translations. The best known of these, the Jung Codex (named in honour of the psychoanalyst Carl Jung by those who purchased it for his library), includes four important items: a *Letter of James*, recording revelations imparted by the risen Christ to the Apostles; the *Gospel of Truth*, perhaps to be identified with the work of this name attributed by Irenaeus to Valentinus; the *Letter to Rheginus*, a Valentinian work, possibly by Valentinus himself, on the Resurrection; and a *Treatise on the Three Natures*, also Valentinian. Of the other documents from the Naj' Ḥammādī library thus far published, the one that has attracted most attention is the *Gospel of Thomas*, a collection of sayings and parables that are ascribed to Jesus.

A figure of immense significance who is often, though perhaps mistakenly, counted among the Gnostics was Marcion, who after breaking with the Roman Church in 144 set up a successful organization of his own. Teaching that there is a radical opposition between the Law and the Gospel, he refused to identify the God of love revealed in the New Testament with the wrathful Creator God of the Old Testament. He set forth these contrasts in his *Antitheses*, and his adoption of a reduced New Testament consisting of St. Luke's Gospel and certain Pauline epistles, all purged of presumed Jewish interpolations, had an important bearing on the church's formation of its own fuller canon.

Characteristics of apologetic literature

The Apologists. Meanwhile, the orthodox literature of the 2nd and early 3rd centuries tends to have a distinctly defensive or polemical colouring. It is the age of Apologists, and these Apologists had to engage in battle on two fronts. First, there was the hostility and criticism of pagan society. Because of its very aloofness the church was popularly suspected of sheltering all sorts of immoralities and thus of threatening the established order. At a higher level, Christianity, as it became better known, was being increasingly exposed to intellectual attack. The physician Galen of Pergamon (d. *c.* 200) and the Middle Platonist thinker Celsus, who followed the religiously inclined form of Platonism that flourished from the 3rd century BC to the 3rd century AD (*cf.* his devastating *Alēthēs logos*, or *True Word*, written *c.* 178), were only two among many "cultured despisers." But, secondly, orthodoxy had to take issue with distorting tendencies within, whether these took the form of Gnosticism or of other heresies, such as the so-called semi-Gnostic Marcion's rejection of the Old Testament revelation or the claim of

the ecstatic prophet from Phrygia, Montanus (*c.* 157), to be the vehicle of a new outpouring of the Holy Spirit. Christianity had also to define exactly where it stood in relation to Hellenistic culture.

The early Apologists (2nd century). Strictly speaking, the term Apologists denotes the group of 2nd-century writers who defended Christianity against external critics, pagan and Jewish. The earliest of this group was Quadratus, who in about 124 addressed an apology for the faith to the emperor Hadrian; apart from a single fragment, it is now lost. Other early 2nd-century Apologists who are mere names known to scholars are Aristo of Pella, the first to prepare an apology to counter Jewish objections; and Apollinaris, bishop of Hierapolis, said to be the author of numerous apologetic works and also of a critique of Montanism. An early apology that has survived intact is that of Aristides, addressed *c.* 140 to the emperor Antoninus Pius; after being completely lost, the text was rediscovered in the 19th century. The most famous Apologist, however, was Justin, who was converted to Christianity after trying various philosophical schools, paid lengthy visits to Rome, and was martyred there (*c.* 165). Justin's two *Apologies* are skillful presentations of the Christian case to the pagan critics; and his *Dialogue with Trypho* is an elaborate defense of Christianity against Judaism.

Relationship to philosophy

Justin's attitude to pagan philosophy was positive. His pupil Tatian could see nothing but evil in the Greco-Roman civilization. Indeed, Tatian's *Discourse to the Greeks* is less a positive vindication of Christianity than a sharp attack on paganism. His contemporary Athenagoras of Athens, author of the apologetic work *Supplication for the Christians* and a treatise *On the Resurrection of the Dead*, is as friendly as Justin to Greek culture and philosophy. Two others who deserve mention are Theophilus of Antioch, a prolific publicist whose only surviving work is *To Autolycus*, prepared for his pagan friend Autolycus; and the anonymous author of the *Letter to Diognetus*, an attractive and persuasive exposition of the Christian way of life that is often included among the Apostolic Fathers.

As stylists the Apologists reach only a passable level; even Athenagoras scarcely achieves the elegance at which he obviously aimed. But they had little difficulty in refuting the charges of atheism, cannibalism, promiscuity, and other spurious charges popularly brought against Christians, or in mounting a counterattack against the debasements of paganism. More positively, they strove to vindicate the Christian understanding of God and specific doctrines such as the divinity of Christ and the resurrection of the body. In so doing, most of them exploited current philosophical conceptions, in particular that of the *Logos* (Word), or rational principle underlying and permeating reality, which they regarded as the divine reason, become incarnate in Jesus. They have been accused of Hellenizing Christianity (making it Greek in form and method), but they were in fact attempting to formulate it in intellectual categories congenial to their age. In a real sense they were the first Christian theologians.

Eristic (missionary) Apologists. As the 2nd century advanced, a more confident, aggressive spirit came over Christian Apologists, and their intellectual and literary stature increased greatly. Clement of Alexandria (150–215), for example, while insisting on the supremacy of faith, freely draws on Platonism and Stoicism to clarify Christian teaching. In his *Protrepticus* ("Exhortation") and *Paedagogus* ("Instructor") he urged pagans to abandon their futile beliefs, accept the Logos as guide, and allow their souls to be trained by him. In interpreting scripture he used an allegorizing method derived from a Jewish philosopher, Philo (*c.* 15 BC–after AD 40), and against Gnosticism he argued that the baptized believer who studies the Scriptures is the true Gnostic, faith being at once superior to knowledge and the beginning of knowledge.

Rapprochement with and reaction to Greco-Roman culture

The critique of Gnosticism was much more systematically developed by Clement's older contemporary, Irenaeus (*c.* 120/140–*c.* 200/203) of Lyons, in present-day France, in his voluminous *Against Heresies.* While countering the Valentinian dualism that asserted that spirit was good and matter evil, this treatise makes clear the church's growing reliance on its creed or "rule of faith," on the New Testament canon, and on the succession of bishops as guarantors of the true apostolic tradition. Irenaeus was also a constructive theologian, expounding ideas about God as creator, about the Son and the Spirit as his "two hands," about Christ as the new Adam who reconciles fallen man with God, and about the worldwide church with its apostolic faith and ministry, a concept which theology was later to take up eagerly.

More brilliant as a stylist and controversialist, the North African lawyer Tertullian (*c.* 160–after 220) was also the first Latin theologian of considerable importance. Unlike Clement, he reacted with hostility to pagan culture, scornfully asking, "What has Athens to do with Jerusalem?" His *Apology* remains a classic of ancient Christian literature, and his numerous moral and practical works reveal an uncompromisingly rigid moral view. Although later becoming a Montanist himself (a follower of the morally rigorous and prophetic sect founded by Montanus), he wrote several antiheretical tracts, full of abuse and biting sarcasm. Yet, in castigating heresy he was able to formulate the terminology, and to some extent the theory, of later Trinitarian and Christological orthodoxy; his teaching on the fall of man, aimed against Gnostic dualism, in part anticipates Augustine.

Roughly contemporary with Tertullian, and like him an intellectual and a rigorist, was Hippolytus (*c.* 170–*c.* 235), a Greek-speaking Roman theologian and antipope. He, too, had a vast literary output, and although some of the surviving works attributed to him are disputed, it is probable that he wrote the comprehensive *Refutation of All Heresies,* attacking Gnosticism, as well as treatises denouncing specifically Christian heresies. He was also the author both of numerous commentaries on scripture and (probably) of the *Apostolic Tradition*, an invaluable source of knowledge about the primitive Roman liturgy. His *Commentary on Daniel* (*c.* 204) is the oldest Christian biblical commentary to survive in its entirety. His exegesis (interpretive method) is primarily typological; *i.e.*, treating the Old Testament figures, events, and other aspects as "types" of the new order which was inaugurated by Christ.

Late 2nd to early 4th century. *Alexandria.* Meanwhile, a brilliant and distinctive phase of Christian literature was opening at Alexandria, the chief cultural centre of the empire and the meeting ground of the best in Hellenistic Judaism, Gnosticism, and Neoplatonism. Marked by the desire to present Christianity in intellectually satisfying terms, this literature has usually been connected with the Catechetical School, which, according to tradition, flourished at Alexandria from the end of the 2nd through the 4th century. Except for the brief period, however, when Origen was in charge of it (202/203–230/231), it may be doubted whether the school was ever itself a focus of higher Christian studies. When speaking of the School of Alexandria, some scholars claim that it is better to think of a distinguished succession of like-minded thinkers and teachers who worked there, and whose highly sophisticated interpretation of Christianity exercised for generations a formative impact on large sectors of eastern Christendom.

The Catechetical School of Alexandria

The real founder of this theology, with its Platonist leaning, its readiness to exploit the metaphysical implications of revelation, and its allegorical understanding of scripture, was Clement, the Christian humanist whose welcoming attitude to Hellenism and critique of Gnosticism were noted above. His major work, the *Stromata* ("Miscellanies"), untidy and deliberately unsystematic, brings together the inheritance of Jewish Christianity and Middle Platonism in what aspires to be a summary of Christian gnosis (knowledge). All his reasoning is dominated by the idea of the Logos who created the universe and who manifests the ineffable Father alike in the Old Testament Law, the philosophy of the Greeks, and finally the incarnation of Christ. Clement is also a mystic for whom the higher life of the soul is a continuous moral and spiritual ascent.

Significance and influence of Origen

But it is Origen whose achievement stamps the Alexandrian school. First and foremost, he was an exegete (critical interpreter), as determined to establish the text of scripture scientifically (*cf.* his *Hexapla*) as to wrest its spiritual import from it. In homilies, scholia (annotated works), and continuous commentaries he covered the whole Bible, deploying a subtle, strongly allegorical exegesis designed to bring out the several levels of significance it contains. As an apologist, in his *Contra Celsum*, he refuted step-by-step the pagan philosopher Celsus' damaging onslaught on Christianity. In all his writings, but especially his *On First Principles*, Origen shows himself to be one of the most original and profound of speculative theologians. Neoplatonist in background, his system embraces both the notion of the pre-existence of souls, with their fall and final restoration, and a deeply subordinationist doctrine of the Trinity: *i.e.*, one in which the Son is subordinate to the Father. For his spiritual teaching, with its emphasis on the battle against sin, on freedom from passions, and on the soul's mystical marriage with the Logos, his *Commentary on Canticles* provides an attractive introduction.

Origen's influence on Christian doctrine and spirituality was to be immense and many-sided; the orthodox Fathers and the leading heretics of the 4th century alike reflect it. Meanwhile, the Alexandrian tradition was maintained by several remarkable disciples. Two of these whose works have been entirely lost but who are reported to have been polished writers were Theognostus (fl. 250–280) and Pierius (fl. 280–300), both heads of the Catechetical School and apparently propagators of Origen's ideas. But there are two others of note, Dionysius of Alexandria (*c.* 200–*c.* 265) and Gregory Thaumaturgus (*c.* 213–*c.* 270), of whose works some fragments have survived. Dionysius of Alexandria wrote on natural philosophy and the Christian doctrine of creation but is chiefly remembered for his dispute with Pope Dionysius (d. 268) of Rome on the correct understanding of the Trinity. In this Dionysius of Alexandria is revealed as a faithful exponent of Origen's pluralism and subordinationism. Gregory Thaumaturgus has left a fascinating *Panegyric to Origen*, giving a graphic description of Origen's method of instruction, as well as a dogmatically important *Symbol* and a *Canonical Epistle* that is in effect one of the most ancient treatises of casuistry (*i.e.*, the application of moral principles to practical questions).

Anti-Origenist reaction. If Origen inspired admiration, his daring speculations also provoked criticism. At Alexandria itself, Peter, who became bishop *c.* 300 and composed theological essays of which only fragments remain, attacked Origen's doctrines of the pre-existence of souls and their return into the condition of pure spirits. But the acutest of his critics was Methodius of Olympus (d. 311), of whose treatises *The Banquet*, exalting virginity, survives in Greek, and others mainly in Slavonic translations. Although indebted to Alexandrian allegorism, Methodius remains faithful to the Asiatic tradition (literal and historical) of Irenaeus—who had come to France from Asia Minor—and his realism, and castigates Origen's ideas on the pre-existence of souls, the flesh as the spirit's prison, and the spiritual nature of the resurrected body. As a writer he strove after literary effect, and Jerome, writing a century later, praises the excellence of his style.

Pre-Nicene Latin literature. Latin Christian literature was slow in getting started, and North Africa has often been claimed as its birthplace. Tertullian, admittedly, is the first Christian Latinist of genius, but he evidently had his humbler predecessors. Latin versions of the Bible, recoverable in part from manuscripts, were appearing in Africa, Gaul, and Italy during the 2nd century. In that century, too, admired works like *I Clement*, *Barnabas*, and the *Shepherd of Hermas* were translated into Latin. The oldest original Latin texts are probably the Muratorian Canon, a late 2nd-century Roman canon, or list of works accepted as scripture, and the *Acts of the Scillitan Martyrs* (180) of Africa.

Origin and gradual emergence of Latin patristic literature

The first noteworthy Roman Christian to use Latin was Novatian (*c.* 200–*c.* 258), the leader of a rigorist schismatic group. His surviving works reveal him as an elegant stylist, trained in rhetoric and philosophy, and a competent theologian. His doctrinally influential *De trinitate* ("Concerning the Trinity") is basically apologetic: against Gnostics it defends the oneness and creative role of Almighty God, against Marcion it argues that Christ is the Son of God the Creator, against Docetism (the heresy claiming that Jesus only seemed the Christ) that Christ is truly man, and against Sabellianism (the denial of real distinctions in the Godhead) that in spite of Christ's being fully divine there is but one God. His rigorous moralism comes out in his *On Public Shows* and *On the Excellence of Chastity* (both once attributed to Cyprian); in *On Jewish Foods* he maintains that the Old Testament food laws no longer apply to Christians, the animals which were classified as unclean having been intended to symbolize vices.

A much greater writer than Novatian was his contemporary and correspondent, Cyprian (*c.* 200–*c.* 258), the statesmanlike bishop of Carthage. A highly educated convert to Christianity, Cyprian left a large corpus of writings, including 65 letters and a number of moral, practical, and theological treatises. As an admirer of Tertullian, he continued some of his fellow North African's tendencies, but his style is more classical, though much less brilliant and individual. Cyprian's letters are a mine of information about a fascinating juncture in church history. His collections of *Three Books of Testimonies to Quirinus*, or authoritative scripture texts, illustrate the church's reliance on these in defending its theological and ethical positions. A work that has been of exceptional importance historically is his *Unity of the Catholic Church*, in which Cyprian contends that there is no salvation outside the church and defines the role of the Roman see (episcopal authority). His *To Demetrianus* is an original, powerful essay refuting the allegation of pagans that Christianity was responsible for the calamities afflicting society.

Defense of Roman episcopal authority

Three writers from the later portion of this period deserve mention. Victorinus of Pettau, converted about 335 from Neoplatonism, was the first known Latin biblical exegete; of his numerous commentaries the only one that remains is the commentary on Revelation, which maintained a millenarian outlook—predicting the 1,000-year reign of Christ at the end of history—and was clumsy in style. Arnobius the Elder (converted by 300) sought in his *Adversus nationes* ("Against the Nations"), like Tertullian and Cyprian before him, to free Christianity from the charge of having caused all the evils plaguing the empire, but ended up by launching a violent attack on the contemporary pagan cults. A surprising feature of this ill-constructed, verbose apology is Arnobius' apparent ignorance concerning several cardinal points of Christian doctrine, combined with his great enthusiasm for his newfound faith.

By contrast, his much abler pupil Lactantius (*c.* AD 240–*c.* 320), like him a native of North Africa, was a polished writer and the leading Latin rhetorician of the day. His most ambitious work, the *Divine Institutes*, attempted, against increasingly formidable pagan attacks, to portray Christianity as the true form of religion and life, and is in effect the first systematic presentation of Christian teaching in Latin. The later *Deaths of Persecutors*, now generally recognized as his, describes the grim fates of persecuting emperors; it is a primary source for the history of the early 4th century and also represents a crude attempt at a Christian philosophy of history.

THE POST-NICENE PERIOD

The flowering of early Christian literature

The 4th and the first half of the 5th century witnessed an extraordinary flowering of Christian literature, the result partly of the freedom and privileged status now enjoyed by the church, partly of the diversification of its own inner life (*cf.* the rise of monasticism), but chiefly of the controversies in which it hammered out its fundamental doctrines.

Arianism, which denied Christ's essential divinity, aroused an all-pervasive reaction in the 4th century; the task of the first two ecumenical councils, at Nicaea (325)

and Constantinople (381), was to affirm the orthodox doctrine of the Trinity. In the 5th century the Christological question moved to the fore, and the Council of Chalcedon (451), completing that of Ephesus (431), defined Christ as one person in two natures. The Christological controversies of the 5th century were extremely complex, involving not only theological issues but also issues of national concerns—especially in the Syrian-influenced East, where the national churches were called non-Chalcedonian because they rejected the doctrinal formulas of the Council of Chalcedon.

Involved in the 5th-century Christological controversy were many persons and movements: Nestorius, consecrated patriarch of Constantinople in 428, and his followers, the Nestorians, who were concerned with preserving the humanity of Christ as well as his divinity; Cyril (*c.* 375–444), patriarch of Alexandria, and his followers, who were devoted to maintaining a balanced emphasis on both of the natures of Christ, divine and human; Eutyches (*c.* 378–after 451), a muddleheaded archimandrite (head of a monastery) who affirmed two natures before and one nature after the incarnation; the Monophysites, who (following Eutyches) stressed the one unified nature of Christ; the moderates and those who sought theological, ecclesiastical, and even political solutions to this highly complex doctrinal dispute, such as Pope Leo I (d. 461). It was a time when the Alexandrian and Antiochene theological schools vied with each other for the control of the theology of the church. In the Syrian East the Antiochene tradition continued in the schools of Edessa and Nisibis, which became centres of a non-Greek national renaissance. The issues of grace, free will, and the fall of man concerned the West mainly. Meanwhile, old literary forms were developing along more mature lines, and new ones were emerging, including historiography, lives of saints, set piece (fixed-form) oratory, mystical writings and hymnody.

Literature of the Arian controversy

The Nicene Fathers. A seesaw struggle between Arians and orthodox Christians dominated the immediate post-Nicene period. Arius (d. 336) himself, Eusebius of Nicomedia (d. *c.* 342), and other radicals occupied the extreme left wing, carrying Origen's views on the subordination of the Son to what became dangerous lengths. Apart from a few precious letters and fragments, their writings have perished. On the extreme right Athanasius (*c.* 293–373), Eustathius of Antioch (d. *c.* 337), and Marcellus of Ancyra (d. *c.* 374 and strongly anti-Origenist) tenaciously upheld the Nicene decision that the Son was of the same substance with the Father. Again, the writings of the two latter figures, except for scattered but illuminating fragments, have disappeared. Most churchmen preferred the middle ground; loyal to the Origenist tradition, they suspected the Nicene Creed of opening the door to Sabellianism (the heresy that defended the oneness of God and viewed the Father, Son, and Holy Spirit as three successive modes of revelation), but were equally shocked by Arianism in its more uncompromising forms. Eusebius of Caesarea (*c.* 260/264–*c.* 340) was their spokesman, and for decades the eastern emperors supported his mediating line.

Eusebius is chiefly known as a historian; his *Ecclesiastical History*, with its scholarly use of documents and guiding idea that the victory of Christianity is the proof of its divine origin, introduced something novel and epoch-making. But he also wrote voluminous apologetic treatises, biblical and exegetical works, and polemical tracts against Marcellus of Ancyra. From these can be gathered his theology of the Word, which was Origenist in inspiration and profoundly subordinationist and which made the strict Nicenes suspect him as an ally of Arius. Such suspicions were unjust, for he upheld Origen's doctrine of eternal generation (*i.e.*, that the Word is generated outside the category of time) and rejected the extreme Arian theses. His influence can be studied in the works of Cyril of Jerusalem (*c.* 315–386?), whose *Catecheses*, or introductory lectures on Christian doctrine for candidates for Baptism, exemplify a pastoral type of Christian literature. Though critical of the Arian positions, Cyril remained reserved in his attitude toward the Nicene theology and at several other points showed affinities with Eusebius.

Significance of Athanasius

Athanasius bestrides the 4th century as the inflexible champion of the Nicene dogma. He had been present at the council, defending Alexander, the theologian-bishop of Alexandria from 313 to 328, who had exposed Arius; and after succeeding Alexander in 328 he spent the rest of his stormy life defending, expounding, and drawing out the implications of the Nicene theology. His most thorough and effective exposition of the Son's eternal origin in the Father and essential unity with him is contained in his three *Discourses Against the Arians;* but in addition he produced a whole series of treatises, historical or dogmatic or both, as well as letters, covering different aspects of the controversy.

It would be misleading, however, to delineate Athanasius exclusively as a polemicist. First, even in his polemical writings he was working out a positive doctrine of the triune God that anticipated later formal definitions. His *Letters to Bishop Serapion*, with their persuasive presentation of the Holy Spirit as a consubstantial (of the same substance) person in the Godhead, are an admirable illustration. Also, his noncontroversial works, such as the relatively early but brilliant apologies, *Discourse Against the Pagans* and *On the Incarnation*; the attractive and influential *Life of St. Antony*, which was to give a powerful impulse to monasticism (especially in the West); and his numerous exegetical and ascetic essays, which survive largely in fragments, sometimes in Coptic or Syriac translations, should not be overlooked.

Contributions to theology and monasticism

The Cappadocian Fathers. Although Athanasius prepared the ground, constructive agreement on the central doctrine of the Trinity was not reached in his lifetime, either between the divided parties in the East or between East and West with their divergent traditions. The decisive contribution to the Trinitarian argument was made by a remarkable group of philosophically minded theologians from Cappadocia (in eastern Turkey)—Basil of Caesarea (*c.* 330–379), his younger brother Gregory of Nyssa (*c.* 335–*c.* 394), and his lifelong friend Gregory of Nazianzus (*c.* 329–*c.* 389/390). Of aristocratic birth and consummate culture, all three were drawn to the monastic ideal, and Basil and Gregory of Nazianzus achieved literary distinction of the highest order. While their joint accomplishments in doctrinal definition were indeed outstanding, each made a noteworthy mark in other fields as well.

So far as Trinitarian dogma is concerned, the Cappadocians succeeded, negatively, in overthrowing Arianism in the radical form in which two acute thinkers, Aëtius (d. *c.* 366) and Eunomius (d. *c.* 394), had revived it in their day, and, positively, in formulating a conception of God as three Persons in one essence that eventually proved generally acceptable. The oldest of Basil's dogmatic writings is his only partially successful *Against Eunomius*, the most mature his essay *On the Holy Spirit*. Gregory of Nyssa continued the attack on Eunomius in four massive treatises and published several more positive dogmatic essays, the most successful of which is the *Great Catechetical Oration*, a systematic theology in miniature. The output of Gregory of Nazianzus was much smaller, but his 45 *Orations*, as well as being masterpieces of eloquence, contain his classic statement of Trinitarian orthodoxy. Basil's vast correspondence testifies to his practical efforts to reconcile divergent movements in Trinitarian thinking.

Basil is famous as a letter writer and preacher and for his views on the appropriate attitude of Christians toward Hellenistic culture; but his achievement was not less significant as a monastic legislator. His two monastic rules, used by St. Benedict and still authoritative in the Greek Church, are tokens of this. Gregory of Nazianzus, too, was an accomplished letter writer, but his numerous, often lengthy poems have a special interest. Dogmatic, historical, and autobiographical, they are often intensely personal and lay bare his sensitive soul. On the other hand, Gregory of Nyssa, much the most speculative of the three, was an Origenist both in his allegorical interpretation of scripture and in aspects of his eschatology. But he is chiefly remarkable as a pioneer of Christian

mysticism, and in his *Life of Moses, Homilies on Canticles,* and several other books he describes how the soul, in virtue of having been created in the divine image, is able to ascend, by successive stages of purification, to a vision of God.

Uncritical theological views

A figure who stood in sharp contrast, intellectually and in temperament, to the Cappadocians was their contemporary, Epiphanius (*c.* 315–403) of Salamis, in Cyprus. A fanatical defender of the Nicene solution, he was in no sense a constructive theologian like them, but an uncritical traditionalist who rejected every kind of speculation. He was an indefatigable hammer against heretics, and his principal work, the *Panarion* ("Medicine Chest"), is a detailed examination of 80 heresies (20 of them pre-Christian); it is invaluable for the mass of otherwise unobtainable documents it excerpts. Conformably with Epiphanius' contempt for classical learning, the work is written in Greek without any pretension to elegance. His particular *bête noire* was Origen, to whose speculations and allegorism he traced virtually all heresies.

Monastic literature. From the end of the 3rd century onward, monasticism was one of the most significant manifestations of the Christian spirit. Originating in Egypt and spreading thence to Palestine, Syria, and the whole Mediterranean world, it fostered a literature that illuminates the life of the ancient church.

Both Anthony (*c.* 250–355), the founder of eremitical, or solitary, monasticism in the Egyptian desert, and Ammonas (fl. *c.* 350), his successor as leader of his colony of anchorites (hermits), wrote numerous letters; a handful from the pen of each is extant, almost entirely in Greek or Latin translation of the Coptic originals. Those of Ammonas are particularly valuable for the history of the movement and as reflecting the uncomplicated mysticism that inspired it. The founder of monastic community life, also in Egypt, was Pachomius (*c.* 290–346), and the extremely influential rule that he drew up has been preserved, mainly in a Latin translation made by Jerome.

Though these and other early pioneers were simple, practical men, monasticism received a highly cultivated convert in 382 in Evagrius Ponticus. He was the first monk to write extensively and was in the habit of arranging his material in groups of a hundred aphorisms, or "centuries," a literary form that he invented and that was to have a great vogue in Byzantine times. A master of the spiritual life, he classified the eight sins that undermine the monk's resolution, and also the ascending levels by which the soul rises to wordless contemplation. Later condemned as an Origenist, he was deeply influential in the East, and, through John Cassian (360–435), in the West as well.

Influence of monastic literature on ethics and mysticism

Side by side with works composed by monks there sprang up a literature concerned with them and the monastic movement. Much of it was biographical, the classic example being Athanasius' *Life of St. Antony.* Sulpicius Severus took this work as his model when early in the 5th century he wrote his *Life of St. Martin of Tours* (*c.* 330–397), the first Western biography of a monastic hero and the pattern of a long line of medieval lives of saints. But it was Palladius (*c.* 363–before 431), a pupil of Evagrius Ponticus (d. 399), who proved to be the principal historian of primitive monasticism. His *Lausiac History* (so called after Lausus, the court chamberlain to whom he dedicated it), composed about 419/420, describes the movement in Egypt, Palestine, Syria, and Asia Minor. Since much of the work is based on personal reminiscences or information received from observers, it is, despite the legendary character of many of its narratives, an invaluable source book.

Finally, no work so authentically conveys the spirit of Egyptian monasticism as the *Apophthegmata Patrum* ("Sayings of the Fathers"). Compiled toward the end of the 5th century, but using much older material, it is a collection of pronouncements of the famous desert personalities and anecdotes about them. The existing text is in Greek, but it probably derives from an oral tradition in Coptic.

The School of Antioch. Antioch, like Alexandria, was a renowned intellectual centre, and a distinctive school of Christian theology flourished there and in the surrounding region throughout the 4th and the first half of the 5th century. In contrast to the Alexandrian school, it was characterized by a literalist exegesis and a concern for the completeness of Christ's manhood. Little is known of its traditional founder, the martyr-priest Lucian (*d.* 312), except that he was a learned biblical scholar who revised the texts of the Septuagint and the New Testament. His strictly theological views, though a mystery, must have been heterodox, for Arius, Eusebius of Nicomedia, and other Arians claimed to be his disciples ("fellow Lucianists"), and Bishop Alexander of Alexandria, who denounced them, lists Lucian among those who influenced them. But Eustathius of Antioch (d. *c.* 337), the champion of Nicene orthodoxy, is probably more representative of the school, with his antipathy to what he regarded as Origen's excessive allegorism and his recognition, as against the Arians, of the presence of a human soul in the incarnate Christ.

Literalistic views of the School of Antioch

It was, however, much later in the 4th century, in the person of Diodore of Tarsus (*c.* 330–*c.* 390), that the School of Antioch began to reach the height of its fame. Diodore courageously defended Christ's divinity against Julian the Apostate, the Roman emperor (361–363) who attempted to revive paganism, and in his lifetime was regarded as a pillar of orthodoxy. Later critics detected anticipations of Nestorianism (the heresy upholding the division of Christ's Person) in his teaching, and as a result his works, apart from some meagre fragments, have perished. They were evidently voluminous and wide-ranging, covering exegesis, apologetics, polemics, and even astronomy; and he not only strenuously opposed Alexandrian allegorism but also expounded the Antiochene *theoria,* or principle for discovering the deeper intention of scripture and at the same time remaining loyal to its literal sense.

Significant Antiochene theologians

In stature and intellectual power Diodore was overshadowed by his two brilliant pupils, Theodore of Mopsuestia (*c.* 350–428/429) and John Chrysostom (*c.* 347–407). Both had also studied under the famous pagan Sophist rhetorician Libanius (314–393), thereby illustrating the cross-fertilization of pagan and Christian cultures at this period. Like Diodore, Theodore later fell under the imputation of Nestorianism, and the bulk of his enormous literary output—comprising dogmatic as well as exegetical works—was lost. Fortunately, the 20th century has seen the recovery of a few important texts in Syriac translations (notably his *Commentary on St. John* and his *Catechetical Homilies*), as well as the reconstruction of the greater part of his *Commentary on the Psalms.* This fresh evidence confirms that Theodore was not only the most acute of the Antiochene exegetes, deploying the hermeneutics (critical interpretive principles) of his school in a thoroughly scientific manner, but also an original theologian who, despite dangerous tendencies, made a unique contribution to the advancement of Christology. His *Catechetical Homilies* are immensely valuable both for understanding his ideas and for the light they throw on sacramental doctrine and liturgical practice.

In contrast to Theodore, John was primarily a preacher; indeed he was one of the most accomplished of Christian orators and amply merited his title "Golden-Mouthed" (*Chrysostomos*). With the exception of a few practical treatises and a large dossier of letters, his writings consist entirely of addresses, the majority being expository of the Bible. Here he shows himself a strict exponent of Antiochene literalism, reserved in exploiting even the traditional typology (*i.e.*, treatment of Old Testament events, etc., as prefigurative of the new Christian order) but alert to the moral and pastoral lessons of his texts. This interest, combined with his graphic descriptive powers, makes his sermons a mirror of the social, cultural, and ecclesiastical conditions in contemporary Antioch and Constantinople, as well as of his own compassionate concern as a pastor. Indefatigable in denouncing heresy, he was not an original thinker; on the other hand, he was outstanding as a writer, and connoisseurs of rhetoric have always admired the grace and simplicity of his style in some moods, its splendour and pathos in others.

The last noteworthy Antiochene, Theodoret (*c.* 393–*c.* 458) of Cyrrhus, in Syria, was also an elegant stylist. His writings were encyclopaedic in range, but the most memorable perhaps are his *Remedy for Greek Maladies,* the last and in some respects most effective of ancient apologies against paganism; and his *Ecclesiastical History,* continuing Eusebius' work down to 428. His controversial treatises are also important, for he skillfully defended the Antiochene Christology against the orthodox Bishop Cyril of Alexandria (*c.* 375–444) and was instrumental in getting its more valuable features recognized at the Council of Chalcedon. He was a scholar with a comprehensive and eclectic mind, and his large correspondence testifies to his learning and mastery of Greek prose as well as illustrating the history and intellectual life of the age.

The schools of Edessa and Nisibis. Parallel with its richer and better known Greek and Latin counterparts, an independent Syriac Christian literature flourished inside, and later outside (in Persia), the frontiers of the Roman Empire from the early 4th century onward. Aphraates, an ascetic cleric under whose name 23 treatises written between 336 and 345 have survived, is commonly considered the first Syriac Father. Deeply Christian in tone, these tracts present a primitive theology, with no trace of Hellenistic influence but a firm grasp and skillful use of scripture. But it was Edessa and Nisibis (now Urfa and Nusaybin in southeast Turkey) that were the creative centres of this literature. Edessa indeed had been a focus of Christian culture well before 200; the old Syriac version of the New Testament and Tatian's *Diatessaron,* as well as a mass of Syriac apocryphal writings, probably originated in its neighbourhood.

Syriac Christian literature

The chief glory of Edessene Christianity was Ephraem (*c.* 306–373), the classic writer of the Syrian Church, who established his school of theology there when Nisibis, its original home and his own birthplace, was ceded to Persia under the peace treaty of 363, after the death of Julian the Apostate. In his lifetime Ephraem had a reputation as a brilliant preacher, commentator, controversialist, and above all, sacred poet. His exegesis shows Antiochene tendencies, but as a theologian he championed Nicene orthodoxy and attacked Arianism. His hymns, many in his favourite seven-syllable metre, deal with such themes as the Nativity, the Epiphany, and the Crucifixion, or else are directed against skeptics and heretics. His *Carmina Nisibena* make a valuable source book for historians, especially for information about the frontier wars.

After Ephraem's death in 373, the school at Edessa developed his lively interest in exegesis and became increasingly identified with the Antiochene line in theology. Among those responsible for this was one of its leading instructors, Ibas (d. 457), who worked energetically translating Theodore of Mopsuestia's commentaries and disseminating his Christological views. His own stance on the now urgent Christological issue was akin to that of Theodoret of Cyrrhus—roughly midway between Nestorius' dualism and the Alexandrian doctrine of one nature—and he bluntly criticized Cyril's position in his famous letter to Mari (433), the sole survivor (in a Greek translation) of his abundant works; it was one of the Three Chapters anathematized by the second Council of Constantinople (553).

Influence of the Antiochene school of thought in the Syrian East

The frankly Antiochene posture typified by Ibas brought the school into collision with Rabbula, bishop of Edessa from 412 to 435, an uncompromising supporter of Cyril and the Alexandrian Christology. As well as writing numerous letters, hymns, and a sermon against Nestorius, Rabbula translated Cyril's *De recta fide* (*Concerning the Correct Faith*) into Syriac and also probably compiled the revised Syriac version of the Four Gospels (known as the Peshitta) in order to oust Tatian's *Diatessaron.* On his death he was succeeded by Ibas, who predictably exerted his influence in an Antiochene direction.

Another eminent Edessene writer was Narses (d. *c.* 503), who became one of the formative theologians of the Nestorian Church. He was the author of extensive commentaries, now lost, and of metrical homilies, dialogue songs, and liturgical hymns. In 447, when a Monophysite (a heretical view that Christ had one nature) reaction set in, he was expelled from Edessa along with Barsumas, the head of the school, but they promptly set up a new school at Nisibis on Persian territory. The school at Edessa was finally closed, because of its Nestorian leanings, by the emperor Zeno in 489, but its offshoot at Nisibis flourished for more than 200 years and became the principal seat of Nestorian culture. At one time it had as many as 800 students and was able to ensure that the then prosperous church in Persia was Nestorian. On the other hand, Philoxenus of Mabbug, who had studied at Edessa in the second half of the 5th century and was one of the most learned of Syrian theologians, was a vehement advocate of Monophysitism. His 13 homilies on the Christian life and his letters reveal him as a fine prose writer; but he is chiefly remembered for the revision of the Syriac translation of the Bible (the so-called Philoxenian version) for which he was responsible and which was used by Syrian Monophysites in the 6th century.

The Chalcedonian Fathers. From about 428 onward Christology became an increasingly urgent subject of debate in the East, and excited interest in the West as well. Two broad positions had defined themselves in the 4th century. Among Alexandrian theologians the "Word-flesh" approach was preferred, according to which the Word had assumed human flesh at the incarnation; Christ's possession of a human soul or mind was either denied or ignored. Antiochene theologians, on the other hand, consistently upheld the "Word-man" approach, according to which the Word had united himself to a complete man; this position ran the risk, unless carefully handled, of so separating the divinity and the humanity as to imperil Christ's personal unity.

Apollinarius the Younger (*c.* 310–*c.* 390) had brilliantly exposed the logical implications of the Alexandrian view; although condemned as a heretic, he had forced churchmen of all schools to recognize, though with varying degrees of practical realism, a human mind in the Redeemer. His writings were systematically destroyed, but the remaining fragments confirm his intellectual acuteness as well as his literary skill. The crisis of the 5th century was precipitated by the proclamation by Nestorius (d. *c.* 451), patriarch of Constantinople—pushing Antiochene tendencies to extremes—of a Christology that seemed to many to imply two Sons. Nestorius held that Mary was not only *Theotokos* ("God-bearing") but also *anthropotokos* ("man-bearing"), though he preferred the term *Christotokos* ("Christ-bearing"). In essence, he was attempting to protect the concept of the humanity of Christ. The controversy raged with extraordinary violence from 428 to 451, when the Council of Chalcedon hammered out a formula that at the time seemed acceptable to most and that attempted to do justice to the valuable insights of both traditions.

The rise of the Christological crisis

A number of theologians and ecclesiastics either prepared the way for or contributed to the Chalcedonian solution. Three who deserve mention are Theodoret of Cyrrhus, Proclus of Constantinople, and John Cassian. The first was probably responsible for drafting the Formula of Union (433) that became the basis of the Chalcedonian Definition. Proclus was an outstanding pulpit orator, and several of his sermons as well as seven letters concerned with the controversy have been preserved; he worked indefatigably to reconcile the warring factions. Cassian prepared the West for the controversy by producing in 430, at the request of the deacon (later pope) Leo, a weighty treatise against Nestorius.

But much the most important, not least because they approached the debate from different standpoints, were Cyril (*c.* 375–444), patriarch of Alexandria, and Pope Leo the Great (d. 461). Cyril had been the first to denounce Nestorius, and in a whole series of letters and dogmatic treatises he drove home his critique and expounded his own positive theory of hypostatic (substantive, or essential) union. He secured the condemnation of Nestorius at the Council of Ephesus (431), and his own letters were canonically approved at Chalcedon. A convinced adherent of the Alexandrian Word-flesh Christol-

Contributions of Cyril of Alexandria and Leo the Great

ogy, he deepened his understanding of the problem as the debate progressed; but his preferred expression for the unity of the Redeemer remained "one incarnate nature of the Word," which he mistakenly believed to derive from Athanasius. Leo provided the necessary balance to this with his famous *Dogmatic Letter*, also endorsed at Chalcedon, which affirmed the coexistence of two complete natures, united without confusion, in the one Person of the Incarnate Word, or Christ.

In patristic literature, however, the interest of both Cyril and Leo extends far beyond Christology. Cyril, for example, published lengthy essays on the Trinitarian issue against the Arians, and also enormous commentaries on Old and New Testament books. If the former show little originality, his exegesis marked a reaction against the more fanciful Alexandrian allegorism and a concentration on the strictly typological significance of the text. Leo, for his part, was a notable preacher and one of the greatest of popes. His short, pithy sermons, clear and elegant in style, set a fine model for pulpit oratory in the West; and his numerous letters give an impressive picture of his continuous struggle to promote orthodoxy and the interests of the Roman see.

Non-Chalcedonian Fathers. The Chalcedonian settlement was not achieved without some of the leading participants in the debate that preceded it being branded as heretics because their positions fell outside the limits now accepted as permissible. It also left to subsequent generations a legacy of misunderstanding and division.

5th-century heretical literature

The outstanding personalities in the former category were Nestorius and Eutyches (*c.* 378–after 451). It was Nestorius whose imprudent brandishing of extremist Antiochene theses—particularly his reluctance to grant the title of "Mother of God" (*Theotokos*) to the Blessed Virgin—had touched off the controversy. Only fragments of his works remain, for after his condemnation their destruction was ordered by the Byzantine government, but these have been supplemented by the discovery, in a Syriac translation, of his *Book of Heracleides*. Written late in his life, when Monophysitism had become the bogy, this is a prolix apology in which Nestorius pleads that his own beliefs are identical with those of Leo and the new orthodoxy. Eutyches, on the other hand, an over-enthusiastic follower of Cyril, was led by his antipathy to Nestorianism into the opposite error of confusing the natures. He contended that there was only one nature after the union of divinity and humanity in the Incarnate Word, and he was thus the father of Monophysitism in the strict, and not merely verbal, sense.

After the Council of Ephesus in 431 the eastern bishops of Nestorian sympathies gradually formed a separate Nestorian Church on Persian soil, with the see of its patriarch at Ctesiphon on the Tigris. Edessa and then Nisibis were its theological and literary centres. But a much wider body of eastern Christians, particularly from Egypt and Palestine, found the Chalcedonian dogma of "two natures" a betrayal of the truth as stated by their hero Cyril. For the next two centuries the struggle between these Monophysites and strict Chalcedonians to secure the upper hand convulsed the Eastern Church. Among the Monophysites it produced theologians of high calibre and literary distinction, notably the moderate Severus of Antioch (465–538), who while contending stoutly for "one nature after the union" was equally insistent on the reality of Christ's humanity. His contemporary Julian of Halicarnassus taught the more radical doctrine that, through union with the Word, Christ's body had been incorruptible and immortal from the moment of the incarnation.

In the 7th century, inspired by the need for unity in the face of successive Persian and Arab attacks, an attempt was made to reconcile the Monophysite dissenters with the orthodox Chalcedonians. The formula, which it was thought might prove acceptable to both, asserted that though Christ had two natures, he had only one activity; *i.e.*, one divine will. This doctrine, Monothelitism, stimulated an intense theological controversy but was subjected to profound and far-reaching criticism by Maximus the Confessor (*c.* 580–662), who perceived that, if men are to find in Christ the model for their freedom and individuality, his human nature must be complete and therefore equipped with a human will. The formula was condemned as heretical at the third Council of Constantinople of 680–681.

The post-Nicene Latin Fathers. Latin Christian literature in this period was slower than Greek in getting started, and it always remained sparser. Indeed, the first half of the 4th century produced only Julius Firmicus Maternus, author not only of the most complete treatise on astrology bequeathed by antiquity to the modern world but also of a fierce diatribe against paganism that has the added interest of appealing to the state to employ force to repress it and its immoralities. From Africa, rent asunder by Donatism, the heretical movement that rejected the efficacy of sacraments administered by priests who had denied their faith under persecution, came the measured anti-Donatist polemic of Optatus of Milevis, writing in 366 or 367, whose line of argument anticipates Augustine's later attack against the Donatists.

Much more significant than either, however, was Gaius Marius Victorinus, the brilliant professor whose conversion in 355 caused a sensation at Rome. Obscure but strikingly original in his writings, he was an effective critic of Arianism and sought to present orthodox Trinitarianism in uncompromisingly Neoplatonic terms. His speculations about the inner life of the triune Godhead were to be taken up by Augustine.

Three remarkable figures, all different, dominate the second half of the century. The first, Hilary of Poitiers (*c.* 315–*c.* 367), was a considerable theologian, next to Augustine the finest produced by the West in the patristic epoch. For years he deployed his exceptional gifts in persuading the anti-Arian groups to abandon their traditional catchwords and rally round the Nicene formula, which they had tended to view with suspicion. Often unfairly described as a popularizer of Eastern ideas, he was an original thinker whose scriptural commentaries and perceptive Trinitarian studies brought fresh insights. The second, Ambrose of Milan (*c.* 339–397), was an outstanding ecclesiastical statesman, equally vigilant for orthodoxy against Arianism as for the rights of the church against the state. Both in his dogmatic treatises and in his largely allegorical, pastorally oriented exegetical works he relied heavily on Greek models. One of the pioneers of Catholic moral theology, he also wrote hymns which are still sung in the liturgy.

Significance of Hilary, Ambrose, and Jerome in the Western Church

The third, Jerome (*c.* 347–419/420), was primarily a biblical scholar. His enormous commentaries are erudite but unequal in quality; the earlier ones were greatly influenced by Origen's allegorism, but the ones written later, when he had turned against Origen, were more literalist and historical in their exegesis. Jerome's crowning gift to the Western Church and Western culture was the Vulgate translation of the Bible. Prompted by Pope Damasus (*c.* 304–384), he thoroughly revised the existing Latin versions of the Gospels; the Old Testament he translated afresh from the Hebrew. His historical and polemical writings (the latter full of sarcasm and invective) are all interesting, and his rich correspondence supremely so. As a stylist he wrote with a verve and brilliance unmatched in Latin patristic literature.

The two foremost Christian Latin poets of ancient times, Prudentius (348–after 405) and Paulinus of Nola (*c.* 353–431), also belong to this half-century. Both used the old classical forms with considerable skill, filling them with a fresh Christian spirit. Prudentius' work is both the finer in quality and the more wide-ranging; in his *Battle for the Soul* he introduced an allegorical form that made an enormous appeal to the Middle Ages. Paulinus is also interesting for his extensive correspondence, much admired in his own day, which kept him in close touch with many leading Christian contemporaries.

Augustine as the major theologian of the West

All these figures are overshadowed by the towering genius of Augustine (354–430). The range of his writings was enormous: they comprise profound discussions of Christian doctrine (notably his *De Trinitate*); sustained and carefully argued polemics against heresies (Manichaeism, a dualistic religion; Donatism; and Pelagian-

ism, a view that emphasized man's free will); exegesis, homilies, and ordinary sermons; and a vast collection of letters. His two best known works, the *Confessions* and *The City of God*, broke entirely fresh ground, the one being both an autobiography and an interior colloquy between the soul and God, the other perhaps the most searching study ever made of the theology of history and of the fundamental contrast between Christianity and the world. On almost every issue he handled—the problem of evil, creation, grace and free will, the nature of the church—Augustine opened up lines of thought that men still debate. The prose style he used matched the level of his argument, having a rich texture, subtle assonance, and grave beauty that were new in Latin.

In part recovered in recent years, the works of Pelagius (fl. 405–418) show him to have been a writer and thinker of high quality. Early in the 5th century, when the monasteries of southern Gaul became active intellectual centres, Vincent of Lérins (d. *c.* 450) and John Cassian published critiques of Augustine's extreme positions on grace and free will, proposing the alternative doctrine called Semi-Pelagianism, which held that man by his own free will could desire life with God. This in turn was criticized by able writers like Prosper of Aquitaine (*c.* 390–*c.* 463) and the celebrated preacher Caesarius of Arles (470–542) and was condemned at the Council of Orange (529). Cassian, however, a firsthand student of Eastern monasticism, is chiefly important for his studies of the monastic life, based on material collected in the East. The rules he formulated were freely drawn upon a century later by St. Benedict of Nursia (*c.* 480–547), the reformer of Western monasticism, when Benedict composed his famous and immensely influential rule at Monte Cassino.

The decline of Latin and Greek patristic literature

The 6th century marks the final phase of Latin patristic literature, which includes several notable figures, of whom Boethius (480–524), philosopher and statesman, is the most distinguished. His *De consolatione philosophiae* (Concerning the Consolation of Philosophy) was widely studied in the Middle Ages, but he also composed technically philosophical works, including translations of, and commentaries on, Aristotle. Beside him should be set his longer lived contemporary, Cassiodorus (*c.* 490–*c.* 585), who, as well as encouraging the study of Greek and Latin classics and the copying of manuscripts in monasteries, was himself the author of theological, historical, and encyclopaedic treatises. Also notable is Venantius Fortunatus (*c.* 540–*c.* 600), an accomplished poet whose hymns, such as "Vexilla regis" ("The royal banners forward go") and "Pange lingua" ("Sing, my tongue, the glorious battle"), are still sung. Finally, Gregory the Great (*c.* 540–604) was so prolific and successful an author as to earn the title of Fourth Doctor of the Latin Church. Although unoriginal theologically and reflecting the credulity of the age, his works (which include the earliest life of St. Benedict) made an enormous appeal to the medieval mind.

Later Greek Fathers. The closing phase of patristic literature lasted longer in the Greek East than in the Latin West, where the decline of culture was hastened by barbarian inroads. But even in the East a slackening of effort and originality was becoming perceptible in the latter half of the 5th century. A clear illustration of this is provided by the practice of substituting chain commentaries composed of excerpts from earlier exegetes and anthologies of opinions of respected past theologians for independent exposition and speculation.

Yet the picture was not altogether dim. In the strictly theological field, Leontius of Byzantium (d. *c.* 545) showed ability and originality in reinterpreting the Chalcedonian Christology along the lines of St. Cyril with the aid of the increasingly favoured Aristotelian philosophy. Two other writers, very different from him and from each other, revived in the late 5th and early 6th centuries the brilliance of past generations. One was the figure who called himself Dionysius the Areopagite (*c.* 500), the still unidentified author of theological and mystical treatises that were destined to have an enormous influence. Based on a synthesis of Christian dogma and Neoplatonism, his work exalts the negative theology (God is understood by what he is not) and traces the soul's ascent from a dialectical knowledge of God to mystical union with him. The other is Romanos Melodos (fl. 6th century), greatest hymnist of the Eastern Church, who invented the kontakion, an acrostic verse sermon in many stanzas with a recurring refrain. The sweep, pathos, and grandeur of his compositions give him a high place of honour among religious poets.

With Maximus the Confessor and John of Damascus the end of the patristic epoch is reached. Maximus was a major critic of Monothelitism; he was also a remarkable constructive thinker whose speculative and mystical doctrines were held in unity by his vision of the incarnation as the goal of history. Writing early in the 8th century, John was chiefly influential through his comprehensive presentation of the teaching of the Greek Fathers on the principal Christian doctrines. But he was more than a mere systematizer. In constructing his synthesis he added at many points a finishing touch of his own; his writings in defense of images, prepared to counter the Iconoclasts (those who advocated destruction of religious images, or icons), were original and important; and he was the author of striking poems, some of which found a place in the Greek liturgy.

CONCLUSION

The vitality of patristic literature

Even such a survey as the present, concerned rather to highlight the most significant movements and figures than to provide an exhaustive catalog of writers and their works, brings out the extraordinary vitality of the patristic literature. For four or five hundred years, when secular culture was slowly but steadily in decline, the patristic writers breathed new life into the Greek and Latin languages and created Syriac as a literary medium. Even when the period came to an end, the halt was really only a temporary pause until the impulses behind it could force other outlets. The literature of the later Byzantine Empire looked back to and drew nourishment from the golden centuries of the Fathers, while Latin Christian letters experienced more than one renascence in the Middle Ages.

The range and variety, too, of the literature are impressive. Its overwhelmingly theological concern necessarily imposed understandable but serious limitations, but, when these have been allowed for, the Christian writers must be acknowledged to have been remarkably successful at molding the traditional literary forms to their new purposes and also at improvising fresh ones adapted to their special situations. Aesthetically considered, patristic literature contains much that is mediocre and even shoddy, but also a great deal that by any standards reaches the heights. And it has a unique interest as the creation of an immensely dynamic and far-reachingly important religious movement during the centuries when it could dominate the whole of life and society.

BIBLIOGRAPHY

Texts: The most complete collection of Greek and Latin patristic writings in the original is J.P. MIGNE, *Patrologiae cursus completus*; there is a Greek series with Latin translations, 161 vol. (1857–66), and a Latin series, 221 vol. (1844–55). More modern collections, less complete to date but providing better critical texts, are: for the Greek Fathers, *Die griechischen christlichen Schriftsteller der ersten drei Jahrhunderte* (1897–); for the Latin Fathers, *Corpus scriptorum ecclesiasticorum latinorum* (1866–), and *Corpus Christianorum* (1954–); for both, *Sources chrétiennes* (1941–), with French translations. Translations of Greek works into Oriental languages are published in *Corpus scriptorum christianorum orientalium* (1903–), and in *Patrologia orientalis* (1907–). Many texts are published in critical editions in *Texte und Untersuchungen zur Geschichte der altchristlichen Literatur* (1882–).

Translations: The following series, while consisting only of selections, provide English translations of patristic works: "Library of the Fathers," 43 vol. (1839–81); "The Ante-Nicene Christian Library" (1864–); "Select Library of Nicene and Post-Nicene Fathers of the Christian Church" (1887–92); "Ancient Christian Writers: The Works of the Fathers in Translation" (1946–); and "The Fathers of the Church" (1947–).

General: The most important of the systematically arranged manuals on the patristic literature available in English are: B. ALTANER, *Patrologie*, 7th ed. (1966; Eng. trans., 1960); F.L. CROSS, *The Early Christian Fathers* (1960); and J. QUASTEN, *Patrology*, 3 vol. (1950–60). See also P. DE LABRIOLLE, *Latin Christianity* (Eng. trans. 1924); H. VON CAMPENHAUSEN, *The Fathers of the Greek Church* and *The Fathers of the Latin Church* (Eng. trans., 1963 and 1964).

(J.N.D.K.)

Paul III, Pope

Alessandro Farnese, a scion of one of the great *condottieri* families of Tuscany, reigned as Pope Paul III from October 13, 1534, to November 10, 1549. Known as the "Farnese pope," he was—in his secular life-style, in his promotion of family interest, and in his patronage of the arts—a typical representative of the Renaissance papacy. He nevertheless manifested a true piety and a concern for church reform that culminated in the calling of the Council of Trent in 1545, inaugurating the resurgence of the Roman Catholic Church after the Protestant Reformation.

Paul III, contemporary medallion. In the coin collection of the Vatican Library.
Leonard von Matt—EB Inc.

The Farnese family

Background and early years. Born at Canino, February 29, 1468, Alessandro was the son of Pier Luigi Farnese and Giovannella Gaetani. In service to the papacy since the 12th century, the Farnese family had extended its possessions from a stronghold on Lake Bolsena south and westward to include most of the fiefs between Perugia, Orvieto, Sermoneta, and the sea. In 1417 Ranuccio Farnese (the Elder), one of the most celebrated *condottieri* (mercenary soldiers) of his time, had been made a Roman senator by Pope Martin V. Ranuccio's son Pier Luigi, by marriage with the Gaetani heiress, solidified the Farnese position in the Roman nobility. In 1489, Pier Luigi's daughter Giulia la Bella married Orsino Orsini, a relative of the Spanish cardinal Rodrigo Borgia (Borja), and became a favourite at the papal court. Her brother Bartolommeo became lord of Montalto; her other brother, Alessandro, was destined for the church.

Sensitive and talented, Alessandro Farnese was entrusted to the Humanist Pomponio Leto for his early education and then joined the Medici circle in Florence under Lorenzo the Magnificent. There he was associated with Giovanni de' Medici (the future Pope Leo X) and attended the University of Pisa.

Because of an obscure family quarrel, Alessandro's early sojourn in Rome was interrupted by a short prison term under Pope Innocent VIII. But his career was assured when Cardinal Rodrigo Borgia became his patron. On Rodrigo's election to the papacy (taking the name Alexander VI), he made Alessandro treasurer of the Roman Church, and a year later, on September 20, 1493, created him a cardinal deacon. Gossip traced Alessandro's rapid preferment to the intimacy between his sister Giulia and the Borgia pope, and Alessandro was referred to as the "petticoat cardinal."

Although a prelate, Alessandro did not become an ordained priest until 1519. Meanwhile, he conducted himself like a Renaissance nobleman. Of wide artistic tastes and philosophic interests, he increased his revenues with multiple benefices. He travelled on diplomatic missions, enjoyed the hunt, and delighted in majestic religious and secular ceremonies. Favoured also by Pope Leo X, he used his wealth to enhance his family position and constructed the famous Palazzo Farnese, on the Via Giulia in Rome. Moreover, despite his unfeigned personal piety, the Farnese cardinal kept a wellborn Roman mistress by whom he fathered four children—Pier Luigi, Paolo, Ranuccio, and Costanza. (Later, as Pope Paul III, he provoked serious charges of nepotism by using his papal influence to further the interests of his children and their families, going so far in one celebrated incident as to appoint two of his grandchildren, still in their teens, to the cardinalate.)

Reform of his private life

In 1509 Pope Julius II invested Cardinal Alessandro Farnese with the bishopric of Parma. Selecting Bartolomeo Giudiccioni as his vicar general, the Cardinal took seriously the obligation of governing the diocese and decided to change his private way of life. In May 1512, he served as Julius' legate for the Fifth Lateran Council in Rome; then, having discontinued his liaison with his mistress in 1513, he put the reform decrees of that council into effect in Parma with a visitation in 1516 and, three years later, with a synod. In June 1519 he was ordained a priest and said his first mass on Christmas of that year. Thereafter, his private life was without reproach, and the Cardinal was identified with the reform party in the Roman Curia.

Achievements as pope. The Farnese cardinal's diplomatic skills made him an invaluable aid to the five pontiffs in whose election he participated—Pius III, Julius II, Leo X, Adrian VI, and Clement VII—before he himself emerged as the Roman pontiff on October 13, 1534. At the age of 67, Pope Paul, though apparently frail, was a man of great charm and determination. He was described in diplomatic reports as shrewd and affable, deliberately slow of speech yet loquacious, expressing himself in an elegant Italian or Latin with learned allusions, and scrupulously refraining from tying himself down to a definite "yes" or "no" until the final settlement of an issue—but then able to act with swift, uncompromising dispatch.

Of medium height, spare of figure, with an aquiline nose, ruddy complexion, and aristocratic hands, Paul III was portrayed by Titian in 1543 at age 75 in the full vigour of his pontificate. Two later Titian portraits depict the ravages of age on the pontiff but reveal the depth of intelligence and strength that accompanied him to his last breath at 82.

The pontiff kept himself in good health by frequent excursions in Rome and the countryside, supervising urban projects and fortifications. He encouraged agriculture and provided for new food supplies. His coronation was accompanied by tournaments and pageants, signalling the end of the austerity imposed by the sack of Rome in 1527. In 1536, he authorized the revival of the carnival and rearranged the main thoroughfare in Rome for the visit of the Emperor Charles V, restoring the panoply of traditional ceremonies for the reception of princes and ambassadors. His lavish policies brought prosperity to Rome and the Papal States.

Despite charges of paganism levelled against his pontificate for its secular extravagances—even astrologers were admitted to the papal court—Pope Paul was determined to reform the church. Aware, however, of the setback suffered by Pope Adrian VI's precipitate reform policy a decade earlier, he proceeded, in the face of great internal opposition, with a slow but deliberate call for conversion of the Roman clergy and curia, as well as a reorganization of the papal offices. Immediately upon his election he announced his intention to hold a council and summoned the papal ambassadors Girolamo Aleandro and Pietro Paolo Vergerio from Venice and Vienna, respectively, for consultation about the dangerous state of the church in the north. He then dispatched Vergerio to Austria and Germany on a two-year sojourn to enlist prelates and

The plan for a reform council

princes in the project of holding a council in Mantua or Turin. The Protestants for years had been clamouring for such an assembly on German soil, free of Roman domination. The papacy, however, had feared the calling of a general council would compromise its authority. Paul, however, proceeded with preparations for the council even after it was rejected by Martin Luther and the Protestant leaders.

In a series of consistories, or consultative assemblies, he created cardinals of proven virtue throughout Europe. He also encouraged the foundation of new religious orders and congregations, such as the Theatines, Somaschi, Barnabites, and the Ursuline nuns. Particularly important was his confirmation of the new Jesuit order, which was to provide the papacy with one of its principal instruments in promoting the Counter-Reformation.

Pope Paul's greatest problems were caused by his relations with Emperor Charles V and the French king Francis I, whom he tried to persuade to cease their inveterate wars and turn their forces against the Ottoman Turks, who menaced the coasts of Italy as well as the outposts of Christendom in the East. He encouraged the Emperor to suppress the Lutheran Schmalkaldic League, urged the French king to eliminate the Huguenots, and employed tortuous diplomatic skill to avoid siding with either monarch. In 1538 he journeyed to Nice in an attempt to bring them together. That same year, he excommunicated the English king Henry VIII, who had declared himself head of the English Church. (An earlier sentence of excommunication under Clement VII had been suspended.) Using the military skill of Pier Luigi (his son by his former mistress) and the diplomacy of his grandson Cardinal Alessandro, Paul asserted papal control over central Italy, skillfully avoiding encirclement by both the imperial and French forces.

In May 1536 Pope Paul published a bull of convocation for his proposed council to be held in Mantua. He also authorized a select group of cardinals to draw up a report on the abuses within the church. Guided by Cardinal Gasparo Contarini, this group denounced the ordination of poorly prepared priests, the selection of incompetent bishops, the accumulation of benefices, and the decadence of the religious orders, preaching, and the care of souls. The report, however, fell into Protestant hands and was used by Luther in a violent attack on the Roman Church and the papacy. Nevertheless, the Pope pursued his plans to hold the council, scheduled to open on May 23, 1537, at Mantua. With infinite patience, Paul sought to overcome the opposition of Emperor, kings, prelates, and princes, proroguing and postponing the council's opening again and again over the course of nine years, but finally succeeding in having it inaugurated by his legate, Cardinal Giovanni del Monte, in Trent on December 13, 1545.

In deference to the clamouring of the Protestants, the Emperor insisted that the council confine itself mainly to dealing with discipline and reform. Nevertheless, the Pope's decision that doctrinal matters be given precedence prevailed, and, in its early sessions, the Council of Trent hammered out decrees on the canon of the Scriptures, original sin, justification, and the sacraments, as well as on reform. Fears of the plague and the menace of an attack by armed Protestant forces induced the Pope to accept the council's transfer to Bologna in February 1548. But the Emperor forbade the Spanish and German prelates to go to Bologna, and the Pope had to suspend the Council on September 17, 1549. Nevertheless, this first phase of the Council of Trent had achieved a substantial step forward, leading to a thorough reform of the Church's teaching and discipline.

Throughout his pontificate, Pope Paul frequently visited trouble spots in the Papal States and beyond. He was in Civitavecchia in 1535 and 1537; visited Lucca and Piacenza on his way to Nice in 1538; appeared in Perugia to pacify the city after his forces broke the power of the Colonna family in 1540; and in 1543 he visited Bologna on his way to Busseto to meet the Emperor.

Patronage of the arts

As a patron of the arts, Pope Paul restored the University of Rome, increased the subsidies and importance of the Vatican Library, and showed favour to theologians and canonists but did not neglect the fine arts. He cajoled Michelangelo into finishing the fresco "The Last Judgment" in the Sistine Chapel, decorating the Pauline Chapel, and completing the plans for the construction of the new St. Peter's Basilica. He used Antonio da Sangallo the Younger and a host of architects to renew the fortifications of Rome and the Papal States, continued the construction of the Sala Regia (Royal Hall) in the Vatican, and ordered the reconstruction of the buildings on the Capitoline Hill.

In the midst of grave family, political, and military setbacks, the Pope visited the Quirinal Palace in Rome in early November 1549 and was taken with a raging fever. Clear-minded to the end, he received the last sacraments and died on November 10, in his 82nd year. On his deathbed he is reported to have repented of his nepotism.

Whatever the faults of his early career and the political intrigues of his pontificate, Pope Paul III was remembered by contemporaries as "good hearted, obliging and supremely intelligent . . . worthy to be described as magnanimous." He led the Church out of the decadent splendour of the Renaissance into the austere rejuvenation of the post-Reformation epoch. His grandiose tomb in St. Peter's by Michelangelo's pupil Guglielmo della Porta befits the place he occupies in the church's history.

BIBLIOGRAPHY. LUDWIG PASTOR, *Geschichte der Päpste*, 3rd ed. (1908; Eng. trans., *The History of the Popes from the Close of the Middle Ages*, vol. 11–12, 1923), the most reliable, detailed biography based on archival sources; LEON DOREZ, *La Cour du Pape Paul III*, 2 vol. (1932), a documented account of finances and artistic interests; W.H. EDWARDS, *Paul der Dritte, oder die geistliche Gegenreformation* (1933), a detailed analysis of his religious and spiritual activities; CARLO CAPASSO, *Paolo III, 1534–1549*, 2 vol. (1923–24), an expansive, literary account; PIERRE JANELLE, *The Catholic Reformation*, pp. 53–68 (1949, reprinted 1963), brief, comprehensive, and reliable.

(F.X.M.)

Paul VI, Pope

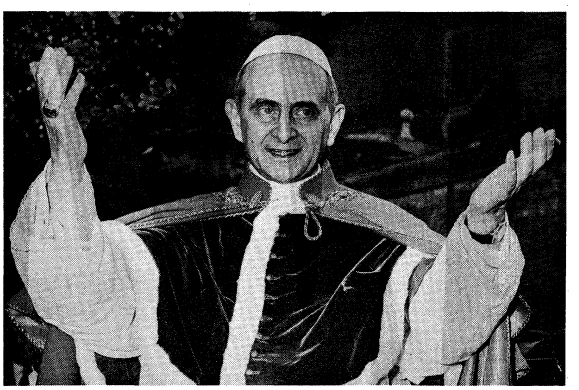
Pope Paul VI (Giovanni Battista Montini), 262nd pope of the Roman Catholic Church, governed through most of the second Vatican Council (1962–65) and in the immediate postconciliar period. His pontificate was confronted with the many problems and uncertainties of a church facing a new position and role in the contemporary world.

Early Vatican career

Early life and career. Born at Concesio (Brescia), September 26, 1897, the son of a middle class lawyer—who was also a journalist and local political figure—and of a mother belonging to the same social background, young Montini was in his early years educated mainly at home because of frail health and later in Brescia. Ordained priest on May 29, 1920, he was sent by his bishop to Rome for higher studies and was eventually recruited for the Vatican diplomatic service. His first assignment, in May 1923, was to the staff of the apostolic nunciature (papal ambassador's post) in Warsaw, but persistent ill-health brought him back to Rome before the end of that same year. He then pursued special studies at the Ecclesiastical Academy, the training school for future Vatican diplomats, and at the same time resumed work at the Vatican Secretariate of State, where he remained in posts of increasing importance for over 30 years.

Leisure moments away from the office were dedicated to the spiritual care of young students from the University of Rome. He thus formed close personal contacts and friendships with many promising young men who later emerged as prominent leaders on the Italian political scene after World War II. Among these were: Giovanni Gronchi, later president of Italy; the future prime ministers of Italy Mario Scelba and Aldo Moro; and future governmental cabinet members. This university student group, under the guidance of their young chaplain Montini, undertook visits to the poor and practical efforts to assist the poor in their needs.

In 1939 Montini was appointed papal undersecretary of state and later, in 1944, acting secretary for ordinary (or nondiplomatic) affairs. He declined an invitation to be



Paul VI.
Wide World Photos

elevated to the Sacred College of Cardinals in 1953. In the beginning of November 1954, Pope Pius XII appointed him archbishop of Milan and Pope John XXIII named him cardinal in 1958. He was elected pope on June 21, 1963, choosing to be known as Paul VI.

Social and ecclesiastical concerns

Vatican II and Paul VI's pontificate. The Montini pontificate began in the period following the difficult first session of the second Vatican Council, in which the new pope had played an important, though not spectacular, part. His lengthy association with university students in the stormy atmosphere of the early days of the Fascist regime in Italy, plus the generally philosophical bent of his mind—developed by a long-standing habit of extensive and reflective reading—enabled him to bring to the perplexing problems of the times an academic understanding, coupled with the knowledge derived from long years of practical diplomatic experience. Paul VI guided the three remaining sessions of the second Vatican Council, often developing points he had first espoused as cardinal archbishop of Milan. His chief concern was that the Roman Catholic Church in the 20th century should be a faithful witness to the tradition of the past, except when tradition was obviously anachronistic.

Upon the completion of the council (December 8, 1965), Paul VI was confronted with the formidable task of implementing its decisions, which affected practically every facet of church life. He approached this task with a sense of the difficulty involved in making changes in centuries-old structures and practices—changes rendered necessary by many rapid transformations in the social, psychological, and political milieu of the 20th century. Montini's approach was consistently one of careful assessment of each concrete situation, with a sharp awareness of the many varied complications that he believed could not be ignored. This prevalently philosophical attitude was often construed by his critics as timidity, indecision, and uncertainty. Nonetheless, many of Paul VI's decisions in these crucial years called for courage. In July 1968, he published his encyclical *Humanae Vitae* ("Of Human Life"), which reaffirmed the stand of several of his predecessors on the long-smoldering controversy over artificial means of birth prevention, which he opposed. In many sectors this encyclical provoked adverse reactions that may be described as the most violent attacks on the authority of papal teaching in modern times. Similarly, his firm stand on the retention of priestly celibacy (*Sacerdotalis Caelibatus*, June 1967) evoked much harsh criticism. Paul VI later likened the large numbers of priests leaving the ministry to a "crown of thorns." He also was disturbed by the growing numbers of religious men and women asking for release from vows or who were abandoning out of hand their religious vows.

From the very outset of his years as pope, Paul VI gave clear evidence of the importance he attached to the study and the solution of social problems and to their impact on world peace. Social questions had already been prominent in his far-reaching pastoral program in Milan (1954–63). During those years he had travelled extensively in the Americas and in Africa, centring his attention mainly on concern for workers and for the poor. While visiting the United States in 1960, the then Cardinal Archbishop of Milan presented to Pres. Dwight D. Eisenhower a statue of an angel shattering the chains of a captive, thus indicating his constant interest in the relief of oppression in any form. Such problems dominated his first encyclical letter *Ecclesiam Suam* ("His Church"), August 6, 1964, and later became the insistent theme of his celebrated *Populorum Progressio* ("Progress of the Peoples"), March 26, 1967. This encyclical was such a pointed plea for social justice that in some conservative circles the Pope was accused of Marxism.

Apostolic journeys. In an address to the council fathers at the end of the first session of the second Vatican Council, the then Cardinal Montini formulated a question that may be called the theme of his pastoral service as pontiff: "Church of Christ, what say you of yourself?" In an effort to answer this fundamental question, Paul VI undertook a series of apostolic journeys that were unparalleled occasions for a pope to set foot on every continent. His first journey was a pilgrimage to the Holy Land (January 1964), highlighted by his historic meeting with the Greek Orthodox patriarch of Constantinople, Athenagoras, in Jerusalem. At the end of that same year he went to India, the first pope to visit Asia. The following year (October 4, 1965) he travelled to the headquarters of the United Nations in New York where he delivered a moving plea for peace to the General Assembly in special session. In 1967 he undertook short visits to Fátima (Portugal) and to Istanbul and Ephesus (Turkey), a journey that had special ecumenical significance: a second meeting with Athenagoras in the patriarch's own episcopal city (Constantinople). In August 1968, the Pope went to Bogotá (Colombia), with a brief stopover in the Bermudas on the return flight to Rome. He appeared before the International Labour Organisation and the World Council of Churches in Geneva (Switzerland), in June 1969. The following month he was in Uganda, East Africa. In the autumn of 1970, he undertook the longest papal journey in modern history: ten days spent in visits to Teheran (Iran), East Pakistan, the Philippines, Western Samoa, Australia, Indonesia, Hong Kong, and Ceylon, each stop bringing Paul VI into personal contact with different peoples of the world. His arrival in Manila almost ended in tragedy when an attempt was made on his life within minutes of his descent from the plane, but with no serious injury.

Papal journeys and concerns

The themes treated by Paul VI on these trips were basically the same: world peace, social justice, world hunger, illiteracy, the brotherhood of man under God, and international cooperation. These subjects reflect the character and the ideals of Paul VI, often called the "pilgrim pope," for whom "peace and brotherhood are synonymous."

Social and ecumenical interests. On January 6, 1971, in the Clementine Hall in the Vatican, Paul VI conferred the Pope John XXIII Peace Prize on the Albanian-born Mother Mary Teresa Bojaxhiu, who had spent most of her life in India, where she had founded a special religious congregation of women dedicated to the alleviation of the countless ills of the poorest classes in the country, gathering the dying off the streets and doing what she could to assist them in every way possible. Paul VI declared on this occasion that the award was intended to centre attention on how even a humble individual without means can further world peace without fanfare, simply by proving in day-to-day action that "every man is my brother," the theme that he himself had chosen for the observance of the 1971 Day for World Peace. Here, as in other instances, Paul's aim was to confront the world at large with the inescapable problems of justice and peace while at the same time proving conclusively that even these apparently insoluble problems can and must be settled with realistic courage and individual perseverance. He was relentlessly anxious to arouse the universal conscience of mankind to the practical solution of the problems involved in the realization of the ideal of charity and justice. As archbishop of Milan, he had been accustomed to go out every Friday afternoon, in memory of the Passion of Christ, to visit some poor invalid or someone tried by suffering, to make his own personal contribution to helping his fellowmen.

Ecumenical concerns

Paul VI's human concern found further expression in his efforts to lessen the long-standing tensions between the church of Rome and other churches and even with those professing no religion at all. He sought out closer understanding with numerous religious leaders throughout the world, both Christian and non-Christian, placing more emphasis on those aspects that unite the churches rather than on those that divide. To show that mutual acquaintance is at the very foundation of any plans or hopes for unity, Pope Paul met with prominent religious leaders from various communities in Great Britain, the United States, and the Soviet Union, as well as other countries. Paul VI also set up a special secretariat for nonbelievers, stressing the need of understanding and endeavouring to solve the problems posed by atheism. Under his guidance the Roman Catholic Church drastically revised its legislation governing marriages between its own members and those who profess other faiths, expressing a firm desire to diminish the threat of human tragedy following possible clashes of individual consciences. For this reason Paul VI's *motu proprio* (a type of papal document) was welcomed and praised for its understanding of human problems and its desire to find a satisfactory solution to the problem of mixed marriages without demanding of either side any renunciation of basic principles of conscience. In the rise of modern ecumenism, Paul VI saw excellent opportunities to encourage world brotherhood, which, he hoped, might enable all men to continue their efforts for human well-being in their pursuit of happiness in unity of faith in God. On May 15, 1971, commemorating the 80th anniversary of Pope Leo XIII's encyclical *Rerum Novarum* on the reform of the social order, Pope Paul issued a forceful apostolic letter, "Octogesima Adveniens" with particular insistence on the necessity of involvement of all men in the solution of the problems of justice and peace. He died at Castel Gandolfo on August 6, 1978.

BIBLIOGRAPHY. J.L. GONZALEZ and T. PEREZ, *Pablo VI* (1964; Eng. trans. by E.L. HESTON, 1964), is a biography essential to an understanding of the man; substantial extracts from his discourses and other public pronouncements are included.

(E.L.H.)

Paul the Apostle, Saint

Paul was a 1st-century Jew who, after being the bitterest enemy of the Christian Church, became its leading missionary and possibly its greatest theologian. His letters, the earliest extant Christian documents, antedate the Gospels of the New Testament. More than half of the Acts of the Apostles deals with his career, and this, together with the letters written by him or in his name, comprises one-third of the New Testament. His efforts and his vision of a world church were responsible for the rapid spread of Christianity and for the speed with which it became a universal religion. None of the followers of Jesus did more than he to establish the patterns of Christian thought and practice.

Sources. For Paul's life there are no reliable sources outside the New Testament. The primary source is his own correspondence, of which Romans, I and II Corinthians, and Galatians are acknowledged to be genuine by all scholars; Philippians, Colossians, I and II Thessalonians, and Philemon by most. About Ephesians, opinion is divided, but it contains little biographical material. The Pastoral Letters (to Timothy and Titus) were written by a disciple of Paul but probably contain Pauline fragments. The letters alone, however, provide no connected story. For that it is necessary to rely on Acts, written some 30 years after Paul's death. Because its evidence sometimes conflicts with that of the letters, some scholars question the historicity of Acts. The general belief is, however, that Acts was written by Paul's companion the evangelist Luke, who drew on his own diary for much of the story.

Life. *Early life.* Paul's birthplace, Tarsus in Cilicia, a district of Asia Minor lying on the main trade route between East and West, was a cosmopolitan university city, which had been the home of famous Stoic philosophers. Paul was proud of his native city and manifested his debt to its Greek culture in his command of idiomatic Greek, in the occasional use of philosophical terms, and in a wealth of metaphors drawn from city life. He was proud, too, of the Roman citizenship inherited from his father; he used his Roman name Paulus in preference to his Jewish name Saul, and he found in the world empire of Rome a model for his later faith in a universal Christian commonwealth. Yet his formal education must have been strictly Jewish. He grew up with a knowledge of Hebrew and under the scrupulous regimen of the Pharisees, a religious and political Jewish party that emphasized moral purity and reinterpreted the Torah, or Law, according to the needs of the time. His subsequent conversion never robbed him of his pride in the ancestral traditions absorbed in childhood.

St. Paul, from the Pisa polyptych by Masaccio, 1426. In the Museo Nazionale di San Matteo, Pisa, Italy. His attributes, the book and the sword, symbolize the word of God (the book), which is the weapon (the sword) employed by the Holy Spirit.
Alinari

The notion that Paul had an unhappy adolescence, tortured by religious doubts and an uneasy conscience, is based on a misunderstanding of the Letter of Paul to the Romans, chapter 7. Though in the guise of autobiography, this chapter is a Christian analysis of religious legalism. Paul's explicit references to his early life are free from any suggestion of inner struggle. He excelled all his contemporaries in his zeal for the Law of Moses, and by its standards his life was blameless. His picture—in Romans, chapter 2, verses 17–20—of the pious Jew exulting in the Law, proud of his God, confident of being guide to the blind, is a self-portrait.

Paul's training as a rabbi

According to Acts, Paul was trained as a rabbi under Gamaliel I, a renowned teacher of the Law, and this is borne out by his frequent use of rabbinic methods of exegesis (*i.e.*, interpretation of a scriptural text) and his knowledge of midrashic legends (*i.e.*, commentaries or explanations of a scriptural text in the form of edifying lessons). Like most rabbis he also learned a trade—tentmaking—by which throughout his missionary career he was regularly to earn his own living. It is unlikely that he ever met Jesus. In Jerusalem, however, he learned enough about Jesus to regard him as a menace to Pharisaic Judaism, for Paul first appears on the scene of history as a persecutor of the Christian Church. In the judgment of Paul the Pharisee, Jesus had broken the Law and taught others to break it and had been justly condemned under the curse that the Old Testament book of Deuteronomy pronounces on lawbreakers.

Conversion. Through a vision on the road to Damascus, Paul became convinced that this crucified Jesus was alive again, vindicated by the Resurrection. Paul came to believe that the curse of crucifixion had been real enough, but it had been borne vicariously. Out of love for men Christ had identified himself with them and now invited them to be identified with him in his new life. In Galatians, he says: Christ "loved me and gave himself for me." For Paul, Christ, not the Law, was the full, final revelation of God's nature and purpose; and all of his passionate devotion was transferred to this new centre. Along with his conversion came his call to be Apostle to the Gentiles and to break down the barrier of prejudice and hostility that the Law had erected between the Jewish people and their neighbours.

Immediately after this experience, Paul withdrew into solitude in Arabia, no doubt to think over the implications of what had happened to him. In later controversy, he insisted that he had received directly from Christ not only his apostleship but his gospel, which declared that a man's standing with God depends solely on his faith in what God has done in Christ and not in any sense on his own effort or achievement. He also referred, however, to traditions received from those who were Christians before him and displayed a full acquaintance with the teaching of Jesus recorded in the Gospels. Much of this he must have learned at Damascus after his return from Arabia.

Visits to Jerusalem and Antioch. Three years later Paul spent two weeks in Jerusalem visiting Peter and James. According to Acts he was forced to leave Damascus by a threat against his life and escaped over the wall of the city in a basket. He himself mentions this escape in II Corinthians, adding that the threat came from a governor appointed by the Nabataean king, Aretas. The difficulty with this account is that Damascus was under Roman rule at least until AD 37, and it is not possible to defer Paul's first visit to Jerusalem as late as that without throwing the chronology of his career into confusion. Possibly the escape belongs to a later, unrecorded visit to Damascus.

From Jerusalem Paul returned to his native Cilicia. It was 14 years later (or possibly 14 years from his conversion) before he was to be in Jerusalem again. If this information from the Letter of Paul to the Galatians is fitted into the framework of Acts, Paul is out of view for ten years, until—according to Acts, chapter 11—the missionary Barnabas finds him and takes him to Antioch. This hiatus can be avoided only by abandoning the order of Acts altogether and assuming that the second visit took place after his missionary work in Asia Minor and Greece. This radical solution, however, creates more problems than it purports to solve and has won little support from biblical scholars. Rather, it seems necessary to imagine Paul engaged in strenuous missionary work in Cilicia and Syria and to assign to this period many of the experiences listed in II Corinthians, chapter 11, verses 23–27, for which no place can be found in the subsequent narrative of Acts.

The first planned Christian missions

Thus far, Christianity had been disseminated beyond Palestine without plan or direction from its leaders. Christians had been scattered by persecution and had carried the gospel with them, preaching at first to fellow Jews but finding a growing response also among Gentiles. The idea of a planned mission first arose at Antioch. The missionaries, Barnabas and Paul, before setting out on their travels, went to Jerusalem, accompanied by Titus. They had a private consultation with the leaders of the Christians there—James, Peter, and John—and reached agreement on future missionary policy. In view of Paul's vehement assertion that nobody else was party to the discussion, this meeting cannot be identified with the public conference of Acts, chapter 15. Acts, chapter 11, mentions an earlier visit of Paul and Barnabas to Jerusalem to deliver a relief fund for a famine that occurred probably in AD 46, and it is best to assume that they took advantage of that occasion to lay their plans before their Jerusalem colleagues.

Shortly after this Peter visited Antioch. One question that had not been discussed at Jerusalem was the relationship between Jewish and Gentile Christians in the new community. Jews were forbidden by their Law to eat with Gentiles. The church at Antioch had been disregarding this rule and holding common meals and Eucharists, until a protest came from James that this practice at Antioch was making things difficult and even dangerous for the church in Jerusalem. The Jewish members, including Peter and Barnabas, accordingly abandoned their liberalism for fear of Jewish reprisals against the mother church. According to Galatians, it was left for Paul to show his mettle by insisting that not only church unity but the very truth of the gospel was at stake.

First missionary journey. Paul and Barnabas, accompanied by Barnabas' cousin John Mark, now set out for Cyprus on a preaching tour, beginning at Salamis, the principal city of the island, and ending at Paphos, the seat of government, where they had an encouraging interview with the governor, Sergius Paulus. They then crossed to the mainland and landed at Perga (near modern Murtana, Turkey). There, to Paul's annoyance, Mark left the party for home. Paul himself was ill, probably with an attack of the recurrent illness that in II Corinthians he called his "thorn in the flesh." He had to change his plans (presumably, in view of his later history, he already had his sights on Ephesus—an ancient Ionian city in western Asia Minor) and go to recuperate in the more healthful uplands of Anatolia, or Asia Minor. This reconstruction of events is known as the "South Galatia" theory—*i.e.*, that Galatians was written to the churches founded during the mission of Acts, chapters 13 and 14—in contrast to a "North Galatia" theory that is no longer held by many scholars.

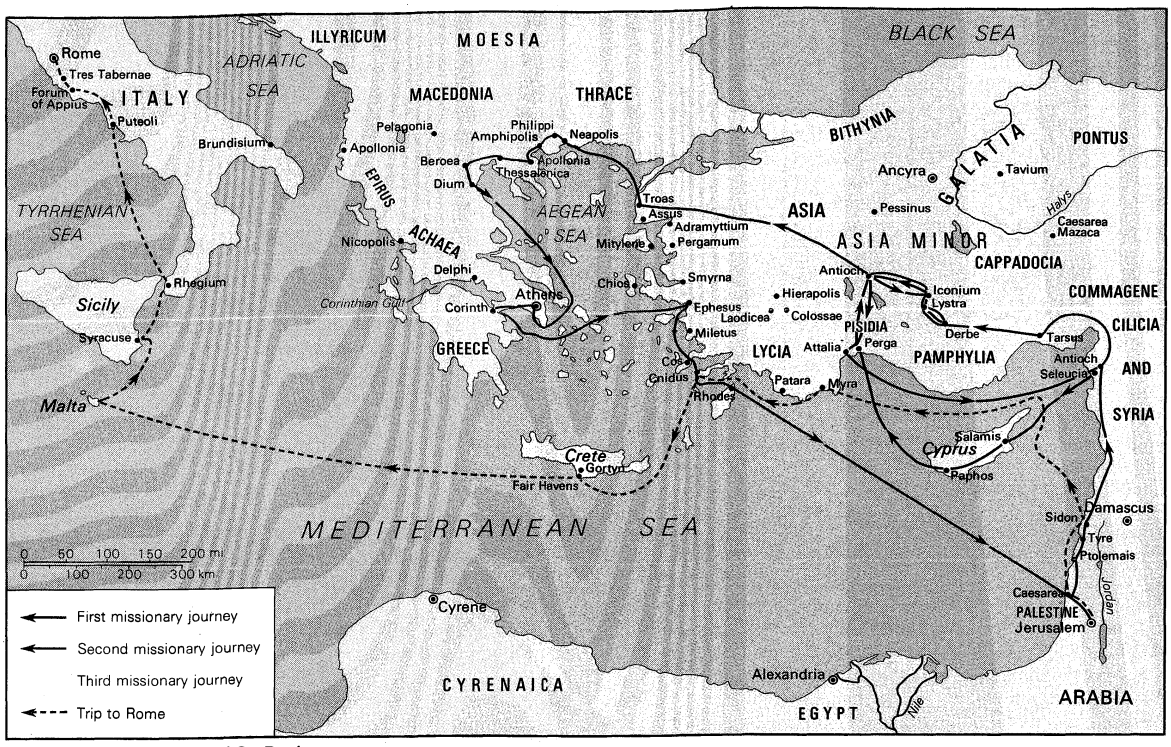
The Galatian mission started in the Roman colony of Pisidian Antioch (near modern Yalvaç, Turkey), where the missionaries preached in the synagogue until Jewish hostility first compelled them to turn to the Gentiles and finally drove them from the city. Their visit to Iconium (modern Konya, Turkey) followed the same pattern. At Lystra (near modern Hatunsaray, Turkey) they were first mistaken for local gods, but subsequently Paul was stoned and left for dead. Yet, after reaching Derbe (the site is disputed), they were able to retrace their steps and revisit the churches they had founded before returning to Syrian Antioch.

The Jerusalem conference

The success of this tour in gaining Gentile converts made a settlement of the dispute over table fellowship more urgent than ever; and so a conference was held in Jerusalem, which issued a letter asking Gentile Christians to relieve Jewish Christians of embarrassment by observing some of the Jewish rules of ceremonial purity. The historicity of this conference, or at least of Paul's participation in it, has been impugned on the grounds that Paul never mentions the apostolic letter or the solution propounded in it. Nevertheless, the probability is that events were moving so fast that this compromise was out-of-date almost as soon as it was promulgated.

Second missionary journey. Paul now proposed a new tour but refused to take Mark. Barnabas stood by his cousin, and the partners separated. Barnabas and Mark returned to Cyprus, and Paul chose as his new colleague Silas, who like himself was a Roman citizen, bearing the Roman name of Silvanus. Together they visited the churches of Syria, Cilicia, and Galatia. At Lystra they were joined by Timothy, a young Christian with a Jewish mother and a Gentile father. At this point it is possible to begin to see emerge Paul's missionary strategy of concentrating on large centres of Roman influence. Aiming again for Ephesus, he was prevented "by the Holy Spirit" from entering the Roman province of Asia, turned north toward the large cities of Bithynia (northwestern Asia Minor), was diverted a second time, and so came to Troas, where he had a vision of a Macedonian asking for help.

So Christianity gained its first foothold in Europe, with the formation of a small church in the Roman colony of Philippi in Macedonia. From the serene letter written to this church at the end of his life, it seems that they never caused him any of the ethical, theological, or disciplinary difficulties he had had to face elsewhere. Instead, they fre-



The missionary journeys of St. Paul.

quently sent money out of slender resources to help with the expenses of his work in other cities and ultimately for his own relief in prison in Rome. After the cure of a psychic slave girl, Paul and Silas were imprisoned on a charge of anti-Roman practices but were released with apologies when they revealed their Roman citizenship and complained of the illegality of their treatment. At Thessalonica (modern Thessaloníki, Greece) they preached for three weeks before being interrupted by a riot and a charge of treasonable adherence to a rival emperor, from which they escaped by discreet departure. Hostile Jews, however, pursued them to the neighbouring town of Beroea and cut short their work there also. Paul went on to Athens, leaving Silas and Timothy to follow later.

Work at Athens and Corinth

At Athens, Paul addressed the council of the Areopagus (formerly the supreme court of Athens; at the time of Paul a body with wider and vaguer powers) and converted one of its members, but nothing is said about a church. The reason appears in the letter that he wrote soon afterwards to Thessalonica. He was deeply concerned about the fate of the recent converts he had left behind there, exposed to persecution and without adequately instructed leadership. He was prevented from going back "by Satan," probably a return of his recurrent illness. When Timothy arrived, Paul, with no thought for his own condition, sent him back for news. He himself went on to Corinth, still convalescent.

In Corinth, Paul lodged with a Christian couple, Aquila and Priscilla, tentmakers like himself. Aquila was a Jew from Pontus (northeastern Asia Minor), but he and his wife had been living in Rome until the previous year, when an edict of the emperor Claudius had expelled all Jews from the capital. These two became lifelong friends of Paul and on one occasion risked their lives for him. Soon Silas and Timothy arrived with the cheering news that the Christians in Thessalonica were loyal and in good heart. The only cause for anxiety was that some of them had misunderstood Paul's preaching about the advent of Christ, taking his language about its imminence with a literalness he never intended. To deal with this situation, he wrote his first letter to them. Some weeks later, when he heard that members of the church had actually given up their daily work to prepare for the end, he wrote his second, somewhat sterner, letter accusing them of truancy.

After a year and a half in Corinth, Paul was arraigned before the new governor Gallio on a charge of practicing an illicit religion. The case was dismissed, and for the author of Acts this episode was the strongest argument in his apology for the Christian religion. For the historian it has an even greater importance, providing the one fixed date in Paul's biography, from which the rest of the chronology has to be worked out, backward and forward. An inscription found at Delphi proves that Gallio's year of office began on July 1, AD 51. Paul must have arrived in Corinth early in 50. "Some time afterwards" (the author of Acts has a tantalizing way of adding vague indications of time to precise ones) Paul left Corinth for Ephesus, Caesarea (on the coast of Palestine), and Jerusalem and so returned to base at Antioch. Of Silas no more is heard until he reappears in the final salutation of the First Letter of Peter.

Third missionary journey. From Antioch Paul set out on a further tour of the Galatian churches, after which he at last succeeded in reaching Ephesus. Here he was to stay for three years, longer than he devoted to any other city. At first he taught in the synagogue, then, when that was closed to him, in a hired lecture hall. Luke was not with him and records in Acts none of the events of this long period except the discrediting of a Jewish exorcist and a dangerous riot started by the silversmiths' guild because Paul's success was having an adverse effect on their market. From Paul's letters, however, some of the gaps can be filled. He gathered around him a team of colleagues who evangelized the surrounding province of Asia. It must have been at this time, for example, that the churches of the Lycus Valley were founded, at Colossae, Hierapolis, and Laodicea. Here, as elsewhere, missionary work had its dangers. Aquila and Priscilla, who seem to have made their home in Ephesus, "risked their necks" for Paul. He speaks, too, in II Corinthians, of a mysterious "trouble" that made him despair of life and, in I Corinthians, of fighting with wild beasts at Ephesus. It is possible that he spent some of the time in prison.

Much of his attention was taken up with the church in Corinth, to which he wrote four letters and paid one visit, unrecorded in Acts. The first letter has not survived, but in reply to it he received a letter that raised questions about various matters of faith and conduct, marriage and divorce, the eating of meat slaughtered in pagan ritual, the right of women to lead public worship, ecstatic speech, and the resurrection of the body. The writers seem to

Letters to the church in Corinth

have been those whom Paul refers to as "the strong party," and they did not simply invite instruction but stated a case and expected him to agree with it. Paul had also received a verbal report of a more disquieting nature from "Chloe's people," who told him of party divisions within the church, litigation between its members, disorderly conduct at the Lord's Supper, or Eucharist, and a case of incest that had gone unrebuked. Paul's answer to all this was the letter known as I Corinthians. Before long the troubles in Corinth were exacerbated by the arrival of newcomers with letters of introduction from another church, who questioned Paul's credentials and undermined his authority. Paul decided to deal with this crisis in person; but his visit was not a success, and he came out of the encounter badly. Back in Ephesus he wrote a third letter (of which II Corinthians, chapters 10–13 may be a part), so severe that he regretted it as soon as Titus, his courier, had left. Titus was to return through Macedonia, and Paul was to meet him at Troas, where he planned to spend some time in evangelism. So great, however, was his anxiety about Corinth that he could not settle there. He went on to Macedonia, where Titus met him with the good news that the severe letter had had its desired effect. All the pent-up feelings of the past weeks were then poured out in the letter known as II Corinthians (or perhaps only chapters 1–9).

Another of Paul's concerns during his stay in Ephesus was the organization of a relief fund for the impoverished church in Jerusalem. Many of his own churches were poor, but he encouraged them to be generous because he saw in this collection a demonstration of unity between Jewish and Gentile churches.

It is notoriously difficult to date Paul's letter to the Galatians; but arguments from language, style, and theology would place it between II Corinthians and Romans, and in that case it must have been written from Macedonia. Acts gives the impression that, after meeting Titus, Paul hurried on to Corinth, stopping only to visit the Macedonian churches in passing. But in Romans Paul claims to have preached the gospel "from Jerusalem as far round as Illyricum"; *i.e.*, the Yugoslav coast. The closing paragraph of the letter to Titus mentions a plan to spend a winter in Nicopolis, at the western end of the Corinthian gulf. If this last passage is one of the genuine Pauline fragments embedded in the Pastoral Letters, these two pieces of evidence together would seem to point to an extended ministry in the northwest part of the Balkan Peninsula.

Letter to the church in Rome

Early in AD 57 Paul paid his last visit to Corinth, and it was probably during the three months he spent there that he wrote his letter to the church in Rome. If chapter 16 of Romans is an integral part of that letter, this origin is certain because that chapter is a letter of introduction for Phoebe, a deacon from Cenchreae, the port of Corinth. Even if that chapter is a separate note addressed to the church of Ephesus, Paul's account of his plans makes it almost certain that he was in Corinth when he wrote. He had not founded the church in Rome (there had been a church there by AD 49 when Aquila and Priscilla were forced to leave); and he wrote this full statement of his faith to prepare the way for a visit, in the hope that the church would then sponsor his projected mission to Spain related in Romans. First, however, he had to go to Jerusalem to accompany the relief fund raised by the Gentile churches.

Paul had intended to travel by ship, but a plot against his life dictated a partial change of plans. He went overland to Troas, stopping for a week in Philippi for the Passover. He then took a passage on a ship that stopped at several ports along the coast and enabled him to meet the Ephesian elders at Miletus (modern Saraylar, Turkey) and take leave of them. After a week's stay with the deacon Philip at Caesarea, he reached Jerusalem.

Arrest and imprisonment. On his last journey to Jerusalem, Paul had been accompanied by representatives of the churches that had contributed to the relief fund. One of these, Trophimus, a Gentile from Ephesus, was recognized by some Asian Jews, who jumped to the conclusion that Paul had taken him into the part of the Temple forbidden to Gentiles. Paul had to be rescued from the ensuing riot by Roman soldiers and escaped further ill-treatment by informing the commanding officer that he was a Roman citizen. While still in protective custody he heard of a plot on his life and informed the commanding officer, who sent him under guard to the governor Felix at Caesarea. Having no evidence for a conviction but unwilling to antagonize the Jewish authorities, Felix kept Paul in prison for two years. When his successor, Festus, arrived, rather than be sent to Jerusalem for trial, Paul appealed to the Emperor.

Paul in Rome

The journey to Rome began in late autumn, and a shipwreck delayed the travellers for three months in Malta, so that they arrived in Rome in the spring of AD 60. There Paul remained for two further years, in house custody awaiting trial. At this point the narrative of Acts comes to an abrupt end, without disclosing the outcome of the trial. As long as the Pastoral Letters were accepted as Pauline, their evidence demanded the hypothesis of an acquittal and second imprisonment. By this theory, Paul must have spent two or three further years in missionary work in Greece, Macedonia, Epirus (northwestern Greece), Asia Minor, and Crete before being arrested again, taken to Rome a second time, and this time sentenced to death. Now that these letters are recognized to be pseudonymous, however, there is no reason to suppose that the verdict at the first trial was favourable. There is good evidence from later authors that Paul died a martyr's death in Rome sometime during the reign of the emperor Nero (AD 54–68).

It is usually assumed, and with good reason, that Paul's remaining letters, Philippians, Colossians, Philemon, and Ephesians (if the last is his), were written from his Roman prison. If so, they provide some evidence to fill out the story of his last years. A slave from Colossae, called Onesimus, had run away from his master Philemon (who happened to be a Christian and a friend of Paul); he came to Rome, met Paul, and was converted by him. Reluctantly, Paul decided he must send him back to his master and wrote a letter to Philemon asking him to receive his former slave as a brother. At about the same time he had another contact with the church in Colossae through its leader Epaphras, who came to consult him about an unorthodox "philosophy" that was gaining some following among Christians at Colossae. To refute this aberrant form of Christianity, Paul wrote his letter to the Colossians. At the same time he wrote a letter that has not survived to the church in Laodicea. If the letter that is known as Ephesians is by him, it was a circular letter to Gentile churches he had not founded or visited, written at the same time as the three letters mentioned above and carried by the same courier, Tychicus.

The church of Philippi, which had sent Paul money in the past to help with the expenses of his missions, now sent one of its members, Epaphroditus, with a further gift. Epaphroditus fell ill on the way and nearly killed himself by completing the journey with the illness upon him. He had intended to stay with Paul and look after him; but Paul, hearing that his friends at home were anxious about his state of health, sent him back and wrote his letter to the Philippians to send with him. The rest is silence, except perhaps for a little farewell note to Timothy, preserved in II Timothy, chapter 4, verses 6–8.

Vivid contrasts of Paul's character

Character. Paul was a man of vivid contrasts. Small and unimpressive in physical stature, he must have had immense stamina to withstand the rigours of travel, beatings, and imprisonments, bouts of intermittent illness, and not least "the anxiety for all the churches" (II Corinthians). Though not a natural orator, he could dominate an audience by incandescence of spiritual power. He was capable both of coarseness and of delicate sensitivity.

Paul's favourite metaphors from athletic stadium, law court, and battlefield reveal him as every inch a competitor, debater, and fighter. He aimed at excellence, whether as a Pharisee or as a servant of Christ. He "gave himself to the Lord" (I Corinthians), determined only to "spend and be spent" (II Corinthians), and to repay "to Greek and barbarian, to wise and foolish" the debt he owed to Christ (Romans). He must always be breaking new

ground (Romans), always pressing on to new discoveries of the mysteries of God (Philippians). Yet he gained the strength that was required for this restless activity from an inner peace that the outside world could not disturb (Philippians).

He could be violent in the assertion of his rugged independence (Galatians), acknowledging no superiors except Christ himself. Yet he had many friends and was never happier than when others shared with him his work, his enthusiasms, and his faith. He calls them fellow workers, fellow soldiers, fellow slaves, yokefellows. In a life devoted to the service of others, he had the grace to accept their service in return and to demand from his colleagues the same high standards that he demanded from himself.

Paul had an eye for detail and would take endless pains over matters of conduct that others thought trivial. Yet he combined with this a comprehensive grasp of the significance of his new faith, believing that it was God's purpose to bring the warring powers of the universe to unity in Christ (Ephesians). It was clear to him that the mighty plan that embraced the whole of history and all nature could yet be reduced to microcosm in God's love for one man.

He could be both stable and volatile, at one time wielding a massive common sense, at another, borne aloft on the wings of ecstasy. He had the gift of tongues (I Corinthians) and of prophecy (Acts) and underwent occasional experiences of vision or trance (II Corinthians). Yet he did not base either his faith or his authority on these but believed that God's activity was most clearly to be seen in normal life and above all in human weakness. He dominates the apostolic age not as a saint or superman but as a normative Christian in whom ordinary human nature was raised to its highest powers. This same contrast characterizes his writing. From humdrum details of conduct he can elicit universal principles and can move in a moment from the prose of argument to the poetry of worship.

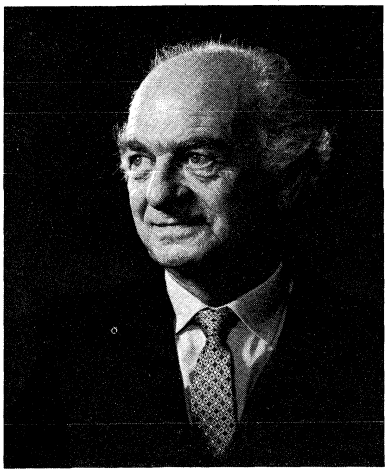
Above all, Paul was a man of God. He saw the hand of God in his own early life, in his conversion, in his apostleship, in his ministry, and even in his illness. His theology is an exposition of the hidden wisdom of God, which had lain behind all history but was now disclosed in Christ. For him, God had been, in Christ, reconciling the world to himself; and man's salvation was God's work from start to finish.

BIBLIOGRAPHY. From the immense bibliography, which includes many commentaries on the Acts and letters, only a small representative selection can be chosen. G.A. DEISSMANN, *St. Paul* (1912); T.R. GLOVER, *Paul of Tarsus* (1925); and J.S. STEWART, *A Man in Christ* (1935), are the most illuminating biographies—the first more social, the second more personal, and the third more theological. H.J. SCHOEPS, *Paul* (1961), tells the story from a radically different point of view. J. KNOX, *Chapters in a Life of Paul* (1954), offers a drastic solution to problems of chronology. W.M. RAMSAY, *St. Paul the Traveller and the Roman Citizen* (1896), is a valuable contribution to historical geography. P.N. HARRISON, *The Problem of the Pastoral Epistles* (1921), is a classic treatment of authorship. W.L. KNOX, *St. Paul and the Church of Jerusalem* (1925); A.M. HUNTER, *Paul and His Predecessors* (1940); and J. MUNCK, *Paulus und die Heilsgeschichte* (1954; Eng. trans., *Paul and the Salvation of Mankind*, 1959), deal with different aspects of Paul's relationship with the other Apostles. ALBERT SCHWEITZER, *The Mysticism of Paul the Apostle* (1931); and W.D. DAVIES, *Paul and Rabbinic Judaism* (1948), are important specialist studies. D.E.H. WHITELEY, *The Theology of St. Paul* (1964), is the most recent comprehensive work on Paul's teaching.

(G.Cai.)

Pauling, Linus

Linus Carl Pauling, American chemist and crusader for peace, achieved scientific renown for his contributions to the understanding of molecular structure, particularly types of chemical bonding (forces holding molecules together) and their relations to the properties of matter and to living organisms. After World War II he became a leader in peace movements, especially in the campaign for nuclear weapons disarmament. He received the Nobel Prize for Chemistry for 1954, "for his research into the nature of the chemical bond and its application to the elucidation of the structure of complex substances," and the Nobel Prize for Peace for 1962, thus becoming the second person to be honoured twice by Nobel awards.



Pauling.

Pauling was born in Portland, Oregon on February 28, 1901, the son of Herman William Pauling, a pharmacist, and his wife Lucy Isabelle (Darling). He received his B.S. in chemical engineering at Oregon State Agricultural College (now Oregon State University), Corvallis, in 1922, then became a graduate assistant at California Institute of Technology, Pasadena, where he took his Ph.D. in physical chemistry in 1925. For two years he was a postdoctoral fellow in Europe, working in the laboratories of such noted scientists as Arnold Sommerfeld in Munich, Niels Bohr in Copenhagen, Erwin Schrödinger in Zürich, and Sir William Henry Bragg in London. He returned to California Institute of Technology as assistant professor of chemistry in 1927, becoming full professor in 1931 and serving as director of the Gates and Crellin Laboratories of Chemistry between 1936 and 1958.

Application of quantum physics to molecular structure

Pauling's chemical work dealt with the many aspects of molecular structure, ranging from simple molecules to proteins. He was among the first to apply the principles of quantum mechanics to the structure of molecules and effectively utilized X-ray diffraction (the alteration of the straight course of X-rays by the interference of an atom or group of atoms), electron diffraction (interference with the course of electrons by atoms), magnetic effects, and the heat involved in forming chemical compounds for the calculation of interatomic distances and the angles between chemical bonds. He was successful in relating the distances and angles between chemical bonds to molecular characteristics and to interaction between molecules.

In order to account for the equivalency of the four bonds around the carbon atom, he introduced the concept of hybrid orbitals, in which electron orbits are moved from their original positions by mutual repulsion. Pauling also recognized the presence of hybrid orbitals in the coordination of ions or groups of ions in a definite geometric arrangement about a central ion. His theory of directed (positive and negative) valence (the capacity of an atom to combine with other atoms) was an outgrowth of his early work, as was the concept of the partial ionic character of covalent bonds—*i.e.*, atoms sharing electrons. His empirical concept of electronegativity, the power of attraction for electrons in a covalent bond, was useful in further clarification of these problems. In the case of compounds the molecules of which cannot be represented unambiguously by a single structure, he introduced the concept of resonance hybrids whereby the true structure of the molecule is regarded as an intermediate state between two or more depictable structures. The resonance theory came under heavy but unsuccessful attack in the U.S.S.R.

in 1951 when doctrinaire scientists of the Communist Party argued that it conflicted with dialectical materialist principles. The ideas on bonding were developed serially in his numerous journal articles during his early career and were consolidated in his book, *The Nature of the Chemical Bond, and the Structure of Molecules and Crystals* (1939), which grew out of lectures he gave in 1937 and 1938. The textbook proved to be one of the most influential of the century.

In 1934 Pauling began to apply his knowledge of molecular structure to the complex molecules of living tissues, particularly in connection with proteins. His studies of the magnetic susceptibility (the ease with which something can be magnetized) of the hemoglobin (the red protein in the red cells of the blood) molecule during oxygenation inaugurated a succession of studies that led to a theory of native (active proteins as found in living organisms), denatured (proteins that through heat or chemical action have broken some of their bonds), and coagulated (solidified) proteins. He became interested in proteins involved in immunological reactions and in 1940, with a German-born biologist, Max Delbrück, developed a concept of molecular complementarity in antibody-antigen reactions (in which the production of antibodies is stimulated in an organism when foreign substances called antigens are introduced). He recognized the importance of hydrogen bonding in protein structure and in interactions between macromolecules (extremely large molecules usually built from repeating groups of smaller molecules). His work with a U.S. chemist, Robert B. Corey, on the structure of amino acids and polypeptides (the chief components of proteins) led him to recognize that certain proteins have structures that resemble a spiral staircase and are called helices.

Late in the 1940s Pauling became interested in sickle-cell anemia when he learned that the red blood corpuscles show their abnormal crescent shape only in venous blood. Intuitively, he reasoned that the cause of the cell deformity must lie in a genetic defect associated with hemoglobin formation. His studies showed that the sickling effect was nullified by the presence of oxygen in the arterial blood.

Pauling also developed a molecular model for the explanation of anesthesia that was made public in 1961, introduced ideas toward the understanding of memory processes, and in 1965 postulated a theory of the atomic nucleus that had certain advantages over other models. His scientific career has been characterized by the application of intuitive hunches aided by a phenomenal memory of chemical facts. Pauling refers to this as the stochastic method (from the Greek "apt to divine the truth by conjecture").

Following the development of nuclear weapons, Pauling became deeply concerned about the possible hazards of exposure to radiation associated with weapons testing. He expressed his view in his book *No More War!* (1958). In January 1958 he brought to the United Nations a petition signed by 11,021 scientists from all over the world urging an end to nuclear weapons tests. In 1963 he left the California Institute of Technology to become a staff member of the Center for the study of Democratic Institutions at Santa Barbara where he largely devoted himself to the study of problems of peace and war. No official reason was given for the award of the Peace Prize for 1962 to Pauling in 1963, but it is widely assumed that he received it for his efforts in behalf of the test ban treaty that was concluded in the same year. His pacifist views estranged him from many scientists with whom he had been closely associated during the years of World War II, when he had served as a civilian with the Office of Scientific Research and Development. Though he was equally opposed to nuclear testing by the United States and the Soviet Union, his loyalty to the U.S. was questioned in some conservative political circles.

Nobel Peace Prize

In 1969 he resigned a position he had held for two years with the University of California, Santa Barbara, in protest against the educational policies of the governor of California and joined the chemistry department of Stanford University, Stanford, California. In addition to winning two Nobel Prizes, an honour that has been conferred on only one other person—Marie Curie, the French physicist—he has been widely honoured in scientific and pacifist circles. He has held guest appointments in many other universities, both at home and abroad. His success as a scientist is based on his capacity for quick insight into new problems, his ability to recognize interrelationships, and the courage to put forward unorthodox ideas. While his concepts have not always been correct, they have always stimulated discussion and investigation.

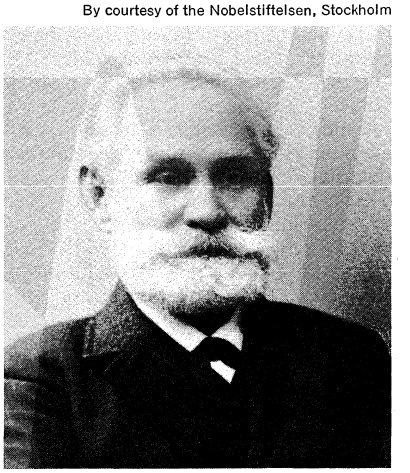
BIBLIOGRAPHY. Pauling's scientific work is stressed by J.H. STURDIVANT in A. RICH and N. DAVIDSON (eds.), *Structural Chemistry and Molecular Biology*, pp. 3-9 (1968). This commemorative volume prepared for Pauling's 65th birthday also has a complete bibliography of his scientific papers and his books. For his Nobel Awards, see *Le prix Nobel en 1954*, pp. 29-32, 66-67 (1955), and the address "Modern Structural Chemistry," pp. 91-99; and *Le prix Nobel en 1963*, pp. 70-82, 113-114 (1964), and the address "Science and Peace," pp. 296-312.

(A.J.I.)

Pavlov, Ivan Petrovich

The Russian physiologist Ivan Petrovich Pavlov is chiefly known for his development of the concept of the conditioned reflex. In a now classic experiment, he trained a hungry dog to salivate at the sound of a bell, which was previously associated with the sight of food. He developed a similar conceptual approach, emphasizing the importance of conditioning, in his pioneering studies relating human behaviour to the nervous system. His Nobel Prize, however, was awarded for his research on the physiology of digestion. After the Bolshevik Revolution and until two years before his death, Pavlov bitterly opposed the Soviet bureaucracy and Marxism.

By courtesy of the Nobelstiftelsen, Stockholm



Pavlov.

Pavlov was born on September 26 (September 14, old style), 1849, the first son of the village priest and the grandson of a peasant, and spent his youth in his native village (Ryazan) in central Russia. He attended a church school and theological seminary, where his seminary teachers impressed him by their devotion to imparting knowledge. In 1870 he abandoned his theological studies to enter the University of St. Petersburg (Leningrad), where he studied chemistry and physiology. After receiving the M.D. at the Imperial Medical Academy in St. Petersburg (graduating in 1879 and completing his dissertation in 1883), he studied during 1884–86 in Germany under the direction of the cardiovascular physiologist Carl Ludwig (in Leipzig), and the gastrointestinal physiologist Rudolf Heidenhain (in Breslau).

Having worked with Ludwig, Pavlov's first independent research was on the physiology of the circulatory system. From 1888–90, in the laboratory of Botkin in St. Petersburg, he investigated cardiac physiology and the regulation of blood pressure.

Experimental physiology

He became so skillful a surgeon that he was able to introduce a catheter into the femoral artery of a dog almost painlessly without anesthesia and to record the influence

on blood pressure of various pharmacological and emotional stimuli. By careful dissection of the fine cardiac nerves he was able to demonstrate the control of the strength of the heartbeat by nerves leaving the cardiac plexus; by stimulating the severed ends of the cervical nerves, he showed the effects of the right and left vagal nerves on the heart.

Pavlov married an attractive pedagogical student in 1881, a friend of the author Fyodor Dostoyevsky, but he was so impoverished that at first they had to live separately. He attributed much of his eventual success to his wife, a domestic, religious, and literary woman, who devoted her life to his comfort and work. In 1890 he became professor of physiology in the Imperial Medical Academy, where he remained until his resignation in 1924. At the newly founded Institute of Experimental Medicine, he initiated precise surgical procedures for animals, with strict attention to their postoperative care and facilities for the maintenance of their health; this excellent care enabled many of them to live for ten to 14 years.

During the years 1890–1900 especially, and to a lesser extent until about 1930, Pavlov studied the secretory activity of digestion. While working with Heidenhain, he had devised an operation to prepare a miniature stomach, or pouch; he isolated the stomach from salivary and pancreatic secretions, while preserving its vagal nerve supply. The surgical procedure enabled him to study the gastrointestinal secretions in a normal animal over its life-span. This work culminated in his book *Lectures on the Work of the Principal Digestive Glands* in 1897. For his advances in the physiology of digestive secretions, Pavlov was awarded the Nobel Prize for Physiology in 1904.

The conditioned reflex

By observing irregularities of secretions in normal unanesthetized animals, Pavlov was led to formulate the laws of the conditioned reflex, a subject that occupied his attention from about 1898 until 1930. He used the salivary secretion as a quantitative measure of the psychical, or subjective, activity of the animal, in order to emphasize the advantage of objective, physiological measures of mental phenomena and higher nervous activity. He sought analogies between the conditional (commonly though incorrectly translated as "conditioned") reflex and the spinal reflex. According to the physiologist Sir Charles Sherrington, the spinal reflex is composed of integrated actions of the nervous system involving such complex components as the excitation and inhibition of many nerves, induction (*i.e.*, the increase or decrease of inhibition brought on by previous excitation), and the irradiation of nerve impulses to many nerve centres. To these components, Pavlov added cortical and subcortical influences, the mosaic action of the brain, the effect of sleep on the spread of inhibition, and the origin of neurotic disturbances principally through a collision, or conflict, between cortical excitation and inhibition.

Beginning about 1930, Pavlov tried to apply his laws to the explanation of human psychoses. He assumed that the excessive inhibition characteristic of a psychotic person was a protective mechanism—shutting out the external world—in that it excluded injurious stimuli that had previously caused extreme excitation. In Russia this idea became the basis for treating psychiatric patients in quiet and nonstimulating external surroundings. During this period Pavlov announced the important principle of the language function in the human as based on long chains of conditioned reflexes involving words. The function of language involves not only words, he held, but an elaboration of generalizations not possible in animals lower than the human.

Opposition to Communism

Pavlov's relationships with the Communists and the Soviet government were unique not only for Russia but also for the history of science. Although he was never a politician, he spoke fearlessly for what he considered the truth. In 1922, during the distressing conditions in the aftermath of the Revolution, he requested permission from Lenin to transfer his laboratory abroad. Lenin denied this request, saying that Russia needed scientists such as Pavlov and that Pavlov should have the same food rations as an honoured Communist. Although it was a period of famine, Pavlov refused: "I will not accept these privileges unless you give them to every one of my collaborators!" In spite of many honours granted him by Soviet officials, he upbraided them openly. After returning from his first visit to the United States in 1923 (the second was in 1929), he publicly denounced Communism, stated that the basis for international Marxism was false, and said that, "For the kind of social experiment that you are making, I would not sacrifice a frog's hind legs!" In 1924, when the sons of priests were expelled from the Military Medical Academy in Leningrad (the former Imperial Medical Academy), he resigned his Chair of Physiology announcing "I also am the son of a priest, and if you expel the others I will go too!" In 1927, distressed that his was the only negative vote in the Academy of Sciences against the newly recommended "Red professors," he wrote to Stalin, protesting that "On account of what you are doing to the Russian intelligentsia—demoralizing, annihilating, depraving them—I am ashamed to be called a Russian!" In the late 1920s, as an anti-Communist gesture, he refused Bukharin, the Communist Commissar of Education, admission to his laboratory, though the laboratory was supported by government funds administered by Bukharin.

During the last two years of his life, Pavlov gradually ceased these excoriations and even stated that he hoped to see the success of the government, at the helm of his country. This change of heart may have been a result of increased government support of science and of his own feelings of patriotism when war with Japan seemed imminent. He was never a Communist, however, nor was he responsible for the technique of brainwashing that has sometimes been ascribed to him.

In personal habits Pavlov was extremely punctual, never missing an appointment, it was claimed, and arriving on time in the laboratory even when there was revolutionary activity on the streets. To a collaborator, who explained his ten-minute delay as a result of the shooting, Pavlov exclaimed, "What difference does a Revolution make when you have experiments to do in the laboratory!" He was a bold, vehement nonconformist both in science and in his personal life; he fiercely took up the cudgel for what he believed regardless of the force of his opposition. Although Pavlov held to scientific agnosticism, he considered true religion beneficial; he said that he envied no one anything except his wife her devout religious faith.

Contributions to science

Pavlov's method of studying the normal, healthy animal in natural conditions made possible his contributions to science. He was able to formulate the idea of the conditioned reflex because of his ability to reduce a complex situation to the simple terms of an experiment. Recognizing that in so doing he omitted the subjective component, he insisted that it was not possible to deal with mental phenomena scientifically except by reducing them to measurable physiological quantities.

Although Pavlov's work laid the basis for the scientific analysis of behaviour, and notwithstanding his stature as a scientist and physiologist, his work was subject to certain limitations. Philosophically, while recognizing the preeminence of the subjective and its independence of scientific methods, he did not, in his enthusiasm for science, clarify or define this separation. Clinically, he accepted uncritically psychiatric views concerning schizophrenia and paranoia, and he adopted such neural concepts as induction and irradiation as valid for higher mental activity. Many psychiatrists now consider his explanations too limited, and some neurophysiologists have taken greater interest in other developments, such as electrophysiology and biochemistry. In contrast to Sherrington, he has had few prominent students outside Russia. His method of working with the normal, healthy, unanesthetized animal over its entire life has not been generally accepted in physiology.

He worked regularly in the laboratory or clinic until his death, from pneumonia, on February 27, 1936, in Leningrad.

BIBLIOGRAPHY. BORIS P. BABKIN, *Pavlov: A Biography* (1949), based on personal and professional knowledge of the author, one of Pavlov's oldest pupils, is the most complete

and reliable account until World War I. In dealing with Pavlov's later life Babkin depends upon a Russian monograph by A.A. Savich, another colleague, and the memoirs of his widow. I.P. PAVLOV, *Lectures on Conditioned Reflexes*, 2 vol., trans. by W.H. GANTT (1929–41), contains a biographical sketch, including personality characteristics, dealing mainly with his career after the Soviet Revolution, based on the translator's six years of professional association with Pavlov. Discussions of Pavlov's scientific achievements and independent attitude to government bureaucracy also by W.H. GANTT are "Physiology Since Pavlov," *New Republic*, 105:728–731 (1941); "Pavlov Through Red-Colored Glasses," *Contemp. Psychol.*, 2:273–274 (1957); and "Ivan Petrovich Pavlov," *Recent Adv. Biol. Psychiat.*, vol. 4 (1962). Gantt's *Russian Medicine* (1937), shows the relation of Pavlov to prominent Russian figures in medicine.

(W.H.G.)

Pavlova, Anna

Only rarely do the great personalities of classical ballet become part of world history; one of the few whose genius was known widely in her lifetime and then did pass into general legend is Anna Pavlova.

Anna Pavlovna Pavlova was of Polish origin but was born in St. Petersburg (now Leningrad). The actual date has been variously quoted; the most authoritative is perhaps the one given in the Russian edition of her husband's biography of her—January 31, 1882.

Culver Pictures

Anna Pavlova.

The place and time could hardly have been better for a child with an innate talent for dancing. Tsarist Russia maintained magnificent imperial schools for the performing arts. Entry was by examination, and, although Pavlova's mother was poor—her father had died when she was two years old—the child was accepted for training at the Imperial School of Ballet at the Mariinsky Theatre (now the Kirov State Theatre) in St. Petersburg in 1891.

Prima ballerina of the Imperial Ballet

Following ballet tradition, Pavlova learned her art from teachers who were themselves great dancers. She graduated to the Imperial Ballet in 1899 and rose steadily through the grades to become prima ballerina in 1906. By this time she had already danced *Giselle* with considerable success.

Almost immediately, in 1907, the pattern of her life began to emerge. That year, with a few other dancers, she went on a tour to Riga, Stockholm, Copenhagen, Berlin, and Prague. She was acclaimed, and another tour took place in 1908. In 1909 the impresario Sergey Diaghilev staged a historic season of Russian ballet in Paris, and Pavlova appeared briefly with the company there and later in London. But her experience of touring with a small group had given her a taste for independence, and she never became part of Diaghilev's closely knit Ballets Russes. Her destiny was not, as was theirs, to innovate but simply to show the beauties of classical ballet throughout the world. While she was still taking leave from the Mariinsky Theatre, she danced in New York and London in 1910, where she was partnered with Mikhail Mordkin.

Once she left the Imperial Ballet in 1913, however, her frontiers were extended. For the rest of her life, with various partners (including Laurent Novikov and Pierre Vladimirov) and companies, she was a wandering missionary for her art, giving a vast number of people their introduction to ballet. Whatever the limitations of the rest of the company, which inevitably was largely a well-trained, dedicated band of young disciples, Pavlova's own performances left those who watched them with a lasting memory of disciplined grace, poetic movement, and incarnate magic. Her quality was, above all, the powerful and elusive one of true glamour.

Pavlova's independent tours, which began in 1914, took her to remote parts of the world. These tours were managed by her husband, Victor Dandré. The repertoire of Anna Pavlova's company was largely conventional. They danced excerpts or adaptations of Mariinsky successes such as *Don Quixote*, *La Fille mal gardée*, *The Fairy Doll*, or *Giselle*, of which she was an outstanding interpreter. The most famous numbers, however, were the succession of ephemeral solos, endowed by her with an inimitable enchantment: *The Dragonfly*, *Californian Poppy*, *Gavotte*, and *Christmas* are names that lingered in the hearts of her audiences, together with her single choreographic endeavour, *Autumn Leaves* (1918).

Famous roles

Pavlova's enthusiasm for ethnic dances was reflected in her programs. Polish, Russian, and Mexican dances were performed. Her visits to India and Japan led her to a serious study of their dance techniques. She compiled these studies into *Oriental Impressions*, collaborating on the Indian scenes with Uday Shankar, later to become one of the greatest performers of Indian dance, and in this way playing an important part in the renaissance of the dance in India.

Since she was the company's raison d'être, the source of its public appeal, and therefore its financial stability, Pavlova's burden was extreme. It was hardly surprising, therefore, that, by the end of her life, her technique was faltering, and she was relying increasingly on her unique qualities of personality. She died on January 23, 1931, at The Hague, in The Netherlands, of pneumonia.

Pavlova's personal life was undramatic apart from occasional professional headlines, as when, in 1911, she quarrelled with Mordkin. She kept her marriage secret for some time, and there were no children; her maternal instincts spent themselves on her company and on a home for Russian refugee orphans, which she founded in Paris in 1920. She loved birds and animals, and her home in London, Ivy House, Hampstead, became famous for the ornamental lake with swans, beside which she was photographed and filmed, recalling her most famous solo, *The Dying Swan*, which the choreographer Michel Fokine had created for her in 1905. These film sequences are among the few extant of her and are included in a compilation called *The Immortal Swan*, together with some extracts from her solos filmed one afternoon in Hollywood, in 1924, by the actor Douglas Fairbanks, Sr.

BIBLIOGRAPHY. VICTOR DANDRE, *Anna Pavlova* (1932), a tribute from her husband written immediately after her death; THEODORE STIER, *With Pavlova Round the World* (1927); and WALFORD HYDEN, *Pavlova* (1931), two accounts of touring with Pavlova written by her musical directors; HARCOURT ALGERANOFF, *My Years with Pavlova* (1957), one of Pavlova's leading dancers writes of his time with her company; A.H. FRANKS (ed.), *Pavlova* (1956), a symposium contributed by people who knew Pavlova, compiled for the 25th anniversary of her death; CYRIL W. BEAUMONT, *Anna Pavlova* (1932); VALERIAN SVETLOV, *Anna Pavlova* (1930); and PAUL D. MAGRIEL (ed.), *Pavlova* (1947), three monographs on her life and work—the last is fully illustrated.

(K.S.W.)

Pedagogy

Traditionally regarded as the art and practice of teaching, pedagogy, because of the growing scientific study of children and the human learning processes and the increasingly numerous analytical studies of educational objectives and curricula, now includes also the notion of the science of teaching. The focus of this article will be on the scientific basis of pedagogy; that is, on principles and ideas supported by objective study.

GENERAL CHARACTERISTICS OF TEACHING

Elements to consider in the teaching situation. In the act of teaching there are two parties (the teacher and the taught) who work together in some program (the subject matter) designed to modify the learners' behaviour and experience in some way. It is necessary to begin, therefore, with observations about the learner, the teacher, and the subject matter and then to consider the significance of group life and the school. It will then be possible to consider the factors and theories involved in modifying a person's behaviour and understanding. These include theories of learning in education, of school and class organization, and of instructional media.

Characteristics of the child and the teacher

A child enters school with little if any attainment in written expression and leaves it capable of learning much from human culture. It was thought originally that this progress was just a matter of learning, memorizing, associating, and practicing. The work of psychologists has revealed, however, that the growth of the pupil's intellectual powers must include a large element of development through different phases, beginning with simple sensorimotor coordination; going on to the beginnings of symbolizing, helped by the growth of language and play; and then on to logical thought, provided the material is concrete; and, finally, in midadolescence, on to the power to examine problems comprehensively, to grasp their formal structure, and to evoke explanation. In his emotional life, the child progresses from direct, immediate, uninhibited reactions to more complex, less direct, and more circumspect responses. The physical growth of the child is so obvious as to need no comment. Any attempt to educate the child intellectually and emotionally and for action must take account of these characteristics. Education must pace development, not follow it and not ignore it. The components in the child's overall educational growth are physical and mental maturation, experience, formal teaching through language, and an urge in the learner to resolve discrepancies, anomalies, and dissonances in his experience.

What is required of a teacher is that he enjoy and be capable of sharing work programs with children, designed to modify their behaviour and experience. This means making relevant experience available to the student at the right time. The teacher must be mature, have humour with a sense of status, be firm yet unruffled, and be sympathetic but not overpersonal. With large classes, the teacher becomes a leader of a group, providing stimulating learning situations.

The subject matter taught also has a marked influence on the total teaching situation. It may be conveniently divided into broad headings of languages, humanities, sciences, mathematics, and arts. Although each group of subjects has something in common with others in terms of the demands it makes on the thinker, each area has also something quite specific in its mode of development. Languages call for verbal learning and production based on oral work, particularly during the early phases. The humanities call for an understanding of cause–effect relations of immediate and remote connections between persons and institutions and man in his environment. The sciences call for induction from experience, though deductive processes are required when the laws of science are formalized into mathematical terms. The humanities and sciences both depend on the ability of the learner to hypothesize. Mathematics calls for the ability to abstract, symbolize, and deduce. An interest in the formal and structural properties of the acts of counting and measuring is fundamental. Arts and literature call for a fairly free opportunity to explore and create.

Characteristics of the school community

A large part of the teacher's role is as a group leader, and the group life of the school and the classroom must influence the teaching situation. Group life shows itself in the dynamic structure of the class—including its manner of reaching group decisions, the hierarchy of its members, the existence of cliques and of isolated individuals —and in its morale and overall response to the school and the rest of the staff. The individual pupil also conducts himself under the influence of the group to which he belongs. His achievements and attitudes are subject to evaluation by the group, leading to support or ostracism, and he sets his standards according to these influences.

In many schools, the range of ages in any class is about one year, and the narrow range makes for some uniformity of subject-matter coverage. But in rural one- and two-teacher schools, groups of children may be heterogeneous by age and ability, and the mode of teaching has to cope with a number of smaller subunits moving along at different rates. The teacher's problem is to coordinate the work of these small, dissimilar groups in such a way that all get attention. Creative free activity has to be practiced by one group while another has more formal instruction from the teacher.

The effect of "streaming," or "tracking"—that is, selecting homogeneous groups by both age and intellectual ability—has promoted much inquiry. The practice evokes extreme opinions, ardent support, and vociferous condemnation. The case for uniformity is that putting a pupil with his intellectual peers makes teaching more effective and learning more acceptable. The case against it draws attention to its bad effects on the morale of those children in the lower streams. This view supports the heterogeneous class on the grounds that the strongest are not overforced, and the weakest gain from sharing with their abler fellows. Experimental evidence on the problem is bewilderingly diverse.

The school community is housed in a physical complex, and the conditions of classrooms, assembly places, and play areas and the existence (or nonexistence) of libraries, laboratories, art-and-craft rooms, and workshops all play their part in the effectiveness of the teaching–learning situation. Severe restrictions may be caused by the absence of library and laboratory services.

The social forces immediately outside the school community also influence the teaching situation. These emanate from home, neighbourhood, and wider social groupings. Teaching is a compact among several groups, including teachers, students, and parents, in the first place, with youth organizations and religious and lay groups playing a secondary role. The overall neighbourhood youth subculture also sets standards and attitudes that a teacher has to take into account in his work.

General objectives of teaching. The classification of the general objectives of teaching in terms of school subject matter is not sufficient to explain the ultimate ends of education. These include, essentially, the promotion of a well-integrated person capable of taking a responsible, active role in society. With such a purpose in mind, one may achieve more insight by choosing a psychological analysis of the objectives into the attainment of intellectual abilities and social insights (cognition), the learning of practical active skills (psychomotor learning), and the development of emotions, attitudes, and values (affective learning).

Development of cognitive, affective, and psychomotor learning

Cognitive growth begins at the level of the infant school, with the acquisition of early language and numerical capabilities, and continues increasingly to dominate education to the secondary and higher levels. But the learner is more than an enlarging reservoir of information. With this acquisition goes a growing power to generalize, abstract, infer, interpret, explain, apply, and create. Cognitive training produces a thinker–observer aware of the modes of thought and judgment making up human intellectual activity. In the final stages, the teacher aims at a thinker, critic, organizer, and creator.

In the development of psychomotor learning, the teacher is concerned with the promotion of coordinated skills and their creative use. Instruction begins with the acts of handwriting and plastic art play, characteristic of earlier

years of schooling. It includes painting, games, workshop skills, and practical science. It has a high prestige value among the pupils themselves and the wider community.

The permeation of emotional learning throughout the whole educative process is not always obvious, in part because very often it is brought about incidentally. Teachers may be self-conscious and self-critical about the deliberate inculcation of emotional responses, which will provide the energy and a mainspring of social life. The acquisition and application of values and attitudes are most marked by the time of adolescence and dominate the general life of the young individual. Theoretical, aesthetic, social, economic, political, hedonistic, and religious values pervade the school curriculum. Literature, art, the humanities, and religious teaching are all directly involved, and the teaching of science and mathematics can bring about a positive attitude toward cognitive and theoretical values.

A person's emotional structure is the pattern of his values and attitudes. Under the influence of instruction and experience, this structure shows three kinds of change. First, the pupil learns to select those situations and problems to which he will make appropriate emotional responses. Second, in general, an increasing range of situations includes happenings more remote from the learner. At first, emotions are aroused by situations directly affecting the child, but as he becomes more mature he is increasingly involved in affairs and causes far removed from his own personal life. Third, his repertoire of emotional responses gradually becomes less immediate, expressive, and linked with physical activity.

The general design of instruction. The scientific analysis of educative processes has led to a more detailed examination of the total act of teaching, which is intended to make the teacher more aware of all that is involved in a piece of instruction.

Foreknowledge about students and objectives. The complete act of teaching involves more than the presentation and development of lesson material. Before he embarks on a fresh stage of instruction, the teacher must be reasonably clear about two things: (1) the capabilities, achievements, strengths and weaknesses, background, and interests of his learners; and (2) the short- and long-term objectives he hopes to achieve in his lesson and series of lessons. These curricular strategies will have to be put into effect in the light of what is known about the students and will result in the actual tactics of the teaching-learning situation.

Educational psychologists give much attention to diagnosing preinstructional achievements, particularly in the basic subjects of language and number, and to measuring intellectual ability in the form of reasoning power. There has been special emphasis on the idea of the student's readiness at various ages to grasp concepts of concrete and formal thought. Numerous agencies produce test material for these purposes, and in many countries the idea has been widely applied to selection for entry to secondary and higher schools; one of the purposes of so-called leaving examinations is to grade students as to their suitability for further stages of education. The teacher himself, however, can provide the most sensitive diagnoses and analyses of preinstructional capacity, and the existence of so much published material in no way diminishes the effectiveness of his responsibility.

The teaching–learning situation. In the actual instruction, a single lesson is usually a part of a longer sequence covering months or more. Each lesson, however, stands to some extent as a self-contained unit within a sequence. In addition, each lesson itself is a complex of smaller teaching–learning–thinking elements. The progress of a lesson may consist of a cycle of smaller units of shorter duration, each consisting of instruction by the teacher and construction by the learner—that is, alternating phases in which first the activity of the teacher and then that of the learner predominates.

The progress of lessons

The lesson or syllabus proper is thus not to be narrowly conceived of as "chalk and talk" instruction. It is better seen as a succession of periods of varying length of instruction by the teacher and of discovery, construction, and problem solving by the pupil. Although the student's own curiosity, experience, and observation are important, so is the cyclic activity of teacher and learner. The teacher selects, arranges, and partially predigests the material to be learned, and this is what is meant by guiding the learner's discovery and construction activity. It is a role the teacher cannot abrogate, and, even in curricula revised to give the learner greater opportunity to discover for himself, there is concealed a large degree of selecting and decision making by the teacher. This is what teaching is about.

Teachers must face the problem of how to maintain curiosity and interest as the chief motivative forces behind the learning. Sustained interest leads the student to set himself realistic standards of achievement. Vital intrinsic motivation may sometimes be supplemented by extrinsic rewards and standards originating from sources other than the student himself, such as examinations and outside incentives, but these latter are better regarded as props to support the attention of the learner and to augment his interest in the subject matter.

Assessment of results. At the end of the lesson proper or of any other unit or program of instruction, the teacher must assess its results before moving to the next cycle of teaching events. Assuming the occurrence of teaching–learning cycles of instruction-construction activity, it follows that there is a built-in process of frequent assessment during the progress of any period of teaching. The results of the small phases of the learner's problem solving provide at the same time both the assessment of past progress and the readiness for further development.

Progress over longer intervals of learning can be measured by more formal tests or examinations within the school or at local administrative level. Postinstructional assessment may have several purposes: to discover when classes or year groups have reached some minimum level of competence, to produce a measure of individual differences, or to diagnose individual learning–thinking difficulties. A wide variety of assessment can be used for this purpose, including the analysis of work produced in the course of learning, continuous assessments by the teachers, essay-type examinations, creative tasks, and objective tests. The content of the assessment material may also vary widely, ranging from that that asks for reproduction of learned material to that that evokes application, generalization, and transfer to new problem situations.

GENERAL THEORIES OF LEARNING IN EDUCATION

Traditional theories. *Mental-discipline theories.* The earliest mental-discipline theories of teaching were based on a premise that the main justification for teaching anything is not for itself but for what it trains—intelligence, attitudes, and values. By choosing the right material and by emphasizing rote methods of learning, according to this theory, one disciplines the mind and produces a better intellect.

In classical times, the ideal product of education was held to be a citizen trained in the disciplined study of a restricted number of subjects—grammar, logic, rhetoric, arithmetic, geometry, music and astronomy. The mode of learning was based on imitation and memorizing, and there was heavy emphasis on the intellectual authority of the teacher, as in the socratic method of question and answer. In later centuries, it was further taken for granted that the study of Greco-Roman literature and philosophy would have a liberalizing effect on the student.

In the hands of the Renaissance Dutch philosopher Erasmus and the Jesuit Fathers, this method of instruction took more sensitive account of the psychological characteristics of young learners. Understanding had to precede learning, and, according to the Jesuits, the teacher's first task was careful preparation of the material to be taught (the prelection). But even with this greater awareness of the learner's needs, the concept of mental discipline still underlay the whole process of instruction. Present-day critics of this classical humanistic approach would challenge the alleged power of mental discipline and the rather exclusive value of Greco-Roman thought.

The theory of "faculty psychology"

The theory of learning involving mental discipline is more commonly associated with Aristotle's "faculty psychology," by which the mind is understood to be composed of a number of faculties, each relatively independent of the others. The principle had its origin in a theory that classified mental and spiritual life in terms of functions of the soul: knowing, feeling, hungering, reasoning, and doing. From the Middle Ages to the early 19th century, the number of recognized faculties grew and included those of judgment, duty, perception, and conception. Since these were associated with certain parts of the cranium by the phrenologists, it was a natural step to assume that learning would consist of the exercise of these "parts," or mental capabilities (though the education of the senses also had a role, in initiating the rational cognitive processes). Certain school subjects were thought to have particular value as agents for exercising certain faculties. Geometry trained the faculty of reason, and history trained the memory. School subjects came to be valuable as much for what faculties they trained as for their own intrinsic worth. This is the learning theory of formal discipline.

Psychological faculties, used as categories, no doubt influenced the study of so-called mental factors. When different cognitive tests are given and the results compared, similarities are found among all the tests and among smaller groups of them. The bases for the similarities are identified as mental factors, including the ideas of intelligence, reasoning, memory, verbal ability, number capacity, and spatial intelligence. The existence of common mental factors underlying different school subjects would support the idea of formal discipline and would lead to the notion of transfer of training, by which exercise in one school subject leads to improvements in learning of another. The transferred elements could be common facts, learning habits, methods of thinking, attitudes, and values. Though much empirical research has been done on transfer of learning, it has yielded mixed results. Some workers hold that transfer has been possible only insofar as there have been identical elements, and even those who claim a transfer of methods generally insist that transfer has little chance of success unless it is actively explained and applied. Learners have to apply methods consciously to the new field in order to succeed. The opposing view would be that each subject is unique and requires its own mode of thought. A more realistic view may be intermediate—namely, that there is both a common and a specific element in each intellectual field, that mental discipline or transfer of training is to some degree possible but only insofar as the similarities and analogies are utilized, that the process is deliberate, and that a residue of specific subject matter remains in each field. This requires specific learning.

Naturalistic theories. A few educational theorists view the education of the child as an unfolding process. The child develops inevitably as a product of nature, and the main function of the teacher is to provide the optimum conditions for this development. This leads to the theory that the child's experience is the essential thing. A Swiss educator, J.H. Pestalozzi, was a leading theorist in this field, and his practical schemes were designed to provide the most appropriate experience for the child's development. In a sense, the modern revival of the potency of experience is an acknowledgement of the developmental element in learning.

Naturalistic emphasis is on creativity

Jean-Jacques Rousseau also started from the assumption that man conforms to nature. Since, more than Pestalozzi, he assumed the certainty of a spontaneous development of powers and faculties, he urged that any form of constraint was to be avoided. Thus it has been held that he saw man as a noble savage growing in isolation in a state of nature. But nature also means a social life. The consequences of Rousseau's basic view have been (1) a reduced emphasis on knowing and greater emphasis on acting and doing, (2) a promotion of positive interests in learning, and (3) an encouragement of the child to depend on his own resources. In their purest form, naturalistic theories are clearly inadequate in the modern world of technology, but their emphasis on spontaneous child activity, as opposed to excessive formal instruction, is a valuable component of the educational process.

Apperception theories. Another theory assumed that human learning consisted essentially of building up associations between different ideas and experiences; the mind, in accordance with the ideas of the 17th-century English philosopher John Locke, was assumed to be at first devoid of ideas. The 19th-century German philosopher Johann Herbart made an important contribution by providing a mental mechanism that determined which ideas would become conscious and which would be left in the subconscious, to be called upon if circumstances warranted it. This was the mechanism of apperception, by which new ideas became associated with existing ideas to form a matrix of association ideas called the apperception mass. New ideas were thus assimilated to the old. A Swiss psychologist, Jean Piaget, argued that such assimilation was not enough, that accommodation of the established ideas to the new experiences was also required.

In any event, ideas such as Herbart's were translated into a sequence of steps presumed to be required to carry out a lesson:

1. Preparation, whereby the teacher starts the lesson with something already known to the class
2. Presentation, introducing new material
3. Association, whereby the new is compared with the old and connected (the stage of apperception)
4. Generalization, whereby the teacher presents other instances of the new idea
5. Application, whereby the ideas are applied to further material, carried out by the child individually (a problem-solving phase)

Though these five steps give the teacher a clear role, they constitute a form of intellectual dominance and could lead to stereotyped lessons restricting the spontaneous creative learning by the pupil. Contemporary curricular revisions, on the contrary, aim at promoting pupil activity.

Conditioning and behaviourist theories. *Classical and operant conditioning.* In the act of classical conditioning, the learner comes to respond to stimuli other than the one originally calling for the response (as when dogs are taught to salivate at the sound of a bell). One says in such a situation that a new stimulus is learned. In the human situation, learning to recognize the name of an object or a foreign word constitutes a simple instance of stimulus learning. Such an event is called sign learning, because, in knowing the sign for something, a person to some extent makes a response to the sign similar to that that he would make to the object itself. Learning new vocabularies, new terms and conventions, or algebraic and chemical symbols all involve some degree of classical conditioning. It is thought probable that one trains the emotions in the same way, for a person may learn to feel pleasure not only when he meets the original situation causing the pleasure but also when he sees some wider context associated with it. This idea is important in school teaching and helps in a general way to explain children's positive and negative feelings toward school, feelings that may have arisen originally from difficulties in learning specific school subjects.

Stimulus learning

Operant, or instrumental, conditioning is so-called because, in making his response, the learner provides the instrument by which a problem is solved. This learning is more important to schoolwork, for teachers are concerned ultimately with drawing forth new responses from their students. Learning is active, and, after the early acquisition of vocabulary, terminology, and rules (by stimulus learning), the learner must use this material in problem-solving responses. By reinforcement (*e.g.*, a reward), both sorts of learning can be combined.

Conditioning theories are not wholly adequate to explain school learning, since the learner is not simply a responder. Intervening between the stimulus and the response is the learner's total conscious structure, made up of the results of experience, previous teaching, attitudes, and his own capacity to comment upon and edit his own response. Simple reinforcement is also inadequate in that

the stimulus and the response are not linked in an exclusive one-to-one basis. Several stimuli may evoke a single response, and several responses may be made to a particular stimulus. These form the behavioral bases for the formation of concepts and transfer effects from one topic to another. The two basic modes of stimulus-response learning provide a ground analysis of school learning, but the complexity of academic achievement calls for much elaboration on the simple model.

Cognitive theories of learning. Cognitive theories are appropriate to the school situation, for they are concerned with knowing and thinking. They assume that perceiving and doing, shown in manipulation and play, precede the capacity to symbolize, which in turn prepares for comprehensive understanding. Although the sequence of motor-perceptual experience followed by symbolic representation has been advocated for a long time, Jean Piaget offered the first penetrating account of this kind of intellectual growth. His views have exercised great influence on educators.

Cognitive theories of learning also assume that the complete act of thought follows a fairly common sequence, as follows: arousal of intellectual interest; preliminary exploration of the problem; formulation of ideas, explanations, or hypotheses; selection of appropriate ideas; and verification of their suitability.

The emphasis of optimum sequence and pacing of learning

Teaching based on cognitive theories of learning recognizes, first, the growth in quality of intellectual activity and capitalizes on this knowledge by organizing instruction to anticipate the next stage in development but does not await it; otherwise there would be no instruction; *i.e.*, instruction should pace development but not outstrip it. Second, it seeks to tune the learning situation to the sequences of the complete act of thought and to arrange, simplify, and organize the subject matter accordingly. Some educators emphasize strongly the arousal phase; in many modern science curricula there is, thus, the idea of inquiry training, which tries to arouse in the child a spontaneous rather than a directed interest. Other educators are concerned more with the middle intellectual phases of the thinking sequence—especially the playing with hypotheses or hunches and the working with organizing ideas and concepts.

Once started, the motivation of cognitive learning depends less on notions of reinforcement and more on standards of intellectual achievement generated by the learner himself. Accordingly, the learner may begin to have aspirations and to set himself future standards that are influenced by his past performances and those of his fellows.

Maturation and readiness theories. Readiness theories of learning lean heavily on the concept of maturation in stages of biological and mental development. It is assumed that a child passes through all stages of development in reaching maturity. The teacher finds out what a child is ready for and then devises appropriate materials and methods. Much of the work on reading skills, for instance, makes use of the readiness concept. The Italian educator Maria Montessori claimed that "periods of sensitivity," corresponding to certain ages, exist when a child's interest and mental capacity are best suited to acquiring knowledge of such things as textures and colours, tidiness, and language.

Insofar as Piaget offered a learning theory, it was based on the idea of readiness. But his approach to development does not overemphasize maturation and readiness, for he pointed out that, after the first few months of life, maturation is marginal in its effects, whereas experience is essential. Development through different intellectual phases, he believed, is necessarily coincident with relevant active experience; readiness is actively promoted, not passively entered, and the teacher must endeavour to be a step ahead of any particular level of readiness.

Structural theories of learning. The second half of the 20th century saw a revival of the concept of the structured wholeness of experience, which Gestalt psychologists had first introduced early in the century. The whole of experience, in this view, is more than the sum of its parts. In educational terms, a new experience—such as a new historical text, an exposition in science, or a problem rider in geometry—begins by seeming relatively formless and unstructured. The learner, who does not yet know his way about the material, begins by seizing upon what appear to him to be important features or figures. He then reformulates the experience in these new terms. The insight gradually becomes more and more structured until finally he reaches an understanding or a solution to the problem. It may be that, in all these processes, the learner may try anything he can think of, usually in a haphazard way.

Learning as an act of structuring and relating

Piaget improved upon Gestalt notions by suggesting a thought structure of a more adaptable nature—one that becomes more differentiated and intuitive with experience. He listed three psychological properties of a structure: wholeness, relationship between parts, and the principle of homeostasis, whereby a mental structure adjusts itself to new experience by assimilation and accommodation. This kind of structuralism found quite independent advocates in other fields. In language, for example, an American, Noam Chomsky, believes that there are innate language structures in the young individual, just as Piaget insists that there are thought structures.

A belief in the structural nature of experience would conceive of the teacher as an encourager, example provider, coanalyzer, and cobuilder of mental structures that originate in the learner in a relatively undifferentiated state. The learner is assumed to be active in forming structures and to be making the best he can of the situation he experiences. The teacher's task is to help and moderate this process of the learner's active construction. This notion works easily and well with able children but entails careful selection with less able students.

Others have also stressed the structural nature of advanced cognitive learning. Each area of human knowledge, in this view, is said to have its own unique structure composed of its concepts and their relationships and its own basic modes of progress. It is suggested that teaching a school subject should not lead to too much tampering with the inherent structural order of the subject but should follow the structure and lines of development of the subject itself. Teaching should not be contrived and artificial. Thus, economics should be taught as an economist views it or physics as a physicist views it or language as a linguist views it. Although such ideas are generally attractive, they have not been widely translated with any success into actual school practice.

GENERAL THEORIES OF EDUCATIONAL ORGANIZATION

Educational organization rests to some extent on psychological views about learning, but explicitly it is concerned with the grouping of pupils for educational experience and instruction.

Grading and streaming

Pupils in general are organized by age into what are usually termed grades, classes, or forms. Each school is also usually either comprehensive (containing students pursuing various academic, commercial, and vocational curricula) or based on the so-called dual plan (containing only students pursuing a particular curriculum). In some countries, this dual system is actually tripartite: there may be schools for classical academic study, schools for technical or vocational study, and schools for more generalized, "modern," diversified study. Whether comprehensive or dual-plan, schools frequently have some kind of streaming or multitracking whereby students are grouped according to ability so that there are separate classes for the less able and the more able (see also ELEMENTARY AND SECONDARY EDUCATION).

Grading and streaming have recently come in for much criticism. There is a rigidity in the two systems that causes some educators uneasiness, particularly since total education is seen as more than achievement in school subjects. Some countries, notably the United States, have made a start in trying to solve this difficulty by introducing the nongraded school, in which grades are abolished and students are placed individually in "phases" for each subject, through which they progress at their own pace. A similar solution has been to ungrade students for certain basic subjects, such as mathematics and native language,

but to have them rejoin their age peers for other school activities. In such systems there is, nevertheless, a kind of grading by intellectual ability, and egalitarians are apt still to be suspicious of them. There is scarcely any clear evidence of the effectiveness of the wholly nongraded system. It would seem probable that the optimum organization may be to combine grading with nongrading. Although this will involve constructing complex timetables, it will also offer the advantages of other, more rigid systems without introducing too many of their disadvantages. For one thing, retaining some grouping by age seems important as a link to extramural activities, in which age peers tend spontaneously to come together.

The modern interest in resources for learning has led to the concepts of general-purpose classrooms, open-plan teaching, and team teaching. The idea of general-purpose classrooms starts from the assumption that the school curriculum can be divided into a few large areas of allied intellectual interests, such as the humanities, languages, and sciences. The total resources available for teaching in each of these areas, including teachers, are then made available in one common teaching space, and ordinary classroom and lesson-period divisions disappear, to be replaced by a real mobility between teachers and learners as they make use of the different resources available, including library and laboratory facilities and various educational hardware (see below *Instructional media*). In the infant and primary schools, similar ideas are introduced in the open-plan system. At both the primary and the secondary levels, however, there is insufficient evidence on the effectiveness of the systems. The attitude and action of the teacher remains the strongest factor, and he may still require some privacy for his teaching.

Team teaching

Team teaching represents an attempt to make better use of every teacher's potential in any subject area, to create a flexible learning situation, and to make nonstreaming more effective. The normal class of 30 pupils with an individual subject teacher is replaced by a larger group of pupils and a team of teachers, who pool their efforts. Although the team plan may take several forms, it generally assumes some variety of the following elements: (1) large-group instruction, in which the total complement of some 50 to 150 students in the program is periodically taught by one teacher (either the same teacher or several teachers in rotation) in a lecture hall; (2) small-group instruction, which alternates with large-group instruction so as to allow small numbers of students and a member of the teaching team to discuss, report, and exchange ideas; (3) independent study, whereby students are given individual projects or library work; and (4) team planning sessions, in which, daily or weekly, the teachers plan, coordinate, report on, and evaluate their programs. The presumed benefits of team teaching are that it makes better use of each teacher's individual interests and strengths; that it avoids unnecessary replication, particularly in such basic subjects as native literature, in which ordinarily several classes led by different teachers cover the same ground; and that teaching in front of one's colleagues is a beneficial practice providing some evaluative feedback. Also, it is said that the less able children do not feel so segregated as in ordinary streamed classes; although they may gain little from the large-group sessions and individual projects, they seem to make real progress in the small seminar groups, without becoming overaware of their more limited capabilities. The reasons for this are obscure. In any event, the most obvious advantage of team teaching is its flexibility, in affording a great variety of possible combinations of student groupings and of educational resources. The major problem is that team teaching cannot be used in all subject areas. Although it may be useful in such areas as the humanities and the social sciences, its provision for lecture-size audiences does not aid the teaching of such subjects as mathematics, in which there are too many individual differences in ability. The same is true of arts and other subjects. Furthermore, without expert leadership, seminars are apt to degenerate into scenes of rather woolly discussions.

The grouping of children by ability, though still practiced, remains a problem. Formal tests are used to separate students according to their ability, and many people feel that separations by such means are neither reliable nor socially desirable. Even with regard to separating the mentally handicapped, there is growing opinion that wherever possible these children should be given basic instruction in special centres and remedial classes in schools for normal children. Handicapped and normal children would thereby share much of their education. Separation of the sexes is also declining in most countries, as the mixing of girls and boys comes to be recognized as healthy and socializing.

INSTRUCTIONAL MEDIA

In general, instructional media are seen by educators as aids rather than substitutions for the teacher. A teacher spends a disproportionate amount of his time in routine chores—in collecting and assigning books and materials and in marking—that could be partly obviated if aids could be so constructed as to free him to concentrate on the central job of promoting understanding, intellectual curiosity, and creative activity in the learner.

Speaking–listening media. In lectures and recordings, the teacher is able to set out his material as he thinks best, but usually the audience reception is weakly passive since there is not much opportunity for a two-way communication of ideas. Furthermore, in lectures, much of the students' energies may be taken up with note writing. This inhibits thinking about the material. Recordings enable one to store lecture material and to use it on occasions when a teacher is not available, but they are rather detached for young learners and seem to evoke better results with older students.

Language laboratories

Language laboratories are study rooms equipped with electronic sound-reproduction devices, enabling students to hear model pronunciations of foreign languages and to record and hear their own voices as they engage in pattern drills. Most laboratories provide a master control board that permits a teacher to listen to and correct any student individually. Many are equipped to use filmstrips or motion pictures simultaneously with the tape recorders. These laboratories are effective modes of operant learning, and, after a minimum vocabulary and syntax have been established, the learning can be converted into a stimulating form of problem solving.

Visual and observational media. Useful visual materials include objects and models, diagrams, charts, graphs, cartoons, and posters; maps, globes, and sand tables for illustrating topographical items; pictures, slides, filmstrips, motion pictures, and television. Facilities include blackboards, bulletin boards, display cases, tables and areas, museums, flannel boards, and electric boards. Such activities as field trips and the use of visiting authorities (usually called resource people) are considered part of visual and observational programs, and even demonstrations, dramatizations, experiments, and creative activities are usually included.

In general, pictures and diagrams, fieldwork, and contrived experiments and observations are all used as concrete leads to the generalizing, abstracting, and explaining that constitutes human learning. To fulfil this function, however, their use must be accompanied by interpretation by an adult mind.

The teacher must offer careful elaboration and discussion, for children's and adolescents' powers to interpret and infer often go astray and thus must be carefully guided. Visual material by itself may even be a hindrance; a scattering of pretty pictures through a history text, for example, does not necessarily produce a better understanding of history. Similar difficulties are inherent in fieldwork—geographical, biological, archaeological, and geological. What is observed rarely gives the whole story and, in the case of archaeological and geological fieldwork, provides an incomplete picture of the past. The teacher must fill in the gaps or somehow lead his students to do so.

Reading–writing media. Reading and writing have formed the staple of traditional education. This assumes sophisticated language attainments and the capacity to think formally and respond to another mind, for a text-

book is essentially a mode of communication between a remote teacher and a reader. The material in a textbook is a sample of a subject area, simplified to a level suitable for the reader. Because the sampling in both the text and the exercise might be haphazard, and there can be no feedback to the writer, the teacher has to take on the writer's responsibilities.

Programmed instruction

Programmed learning is a newer form of reading and writing. The most basic form of programmed instruction —called linear programming—analyzes a subject into its component parts and arranges the parts in sequential learning order. At each step in his reading, the student is required to make a response and is told immediately whether or not the response is correct. The program is usually structured so that right answers are apt to be extremely frequent (perhaps 95 percent of the time)—in order, so the theory goes, to encourage the student and give him a feeling of success. In another kind of programmed instruction—called branching programming—the student is given a piece of information, provided with alternative answers to questions, and, on the basis of his decision, detoured, if necessary, to remedial study or sent on to the next section of the program. The two types of program differ fundamentally in their attitudes toward errors and the use of them. The brancher uses them to further the learning; the linearist avoids them. The chief value of programmed instruction in general is that it allows a student to learn at his own pace, without much teacher supervision. Its chief defect is that it can quickly become dull and mechanical for the student.

Computer-based instruction. The large storage and calculating capacities of the computer suggest great potential for its use in the classroom. It can give instructions to the learner, call for responses, feed back the results, and modify his further learning accordingly. The computer can also be used to measure each student's attainments, compare them with past performances, and then advise teachers on what parts of the curriculum they should follow next.

In a fully computer-assisted instruction program, the computer takes over from the teacher in providing the learner with drill, practice, and revision, as well as testing and diagnosis. The form of the teaching may be simply linear or branching, or it can be extended to thinking and problem solving by simulation. The limitations at the moment centre on the learner's responses, which are limited to a prescribed set of multiple choices. Free, creative responses, which one associates with the best of classroom situations, cannot yet be accommodated.

BIBLIOGRAPHY. B.S. BLOOM (ed.), *Taxonomy of Educational Objectives: The Classification of Educational Goals*, Handbook 1, *Cognitive Domain* (1956), a comprehensive analysis of the cognitive objectives of education—*i.e.*, those concerned with understanding and interpreting human knowledge; J.S. BRUBACHER, *A History of the Problems of Education*, 2nd ed. (1966), a survey of all aspects of education, particularly valuable for locating information about the theories of education from Greco-Roman times to the present; J.S. BRUNER, *Toward a Theory of Instruction* (1966), Bruner's attempt to translate his ideas about cognitive development into a theory of school instruction, it should be read along with Bruner's earlier *Process of Education* (1960); J.H. FLAVELL, *The Developmental Psychology of Jean Piaget* (1963), a constructive and useful analysis of the entire work, assumptions, methods, and results of the research of Jean Piaget and his coworkers; E.R. HILGARD and G.H. BOWER, *Theories of Learning*, 3rd ed. (1966), a standard work on learning theories, strong not only on behaviourist theories in which the authors deal with every nuance but also on instinct theories and those of the Gestalt, cognitive, and functional schools; R.R. RUSK, *The Doctrines of the Great Educators*, 3rd ed. (1965), a clear account of the contributions to educational theory by leading exponents from early times; L.C. TAYLOR, *Resources for Learning* (1971), a popular discursive comment on the resources for learning, including the role of the teacher, the organization of learning, the educational hardware, and various other modes of learning classes and schools.

(E.A.P.)

Pediments

A pediment is a piedmont plain (one lying near a highland) formed by the bevelling of bedrock that may be mantled with alluvium (unconsolidated sediments) or exposed. The pediment surface generally makes an abrupt angle with a mountain mass or backing hillslope and pediment embayments may extend into the uplands along valleys. The term was first applied by W.J. McGee in 1897, when writing of the Sonoran Desert, and was adopted for general usage in 1922 by Kirk Bryan, a U.S. geomorphologist, who described landforms in Arizona.

Comparable forms occur in humid areas but pediments are characteristic of arid lands, where they accentuate the separateness of island-like ranges and surrounding plains. Well-developed pediments give a high-set appearance to desert uplands. Although the dispersed surface drainage and shallow detrital cover do not favour local water supplies nor irrigation agriculture, smooth, firm pediment surfaces are readily traversed by vehicles and provide good access across the important piedmont tracts.

This article treats the physical characteristics of pediment surfaces and some views on the origin of these features, which is still a matter of some dispute. For discussion of related forms, see ALLUVIAL FANS and DESERTS. For information on processes involved in pediment formation, see FLUVIAL PROCESSES; WEATHERING; and DURICRUSTS. See also LANDFORM EVOLUTION for discussion of the Davisian cycle of erosion, equilibrium of process and form, and other ideas on the origins of landforms.

FORM OF PEDIMENTS

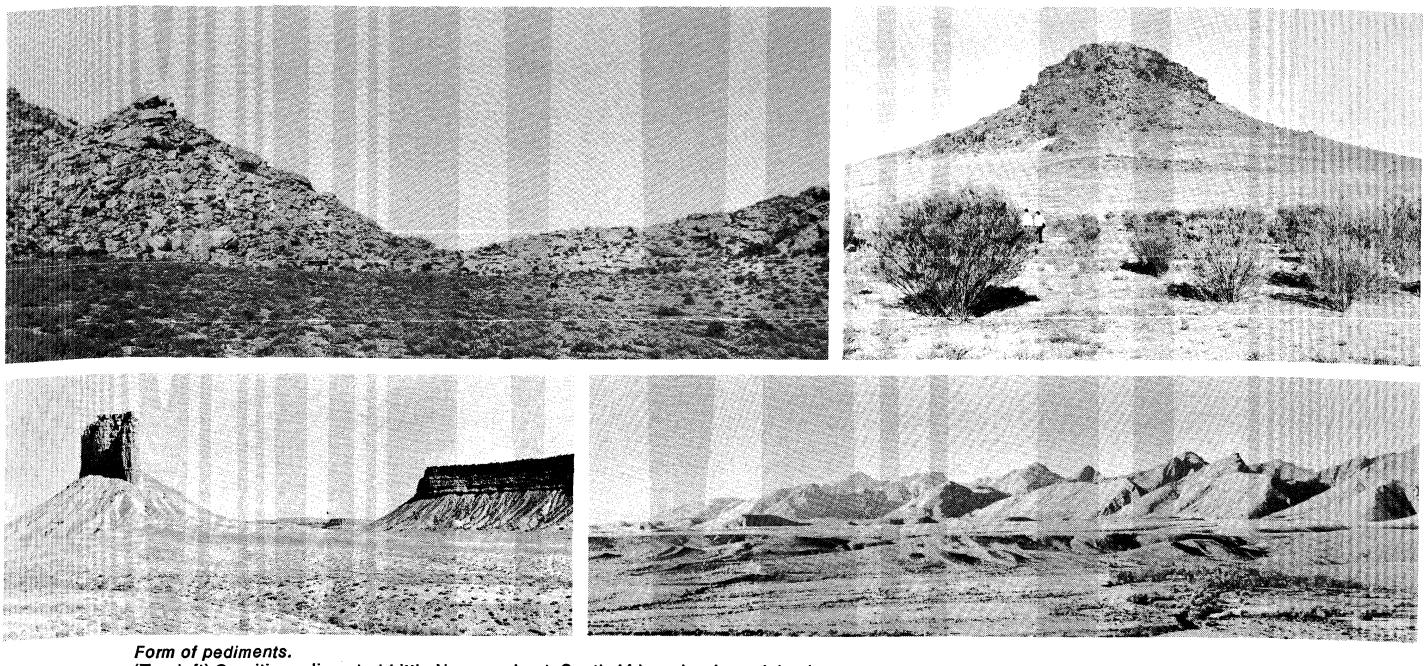
The longitudinal profile of pediments is typically concave; the gradient decreases with distance downslope and approximates a logarithmic curve. Pediment gradients range between 1 in 250 and 1 in 6, and often the smoothness of the slope belies the appreciable steepness of its upper part. Convexity in the lower sector can result from incision of trunk drainage, whereas local convexity in the upper part has been attributed to regrading at a late stage. The uppermost sector may be interrupted by rock outcrops and by small outliers of the upland.

Cross-gradients of undissected pediments are generally small and surfaces are at most smoothly and broadly undulating, with low rises between drainage tracts. Some upper sectors of pediments, however, exhibit as much as 15 metres transverse relief, and gradients are as steep as 1 in 10.

Pediment profile and rock type

The pediment profile is influenced directly by the underlying rock and indirectly by the calibre of detritus yielded from the backing hillslope. The most widespread and best developed pediments are on granite and similar massive rocks; these break down from joint blocks to sand and grit with few intermediate-sized particles. Such pediments have moderate slopes of limited concavity, which persist to the head of the pediment where there is an angular junction with the hill. A mantle of grit and sand usually is present, broken by minor outcrops and tors that are not necessarily concentrated towards the head. Pediments in mixed rocks, and in soft fine-grained rocks with a hard capping, exhibit a smooth and marked concavity. The slope steepens rapidly near the head and passes into the hillslope through a short transitional sector. Such profiles are commonly associated with a mantle of stone derived from the hill and the particle size decreases progressively downslope. Steepening at the head of the pediment is linked with the appearance of coarser colluvial debris. There are few interruptions of profile by rock outcrops, and these generally occur near the head of the pediment. This distinction between granitic and nongranitic pediments is fundamental.

Other differences of gradient express contrasts in erosional and transportive balances and in the stage of pediment evolution. Pediments at the foot of high mountains that yield much coarse debris are generally steeper than those flanking uplands of reduced relief. There is an inverse relation between pediment length and average slope, which suggests the occurrence of regrading during pediment extension. In Arizona, bare rock pediments are steeper than those that are mantled. Pediment slopes in intercanyon sectors are steeper than those along drainage lines, and those along minor channels are steeper than those flanking major channels.



Form of pediments.
(Top left) Granitic pediment at Little Namaqualand, South Africa, showing subdued concavity of profile, angular piedmont junction, and thin mantle without stone. (Top right) Middle Pinnacle, a nongranitic pediment on mixed rocks near Broken Hill, New South Wales, showing markedly concave profile into the hillslope, abundant outcrop, and widespread stone mantle. (Bottom left) Soft rock pediments cut in Mancos shale with a flat capping of Mesa Verde sandstone near Mesa Verde National Park, Colorado. Note the short transitional concavity at the heads and the shallowly dissected lower pediment slopes. (Bottom right) Dissected pediments in the Flinders Range, South Australia, which have a gravel mantle up to ten feet (three metres) thick, across shales and carbonate rocks; steeply dipping limestones and sandstones form the ranges.
By courtesy of (top left) J.A. Mabbutt, (top right, bottom left) T. Langford-Smith, (bottom right) C.R. Twidale

It is claimed that pediment slopes are gentler where aridity is more marked, due to sparser vegetation, and that the slope break against the hill is accentuated. An extreme arid form may be the pinnacle landscapes of the Saharan Tibesti. In this area, hills of horizontally bedded sandstone tower with sheer sides above thinly veneered rock plains, and pediment slopes are absent or limited to a narrow fringe.

Pediment slopes on homogeneous rock tend to be radial in plan view, as on an alluvial cone, and there are fine map examples of arcuate parallel contours with a regular increase in spacing downslope. Where such ideal pediments narrow upslope to an apex in an upland re-entrant, they constitute rock fans. Pediments have been linked to such forms but they are relatively rare. Pediment outlines are variable and are dependent on structural controls. A pediment may open out on softer shale and then narrow to a rock bench where the drainage transverses a resistant sandstone, for example. Some pediment embayments taper gradually upslope whereas many others are blunt-headed. The upper limits of granitic pediments are characteristically regular, with few stream indentations, but the surfaces are smooth and well planed below spurs, in unindented sectors, and in stream embayments.

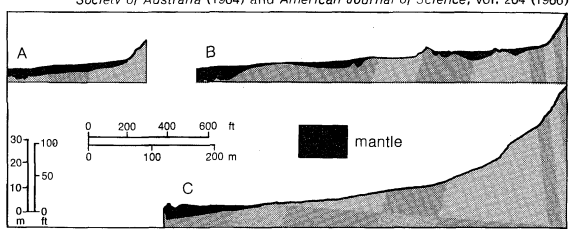
SURFACE FEATURES OF PEDIMENTS

The surfaces first described as pediments were veneered with a metre or so of overburden, such as might be considered to be in transit across the slope. Later workers, particularly those who accept a fluvial origin for pediments, have argued that mantles tens of metres thick can represent the depth of channel scour during floods. In extreme cases, surfaces with a cover of 60 metres have been described as pediments, as along the east front of the Rocky Mountains in Colorado. All agree, however, that the pediment should owe its form predominantly to erosion and that the diagnostic feature is the rock floor.

The pediment mantle

The mantle may occur throughout the slope, as on many granitic pediments, or the rock floor may be exposed near the hill foot. Where depositional basins occur, as in the southwestern United States, the mantle thickens away from the hill base and the mantled pediment passes into an alluvial apron; the term peripediment is used for a locally intervening sector cut across basin fill.

From T. Langford-Smith and G.H. Dury, "A pediment survey at Middle Pinnacle, near Broken Hill, New South Wales," *Journal of the Geological Society of Australia* (1964) and *American Journal of Science*, vol. 264 (1966)



Pediment mantles; mountain front at right.
(A) Complete mantle cover, extending from mountains to lower end of pediment. (B) Mantle filling in irregular pediment surface. (C) Stripped pediment surface with mantle near lower end.

Pediment mantles include a range of materials that reflect the various depositional environments and the source rocks. In interstream sectors the cover tends to be poorly sorted and crudely layered because it is carried from the hillslope and deposited by wash processes, but the mantle becomes finer grained and better sorted downslope. It may include fine aeolian material. Such deposits, termed pedisediments, may grade laterally into typical alluvium near stream channels. On nongranitic pediments there is a mixture of stone from the hill with finer, wash-derived material, and the stone tends to accumulate in a protective surface layer; the mantle on granitic pediments consists chiefly of sand and grit, with stone restricted to the hill margins.

The mantle commonly is weathered, an indication of relatively stable conditions following deposition. Clay enrichment and well-developed soil profiles are common, as are calcareous and gypseous crusts. These duricrusts may protect pediment slopes from dissection.

The upper level of a mantle, or depositional cover, controls pedimentation by restricting downcutting on the ex-

posed slopes above. Rapid accumulation of debris results in burial of the rock plane and rapid removal leads to its dissection. A stable depositional margin allows extension of the subaerial pediment as the hillslope retreats, whereas gentle lowering of the mantle, as in the later stages of diminished relief and sediment yield, favours the exhumation of a suballuvial pediment. Ideal conditions for pediment formation therefore include a subsiding or slowly accumulating basin fill with interior drainage, or a stable or slowly lowering surface adjusted to outgoing drainage, but pedimentation in fact occurs over a wide range of baselevel situations.

Surficial drainage on pediments

There are three characteristic types of surface drainage on pediments. Shallow rills linked in close distributary meshes and slightly impressed in mantle material occupy the upper sectors and tributary embayments of mantled granitic pediments. They head at the hillslope base, where they may connect with slope runnels. Such a rill zone may extend a few hundred metres downslope, depending upon the prominence of the mountain mass, and it generally narrows as the upland is diminished by erosion. The rills transport and deposit material derived from the hillslope.

Small linked channels incised a metre or so into the bedrock or mantle of the pediment form subparallel, slightly anastomosing (braided) systems. On granitic pediments they head at the lower margins of the rill zone and progressively encroach on it; on nongranitic pediments they can occupy the entire surface. They may feed downslope into a depositional plain or into the larger channels described below, and their essential role appears to be the transportation of debris and erosion.

Larger individual channels also occur and they characteristically exhibit shallow trench profiles and the rather straight courses of bedload streams. These channels have sandy floors and loose sandy banks, and have a tendency to braid. On granitic pediments they arise mainly from the confluence of minor channels on the pediment itself; on nongranitic pediments they more commonly form extensions of upland drainage.

MULTIPLE PEDIMENTS

A change from planation to incision in the piedmont can result in the dissection of pediments. Such is the origin of many flat-topped spurs, mesas, and benches with thin cappings of mantle material, as in strike lowlands of the Flinders Ranges of South Australia, where they stand up to 50 metres above the base plains of today. Staged or multiple pediments form spectacular benchlands below some regional escarpments, along the east front of the Colorado Rockies and in some dry intermont basins further west, for example. The succession there has been attributed to uplift and to Pleistocene climatic changes, with pediments planed in drier interglacials and mantled and incised in the wetter glacial periods. In a subhumid nonglaciated environment with intermittent uplift, as in the Canberra region of the Southern Tablelands of southeastern Australia, staged pediments have been described, the lowest of them passing into strath terraces of the incised Murrumbidgee drainage.

PEDIMENT EXTENSION AND PEDIPLANATION

Many desert mountains show evidence of reduction and disintegration by the encroachment of pediments. Comparison of forms indicates that pediment embayments extend headward and eventually link as pediment passes across divides, breaking the upland into isolated smaller masses. Such passes extend the pediment slopes but are generally smoothly convex across the crest. The elimination of an upland watershed should result in a pediment dome with rounded crest and radially concave flanks, as described from the Mojave Desert of California and from Namaqualand in South Africa.

Pediment-like slopes are not restricted to piedmont settings, however, and need not necessarily have resulted from the destruction of upland masses. They have been described along valley sides in the High Plains of eastern Colorado and Kansas, for example.

The term pediplain was introduced for erosional surfaces that formed through the coalescence of pediments, or that were presumed to have formed by the processes of pedimentation. The term has since been extended to describe erosional plainlands of continental extent that are identified with pediments by their multiconcave profiles, absence of deep mantling or weathering, and the persistence of steep hillslopes that are independent of the stage of planation achieved. Such planation has been termed pediplanation, as in southern Africa, and the inference is made that it involves the elimination of relief by retreat of hillslopes, and the extension of piedmont plains by processes seen now to operate on pediments. Pediplanation is held to contrast with the downwearing, rounding, and mantling processes of peneplanation, as envisaged by W.M. Davis, an American geographer.

PEDIMENT-LIKE FORMS IN HUMID REGIONS

Although features accepted as characteristic of pediments attain their best development in deserts, bevelled rock surfaces also occur in humid regions. In humid areas the surfaces are more commonly mantled with deposits and soils and are obscured to greater degree by vegetation. Davis considered rock basements beneath marginal valley-floor strips to be homologous with pediments and he used this similarity to support his view of arid terrain as a variant of the humid model. The differences cited, namely less steep and less rectilinear backing hillslopes, thicker overburden, more deeply weathered bedrock, and lower surface gradients adjusted to finer weathering products, were held to be of degree only. Davis described the Piedmont below the Appalachian Blue Ridge as the humid equivalent of a pediment.

L.C. King, a South African geologist, stressed the uniformitarian nature of hillslopes and maintained that pediments, although best exemplified on hard rocks and in semi-arid environments, appear with varying prominence in all landscapes. He cited the scarp-foot vales of the English Lowlands as pediment forms. J.T. Hack described gravel-capped benches of the Virginia Piedmont as dissected pediments and claimed that the control of piedmont slopes was lithologic rather than climatic, with gentler gradients on shales and carbonate rocks, and steeper slopes on sandstone and quartzite. Hack stressed the similarity of forms in Virginia to those on comparable rocks in the Henry Mountains of Utah, where the climate is much drier.

THEORIES OF PEDIMENT DEVELOPMENT

There are two components in the development of a pediment: the trimming back of the hill and the grading of the rock slope below. The existence of a pediment does not necessarily mean that the hill face has retreated across its length. Where an upland margin accords with a steeply inclined break in rock hardness, it is clear that downwearing at the discontinuity has predominated over slope retreat. Similarly, in granitic terrain, where the uplands are commonly compartmented by steep joints, the piedmont may correspond with a less massive and more readily degraded rock. Most soft-rock pediments, however, are formed on flattish structures with hard rock cappings, such that retreat can occur without changing the lithology of the slope. With these qualifications, it is possible to distinguish between theories of pediment formation that stress backweathering of the hillslope, areal erosion by diffuse runoff, or planation and trimming by laterally migrating stream channels.

Backweathering. Backweathering of the hillslope, with concommitant extension of a rock bench below, was stressed by A.C. Lawson as the dominant factor in pedimentation. The processes invoked in slope retreat were subaerial weathering and the removal of detrius by hillwash and mass movement. The hillslope was predicted to assume and maintain a constant angle during retreat, which was controlled by structure or by the angle of rest of the dominant size of slope debris. The junction of the bench and hill foot, which may have originated as a fault scarp, was determined by the growing detrital apron supplied from above. This apron was held to protect the underlying rock from further degradation. Lawson postulated that the apron must advance more slowly with in-

creasing extent, so restricting its protective role that the rock bench would flatten proportionally and a convex rock profile would result. He claimed that this suballuvial surface would be exposed as a subaerial bench or pediment at a later stage, with decreasing detritus from the diminishing upland. Sheetfloods were held responsible for the transport of debris across the apron and for the ultimate exposure of the rock bench.

Lawson's theory seems applicable in southwest Arizona, where there has been widespread stripping of mantle from rock surfaces, but there is considerable evidence that it by no means offers a universal explanation. Suballuvial profiles, for example, show a range of forms; indentation of the upland front locally at drainage outlets implies a significant role for water action; and soft-rock pediments, which evince the best evidence of slope retreat, are characteristically unmantled in their upper sectors at all stages.

Planation by diffuse runoff. This view received its impetus through the linking of W.J. McGee's early description of pediment landforms with an account of a sheetflood for which an erosive role was claimed. The paucity of subsequent observations suggests that sheetfloods are rare and localized, however, and the commonest form of overland flow on the upper parts of pediments is probably in the many small rills. Rillwash is claimed to be capable of removing and transporting weathered rock with a planatory effect, and combined with weathering, it is also held responsible for the removal of debris from the hillslope and for occasional undercutting at the base.

Lateral planation. The work of migrating stream channels was stressed by Douglas Johnson, who maintained that a zone of predominant transport and planation should exist between the zones of degradation in the upland and deposition in the lowland basin. Swinging of a stream channel about its outlet should form a rock fan similar to an alluvial cone, but such features have rarely been recorded. In granitic terrain, where the piedmont angle is most abrupt and the pediment at its smoothest, lateral planation is particularly weak; hillslopes are rarely undercut, pediment embayments are blunt-headed rather than apical, the slope-pediment contrast is strongest in the straight interstream sectors, and there is restricted channelling in the critical uppermost parts that are mantled with unsorted material. Even on soft-rock pediments traversed by large channels there is little indentation of the hill front, and major outlets are marked by alluvial fans rather than by rock fans. Furthermore, because desert channels are remarkably straight in piedmont sectors, there is restricted scope for planation in the intervening stretches.

Composite explanations. Many would argue that all these processes can act together, with rillwash and backweathering predominant in granitic terrain and stream planation of greater importance on nongranitic and particularly on soft-rock pediments, as attested by the alluvial character of their mantles. Further, the distinction between rillwash and lateral planation by minor channels is easily overdrawn because erosion by the shifting, interbranching, slightly incised rills, which characterize upper pediment sectors, tends to produce some degree of planation.

Control of pedimentation by the mantle. Important problems remain unanswered, particularly the relationship between the mantle and the rock floor, the role of the mantle in pedimentation, and the significance of buried or stripped rock surfaces with respect to changes in régime. The widespread stripped and dissected pediments of southeast Arizona have prompted a suggestion that all the regional pediments are exhumed suballuvial benches, their exposure resulting from a general swing from sheetflooding to channel flow, perhaps due to earth movement. Similarly, pediment mantles in the Rockies have been interpreted as due to a change from drier interglacial to more humid outwash conditions, or to shifts of drainage that bring coarser debris to fine-grained rock environments. Constant invocation of changed environmental conditions to reconcile field evidence becomes unconvincing, however, particularly in areas such as central Australia, where different parts of a single pediment may show signs of active stripping and of burial.

In an arid environment the mantle is not merely protective, for it enhances weathering by retaining moisture beneath the surface. This is particularly true of the piedmont sector, where runoff from the hillslope may cause notching of the hill base and rotting of protrusions of a buried rock floor. A mantle thus forms a weathering level; episodes of stripping will carry away weakened rock and this will lead to planation. Further, a partial mantle provides a smoothly graded surface to which protruding ribs of weathered granite or schist can be trimmed by rillwash.

Such mantle-controlled processes are most important on granitic pediments, where the rock is physically resistant above ground but susceptible to chemical weathering below, and where piedmont lowering is a function of long-continued downwearing under stable conditions rather than of hillslope retreat.

BIBLIOGRAPHY. G.K. GILBERT, *Report on the Geology of the Henry Mountains, Utah*, pp. 120–127 (1877), the first identification of the pediment landform; B.A. TATOR, "Pediment Characteristics and Terminology," *Ann. Ass. Am. Geogr.*, 42:295–317 (1952) and 43:47–53 (1953), a useful summary collating many accounts, mainly from North America; W.J. MCGEE, "Sheetflood Erosion," *Bull. Geogr. Soc. Am.*, 8:87–112 (1897), and K. BRYAN, "Erosion and Sedimentation in the Papago Country, Arizona," *Bull. U.S. Geol. Surv. 730*, pp. 19–90 (1922), classic accounts in which the general significance of the pediment landform was established, the former emphasizing the role of sheetfloods, the latter rillwash and backwearing; A.C. LAWSON, "Epigene Profiles of the Desert," *Univ. Calif. Publs. Geol. Sci.*, 9:23–48 (1915), on the processes of backwearing of the hillslope in pediment formation; TUAN YI-FU, "Pediments in Southeastern Arizona," *ibid.*, vol. 13 (1959), on stripped pediments as a result of climatic change; J.A. MABBUTT, "Mantle-controlled Planation of Pediments," *Am. J. Sci.*, 264:78–91 (1966), on the role of pediment mantles in planation; W.M. DAVIS, "Rock Floors in Arid and Humid Climates," *J. Geol.*, 38:1–26, 136–158 (1930), and J.T. HACK, "Interpretation of Erosional Topography in Humid Temperate Regions," *Am. J. Sci.*, Bradley vol., 258A:80–97 (1960), comparisons of pediment forms in arid and humid settings; L.C. KING, "The Uniformitarian Nature of Hillslopes," *Trans. Edinb. Geol. Soc.*, 17:81–102 (1957), a discussion of the claim that pediments and pedimentation are worldwide, despite modification by climate and rock type.

(J.A.M.)

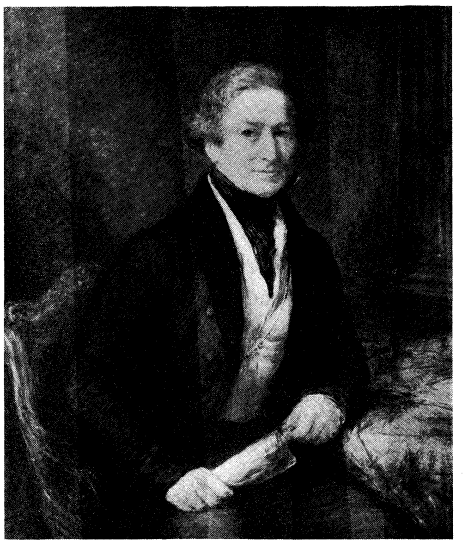
Peel, Sir Robert

British prime minister, founder of the Conservative Party, and the man responsible for the repeal, in 1846, of the Corn Laws that, by imposing heavy tariffs on imported grain, kept farm prices high in England, Sir Robert Peel was one of the outstanding English statesmen of the 19th century. He was born on February 5, 1788, at Bury in Lancashire, the eldest son of a wealthy cotton manufacturer, who was made a baronet by William Pitt the Younger. Robert was educated at Harrow and at Oxford, and, with his father's money, a parliamentary seat was found for him as soon as he came of age, in 1809.

Early political career

As an able young government supporter, Peel received appointment as undersecretary for war and colonies in 1810. Two years later he accepted the difficult post of chief secretary for Ireland. There he made his reputation as an able and incorruptible administrator and at the end of his Irish secretaryship was marked out for early promotion. He had also distinguished himself as the ablest of the "Protestant" party that resisted the admittance of Roman Catholics to Parliament, and in 1817 he gained the coveted honour of election as member of Parliament for Oxford University. Though declining immediate office after his return from Ireland, he was made chairman, in 1819, of the important currency commission that brought about a return to the gold standard.

In the 1822 ministerial reconstruction pursued by Robert Banks Jenkinson, 2nd earl of Liverpool, Peel accepted the post of secretary of state for the home department and a seat in the Cabinet. His first task was to meet the long-standing demands in Parliament for a radical reform of the criminal laws. He then proceeded to a comprehensive reorganization of the criminal code. Between 1825 and



Peel, oil painting by John Linnell, 1838. In the National Portrait Gallery, London.
By courtesy of the National Portrait Gallery, London

1830 he effected its fundamental consolidation and reform, covering three-quarters of all criminal offenses. The rising statistics of crime convinced him that legal reform should be accompanied by improved methods of crime prevention. In 1829 he carried through the Metropolitan Police Act that set up the first disciplined police force for the Greater London area.

When George Canning succeeded Liverpool as prime minister in 1827, Peel resigned on the issue of Catholic emancipation. He returned to office under Arthur Wellesley, 1st duke of Wellington, early in 1828 as home secretary and leader of the House of Commons. Differences with Wellington led to the resignation of several followers of Canning after only four months in office, thus considerably weakening the government. This was followed by the Catholic crisis of 1828–29 that grew out of the renewal of the Irish movement for emancipation in 1823 with the formation of the Catholic Association. Its growing strength culminated in the victory of Daniel O'Connell, the Irish "Liberator," at a by-election for County Clare in 1828. Convinced that further resistance was useless, Peel proffered his resignation and urged the Prime Minister to make a final settlement of the Catholic question. Faced with severe opposition from the King and the Anglican Church, Wellington persuaded Peel, in 1829, to remain in office and assist in carrying through the policy of concession to the Catholics on which they now both agreed. Peel was bitterly attacked for his sudden change of heart and lost his seat for Oxford. Wellington's declaration against parliamentary reform in November 1830 finally brought about the fall of his weak and unpopular government.

Prime minister and Conservative leader

In the reform crisis following Wellington's fall, Peel's position was difficult and ambiguous. Though not opposed to moderate parliamentary reform, he was shocked by the sweeping measure introduced by the ministry of Charles Grey, 2nd Earl Grey, in March 1831. But, on the other hand, he made no effort to conciliate the ultra-Tories, and his refusal to form a new ministry with Wellington and pass a Tory reform bill in 1832 further weakened his standing with his former followers. Yet he was already looking to the growth of conservative opinion in the country, and his moderation in the first reformed Parliament (1833–34) did much to restore his political stature. The premature dismissal of the ministry of William Lamb, 2nd Viscount Melbourne, in November 1834 and Peel's appointment as prime minister gave him an impossible task because he did not possess, and the Conservative Party lacked the organization and men necessary to procure, a majority in the House of Commons (even though the general election of 1835 added considerably to their numbers). Nevertheless, his Tamworth Manifesto was an epoch-making statement of the new Conservative reform principles, and for the first time the party came under his acknowledged leadership. In April 1835, defeated by a combination of Whigs, radicals, and Irish nationalists, he resigned his office. During the next six years, aided by his astute and cautious tactics, the Conservative Party steadily increased in numbers and confidence. Following the general election of 1841, in which he gained a majority of over 70 in the House of Commons, Peel formed an administration that was one of the most memorable of the century.

Peel was faced with war in China and Afghanistan, strained relations with France and the United States, severe commercial distress at home, agitation by the workingmen's reform movement of the Chartists and the Anti-Corn Law League, O'Connell's campaign for the repeal of the union of Ireland and Great Britain, and a five-year accumulation of budgetary deficits. His policy aimed at peace and security abroad, a reduction in the cost of living for the working classes, and encouragement to trade and industry. On the controversial issue of the Corn Laws, to which the landed interest in his party was very sensitive, he brought forward as one of his first measures a new bill drastically reducing the scale of protective duties. In the same year, the bold reintroduction of the income tax (originally instituted during the Napoleonic Wars) established the internal revenue on a sound footing and enabled him to make sweeping reductions of duties on food and raw materials entering the country. The Bank Charter Act of 1844, establishing a tight connection between note issue and gold reserves, completed the foundations of the Victorian banking and currency system. The success of these measures encouraged Peel to launch a second great free-trade budget in 1845. The income tax was renewed, and there was another even more massive round of tariff reductions.

With the return of prosperity, Chartism died down and the Anti-Corn Law League turned to more constitutional methods of agitation. Abroad, a firm but conciliatory policy led to better relations with France. The boundary disputes with the United States were settled by the mission of Alexander Baring, 1st Baron Ashburton, in 1842 and the Oregon treaty of 1846. The same combination of firmness and conciliation was followed in Ireland. Once the threatening campaign for repeal of the union was brought to a halt in 1843 with O'Connell's trial for conspiracy, Peel turned to more constructive measures. A commission was set up to inquire into the relations between landlord and tenant, and a wide scheme for Irish university education was passed into law in 1845.

The liberality of Peel's Irish policy, especially the greatly increased grant to the Catholic seminary of Maynooth, aroused strong Protestant feeling in England, severely straining relations with his own party. The potato disease in 1845, bringing with it the certainty of widespread famine in Ireland, completed the breach. Peel had already come to the conviction that the Corn Laws would have to be abolished sooner or later. His decision in the autumn of 1845 that a relief program for Ireland must be accompanied by the repeal of the Corn Laws split his Cabinet and led to his resignation. But, when Lord John Russell failed to form a free-trade ministry in December 1845, Peel returned to office. After savage parliamentary debates the repeal of the Corn Laws was finally carried through in June 1846. Peel believed that the attempt to preserve the Corn Laws in the changed social and political conditions of Britain would imperil the rule of the aristocracy. Nevertheless, a majority of his party voted against him, and a smaller number joined the opposition to bring about his defeat and resignation later the same month. For the rest of his career he dedicated himself to the support of free-trade principles and the maintenance of Russell's Whig ministry as the only safeguard against a protectionist government. He died on July 2, 1850, as a result of a riding accident.

Assessment

A proud, shy person, Peel was by nature quick-tempered, courageous, stubborn, and often autocratic. With a first-class intellect, an exact memory, and great capacity for work, he was a superb administrator and an outstanding parliamentary debater. Though he has an unchal-

lenged place as founder of the modern Conservative Party, his political outlook was formed in the pre-reform era. He regarded ministers of the crown as servants of the state rather than as mouthpieces for sectional or party views. By insisting on fundamental changes in the national interest, he did much to preserve the continuity of aristocratic parliamentary government in an age of rapid industrial change, social distress, and class conflict. More than any other man, he was the architect of the mid-Victorian age of stability and prosperity that he did not live to see.

BIBLIOGRAPHY. C.S. PARKER (ed.), *Sir Robert Peel from His Private Correspondence*, 3 vol. (1891–99), essential documentary material; LORD MAHON and E. CARDWELL (eds.), *Memoirs of the Rt. Hon. Sir Robert Peel*, 2 vol. (1856–57), Peel's own account of the three political crises of his career; GEORGE PEEL (ed.), *The Private Letters of Sir Robert Peel* (1920), revealing letters to his wife; NORMAN GASH, *Mr. Secretary Peel: The Life of Sir Robert Peel to 1830* (1961) and *Sir Robert Peel* (1972), a full-length biography based on original sources.

(N.G.)

Peirce, Charles Sanders

In the half-century preceding World War I, the pioneering researches of Charles Sanders Peirce, generally known as the founder of Pragmatism, ranged widely in the sciences, mathematics, logic, and philosophy. Most of his work is now known only to specialists, each grasping a small part of it, severed from its connections with the rest. Even his Pragmatism is viewed in relation to that of other Pragmatists rather than to other parts of his own work. A philosopher will know him also for his evolutionary metaphysics (theory of basic reality) of chance and continuity. A mathematician may known him for his contributions to linear algebra. A logician will know him as one of the creators of the algebra of logic—including the logic of relations; quantification theory (on the usages of "every . . . ", "no . . . ", and "some . . . "); and three-valued logic, which admits a third truth value between true and false—and may know him also for his two systems of logical graphs, which he called entitative and existential. A psychologist may discover in him the first modern psychologist in the United States. A worker in semiotic—the general theory of signs—will know him as cofounder of that science. A philologist may encounter him as an authority on the pronunciation of Elizabethan English. A computer scientist may find in one of his letters the first known sketch of the design and theory of an electric switching-circuit computer. But all of this, and much besides, lay beyond the scope of his professional career, which was that of a physical scientist in the service of the United States government. In that capacity, he was known to his contemporaries for his contributions to astronomy, gravimetry (the measurement of weight or density), spectroscopy, metrology (of weights and measures), geodesy, and the mathematical theory of map projections.

Range of his intellectual creativity

Work in the physical sciences for the Coast Survey. Peirce was born in Cambridge, Massachusetts in September 1839, one of the four sons of Sarah Mills and Benjamin Peirce, who was Perkins professor of astronomy and mathematics at Harvard University. After graduating from Harvard College in 1859 and spending one year with field parties of the U.S. Coast and Geodetic Survey, Peirce entered the Lawrence Scientific School of Harvard University, from which, in 1863, he graduated summa cum laude in chemistry. Meanwhile, he had reentered the Survey in 1861 as a computing aide to his father, who had undertaken the task of determining, from observations of occultations of the Pleiades (by the Moon's passing in front of them), the longitudes of American survey points with respect to European ones. Much of his work, at first chiefly astronomical, was done, on assignment by the superintendent of the Survey, in the Harvard College Observatory, in whose *Annals* (1878) his *Photometric Researches* (toward a more precise determination of the shape of the galaxy of stars to which the solar system belongs) appeared.

Work in astronomy and geodetics

Peirce, 1891.

Peirce's father, who was then superintendent of the Survey, obtained an appropriation from Congress to send parties to the Mediterranean Sea to observe the solar eclipse of December 22, 1870. Several months in advance, Charles, who had already observed an earlier eclipse, was sent to look for suitable observational sites along the expected path of totality. He himself then became a member of the party at one of the Sicilian sites.

In 1871 his father obtained an appropriation to initiate a geodetic connection between the surveys of the Atlantic and Pacific coasts. This cross-continental triangulation lent urgency to the need for a gravimetric survey of North America directed toward a more precise determination of the Earth's ellipticity, a project that Charles was to supervise. For that purpose, he returned to Europe four times to obtain apparatus, to conduct experiments, to participate in meetings of the European Geodetic Association and of its permanent commission, and to consult observers and instrument makers. In pursuit of this project, he contributed to the theory and practice of pendulum swinging as a means of measuring the force of gravity. The need to make accurate measurements of lengths in his pendulum researches, in turn, led him to take home with him the so-called Berlin Meter No. 49—the first authoritative line-metre (a metre standardized by matching with the standards of the Normaleichungsamt Imperial Standards Commission, in Berlin) to be received in the United States—and to make a pioneer determination of the length of the metre in terms of a wavelength of light (1877–79). Between 1873 and 1886 Peirce conducted pendulum experiments at about 20 stations in Europe and the United States and (through deputies) at several other places, including Grinnell Land in the Canadian Arctic.

Peirce was, for a time, in charge of the Office of Weights and Measures, in which capacity he recommended to a Congressional committee that the office be expanded, a step toward the eventual establishment of the National Bureau of Standards in 1901.

Though his experimental and theoretical work on gravity determinations had won international recognition for both him and the Survey, he was in frequent disagreement with its administrators from 1885 onward. The amount of time he took for careful preparation of reports was ascribed to procrastination. His "Report on Gravity at the Smithsonian, Ann Arbor, Madison, and Cornell" (written 1889) was never published because of differences concerning its form and content. When consulted about a proposed new apparatus with a short invariable pendulum, he objected that its invariability was not amenable to confirmation. He finally resigned as of the end of 1891, and, from then until his death in 1914, he had no regular employment or income. For some years he was a consulting chemical engineer, mathematician, and inventor; he played a modest part, for example, in the development of the St. Lawrence Power Company at Massena, New York.

Participation in scientific societies

Peirce was elected a fellow of the American Academy of Arts and Sciences in 1867, a member of the National Academy of Sciences in 1877. He presented 34 papers before the latter from 1878 to 1911, nearly a third of them in logic (others were in mathematics, physics, geodesy, spectroscopy, and experimental psychology). He was elected a member of the London Mathematical Society in 1880. He was active in the 1870s in the Philosophical (*i.e.*, scientific) Society of Washington, in the 1870s and 1880s in the American Metrological Society, and in the 1890s in the New York Mathematical Society (later the American Mathematical Society).

Work in logic. Though Peirce's career was in physical science, his ambitions were in logic. By the age of 31, he had published eight papers, a monograph, and a pamphlet—all of them technical—in that field, besides papers and reviews in chemistry, philology, the philosophy of history and of religion, and the history of philosophy. He had also given two series of Harvard University lectures and one of Lowell Institute lectures, all in logic. Though Peirce aspired to a university chair of logical research, no such chair existed, and none was created for him: the day of logic had not yet come. His nearest approach to this ambition occurred at Johns Hopkins University, where he held a lectureship in logic from 1879 to 1884 while retaining his position in the Survey.

Semiotic and logic

Logic in its widest sense he identified with semiotic, the general theory of signs. He laboured over the distinction between two kinds of action: sign action or semiosis, and dynamic or mechanical action. In 1907 he said,

> I am, as far as I know, a pioneer, or rather a backwoodsman, in the work of clearing and opening up what I call *semiotic*, that is, the doctrine of the essential nature and fundamental varieties of possible semiosis.

His major work, unfinished, was to have been entitled *A System of Logic, Considered as Semiotic.*

Though he made eminent contributions to deductive or mathematical logic, he was a student primarily of "the logic of science"; *i.e.*, of induction and of what he called "retroduction" or "abduction," the forming and accepting on probation of a hypothesis to explain surprising facts. His lifelong ambition was to establish abduction and induction firmly and permanently along with deduction in the very conception of logic—each of them clearly distinguished from the other two, yet positively related to them.

It was for the sake of logic that Peirce diversified his scientific researches so extremely, for the logician, he thought, should ideally have an insider's acquaintance with the methods and reasonings of all the sciences.

Peirce gave two further series of Lowell lectures, on the history of science (1892–93) and on "Some Topics of Logic Bearing on Questions Now Vexed" (1903); a series of Cambridge lectures on "objective logic" (1898); a series of Harvard lectures on Pragmatism (1903); and three on "Logical Methodeutic" (1907). He was a principal contributor on many subjects to *The Century Dictionary* (1889–91) and on logic to James Mark Baldwin's *Dictionary of Philosophy and Psychology* (1901–02). He reviewed for *The Nation* (1869–1908).

Work in philosophy. Peirce's Pragmatism was first elaborated in a series of "Illustrations of the Logic of Science" in the *Popular Science Monthly* in 1877–78. The scientific method, he argued, is one of several ways of fixing beliefs. Beliefs are essentially habits of action. It is characteristic of the method of science that it makes its ideas clear in terms first of the sensible effects of their objects, and second of habits of action adjusted to those effects. Here, for example, is how the mineralogist makes the idea of hardness clear: the sensible effect of x being harder than y is that x will scratch y and not be scratched by it; and believing that x is harder than y means habitually using x to scratch y (as in dividing a sheet of glass) and keeping x away from y when y is to remain unscratched. By the same method Peirce tried to give equal clarity to the much more complex, difficult, and important idea of probability. In his Harvard lectures of 1903, he identified Pragmatism more narrowly with the logic of abduction. Even his evolutionary metaphysics of 1891–93 was a higher order working hypothesis by which the special sciences might be guided in forming their lower order hypotheses; thus, his more metaphysical writings, with their emphases on chance and continuity, were but further illustrations of the logic of science.

When Pragmatism became a popular movement in the early 1900s, Peirce was dissatisfied both with all of the forms of Pragmatism then current and with his own original exposition of it, and his last productive years were devoted in large part to its radical revision and systematic completion, and to the proof of the principle of what he by then had come to call "pragmatism."

His "one contribution to philosophy," he thought, was his "new list of categories" analogous to Kant's a priori forms of the understanding, which he reduced from 12 to three: Quality, Relation, and Representation. In later writings he sometimes called them Quality, Reaction, and Mediation; and finally, Firstness, Secondness, and Thirdness. At first he called them concepts; later, irreducible elements of concepts—the univalent, bivalent, and trivalent elements. They appear in that order, for example, in his division of the modalities into possibility, actuality, and necessity; in his division of signs into icons, indexes, and symbols; on the division of symbols into terms, propositions, and arguments; and in his division of arguments into abductions, inductions, and deductions. The primary function of the new list was to give systematic support to this last division. (For the philosophy of Peirce, especially as it relates to Pragmatism, see PRAGMATISM: *History of Pragmatism: The classical Pragmatists.*

Peirce was twice married: first in 1862 to Harriet Melusina Fay, who left him in 1876, and second in 1883 to Mme Juliette Pourtalai (*née* Froissy). There were no children of either marriage. For the last 26 years of his life, he and Juliette lived on a farm on the Delaware River near Milford, Pennsylvania. He called himself a bucolic logician, a recluse for logic's sake. He lived his last years in serious illness and in abject poverty relieved only by aid from such friends as William James—in whose honour Peirce added Santiago (St. James) as his middle name. He died of cancer on April 19, 1914, on his farm. The Peirce house, acquired by the National Park Service in 1972, was to become a national memorial.

Peirce is now recognized as the most original and the most versatile intellect that the Americas have so far produced. The recognition was slow in coming, however, and much of his work is still unexplored.

BIBLIOGRAPHY

The Peirce papers: Collected Papers, ed. by CHARLES HARTSHORNE, PAUL WEISS, and ARTHUR W. BURKS, 8 vol. (1931–58), with a bibliography of Peirce's writings in vol. 8, pp. 249–330. The Peirce papers in the Houghton Library at Harvard University are listed and described by RICHARD S. ROBIN in his *Annotated Catalogue of the Papers of Charles S. Peirce* (1967), and in "The Peirce Papers: A Supplementary Catalogue," *Transactions of the Charles S. Peirce Society*, 7:35–57 (1971). A microfilm edition of the greater part of the collection is available from the Harvard University Library microreproduction service. CAROLYN EISELE has prepared an edition of his mathematical writings (forthcoming).

Studies: Studies in the Philosophy of Charles Sanders Peirce (1952), was edited by PHILIP P. WIENER and FREDERIC H. YOUNG. The *Second Series*, ed. by EDWARD C. MOORE and RICHARD S. ROBIN (1964), contains a supplement to Burks's bibliography of Peirce's own writings (pp. 477–485) and also a bibliography of writings about Peirce (pp. 486–514). There are further supplements in *Transactions . . .*, 2:51–59 (1966). CAROLYN EISELE and VICTOR F. LENZEN have published numerous studies of Peirce's work in the sciences, mathematics, and in the history of science (see the bibliographies mentioned above). Of the many books on Peirce's philosophy, the one that makes most use of unpublished materials and has greatest biographical value is MURRAY G. MURPHEY, *The Development of Peirce's Philosophy* (1961).

Peisistratus

Peisistratus, who was born around the beginning of the 6th century BC, was a major figure in the political, economic, religious, and cultural life of Athens during that century.

Rise to power. In 594 his mother's relative, the reformer Solon, had improved the economic position of the Athenian lower classes, but the Solonian reorganization of the constitution had not eliminated bitter aristocratic contentions for control of the archonship, the chief executive post. As Peisistratus reached manhood, the two major vying factions were called the Plain, led by Lycurgus, and the Coast, led by Megacles.

During a war with the city of Megara about 565, Peisistratus gained military fame by taking the Megarian harbour. He organized his own faction, named the Hillsmen, a group that included noble families from his own district, the eastern part of Attica, and also a very considerable part of the growing population of the city of Athens. At one point Peisistratus slashed himself and the mules of his chariot and made a dramatic entrance into the agora (marketplace) to show how his enemies had wounded him. The people voted him a bodyguard of citizens armed with clubs, with the aid of which he seized the Acropolis and held power briefly in 561. To increase his support he contracted a short-lived marriage with the daughter of Megacles and again acquired temporary power in Athens (probably 556–555), but Lycurgus and Megacles united to force him out.

For several years Peisistratus was an exile in northern Greece. He laid a solid base for his return, exploiting the silver and gold mines of Mt. Pangaeum and gaining the support of conservatives in Thebes, Argos, Naxos, and elsewhere. In 546 he went to Eretria on the island of Euboea, with the force provided by his own funds and by his friends, and from this base invaded Attica. At Pallene, near Mt. Hymettus, he launched a surprise attack on the Athenian army in the heat of midday, while his enemies were gambling or sleeping. After a complete victory, Peisistratus became master of Athens for the third time and remained in power until his death in 527. His sons Hippias and Hipparchus succeeded him.

The occasional modern efforts to interpret the factions involved in the rise of Peisistratus as representing simply different economic interests or purely geographical blocs seem misconceived, nor can one hope to penetrate the legends about Peisistratus to discern his personality. His career down to 546 shows persistence, dexterity, and diplomatic ability; once master, he followed a program that pleased the city population especially but that also appealed to the rural majority.

Tyrant of Athens. Peisistratus was master of Athens by the use of force, so in Greek terms he was a *tyrannos.* He maintained a mercenary bodyguard, composed in part of Scythian archers; he may have disarmed the citizens; and he certainly placed hostages from major families in safekeeping on the island of Naxos. Yet he preserved the constitutional forms of government and made them operate more efficiently. Some aristocrats cooperated and were permitted to hold the yearly post of archon; others went into exile. Once Peisistratus, accused of homicide, appeared before the court on the day of the trial; but his accuser dared not press the charge.

Religious and other developments

His internal policies appear to have been designed to increase the unity and majesty of the Athenian state. Since religion was closely interwoven with the structure of the Greek *polis* or city-state, many of his steps were religious reforms. He brought the great shrine of Demeter at Eleusis under state control and constructed the first major Hall of the Mysteries (Telesterion) for the annual rites of initiation into the cult. Many local cults of Attica were either moved to the city or had branch shrines there. Artemis, for instance, continued to be worshipped at Brauron, but now there was also a shrine to Artemis on the Acropolis. Above all, Athena now became the main deity to be revered by all Athenian citizens. Peisistratus constructed an entry gate (Propylaea) on the Acropolis and perhaps built an old Parthenon under the temple that now stands on the crest of the Acropolis. Many sculptured fragments of limestone from Peisistratid buildings have been found on the Acropolis, and the foundations of a major, unfinished temple can still be seen.

Festivals and literature also flourished in Peisistratid times. The tyrant enhanced the glory of the Panathenaea, a yearly festival to Athena, by accentuating the Great Panathenaea (every four years) with athletic contests and prizes for bards who recited the Homeric epics. After the cult of Dionysus was placed under state sponsorship, prizes were awarded at the yearly Dionysia for the singing of dithyrambs and, from 534, for the performance of tragedies. Poets such as Anacreon lived at the court of Peisistratus and his sons, who also encouraged the collection of oracles and supported the famous soothsayer Onomacritus.

Contribution to the growth of Athens. At this time Athens itself was becoming a city, rather than an agglomeration of villages. Peisistratus improved its water supply by building an aqueduct that fed the Enneakrounos fountain on the edge of the agora. He also beautified and systematized the marketplace itself; 6th-century markers of its borders have been found in agora excavations. Just outside the city, on the banks of the Ilissus stream, he began a temple to Olympian Zeus, but this was not finished until the reign of the Roman emperor Hadrian.

In the countryside, Solon had encouraged the growing of olive trees and vines to produce cash crops; Peisistratus made loans to small farmers for tools and equipment. In a few cases the estates of exiled aristocrats appear to have been broken up, but the major force in reducing aristocratic control over rural Attica seems to have been the regularization of government. Peisistratus instituted a system of travelling judges to provide state trials of rural cases on the spot; he himself made inspection tours.

This extensive cultural and political activity was financed by Peisistratus' revenues from the mines of Mt. Pangaeum and from internal sources. The silver mines of Laurium were state property, and dues were exacted from the growing trade at Athenian harbours. Peisistratus instituted a tax, probably of 10 percent, on agricultural production. On one tour of inspection, according to a famous story, he saw a farmer digging in a field of stones and asked what his income was. When the farmer replied, "Just so many aches and pains; and of these aches and pains Peisistratus ought to take his ten percent," the tyrant remitted all taxes to the frank farmer.

Athenian industry and commerce expanded tremendously in the latter half of the 6th century; the main contribution of Peisistratid rule to these developments was probably the guarantee of internal tranquility and the protection of foreign immigrants.

External policy

Externally, the tyrant pursued a policy of peace, probably because he dared not allow the Athenian citizenry to bear arms in a major war. But at this time the Greek world was also in a temporary state of balance. In the Aegean, Peisistratus helped such friends as Lygdamis of Naxos to become local tyrants. He purified the sacred island of Delos by removing the old graves near its temple of Apollo. His main efforts, however, were concentrated in gaining control of the Hellespont, through which came the exported grain of south Russia. To this end he secured command of Sigeum and installed a younger son, Hegesistratus, as its ruler. More important, he encouraged the Athenian Miltiades to lead a private venture that gained mastery over the Chersonesus.

On the death of Peisistratus, Athens was still much less important politically and militarily than was Sparta. Commercially, states such as Miletus, Corinth, and Aegina were at least as active, and the contemporary tyrant Polycrates of Samos was as important a patron of the arts and letters. Nonetheless, the religious and patriotic unification of Athens had made great progress during Peisistratus' calm, even rule. As Aristotle reports, it became a common saying that the tyranny of Peisistratus had been the age of Cronus, the golden age.

BIBLIOGRAPHY. The main ancient sources are HERODOTUS, bk. 1, ch. 59–64; and ARISTOTLE, *Constitution of Athens*, ch. 13–17. N.G.L. HAMMOND, *A History of Greece to 322 B.C.*, 2nd ed., ch. 6 and 8 (1967), cites other minor sources and gives the political development. On the constitutional side, see C. HIGNETT, *A History of the Athenian Constitution to the End of the Fifth Century B.C.*, ch. 5 (1952). A. ANDREWES, *The Greek Tyrants* (1956), thoughtfully compares Peisistratus with other Greek tyrants.

(C.G.St.)